Object detection based on Yolov4-Tiny and Improved Bidirectional feature pyramid network

Object detection based on Yolov4-Tiny and Improved Bi-directional feature pyramid network

Qi Liu[1 a*], Xiaoyu Fan[1 b*], Zhipeng Xi[1 c*], Zhijian Yin[1 d*], Zhen Yang[1 e*]

[1] Jiangxi Science and Technology Normal University, Nanchang, China

[a*]criafting@aliyun.com, [b*]1020170701@jxstnu.com, [c*]John_boy0618@163.com, [d*]zhijianyin@aliyun.com, [e*] yangzhenphd@aliyun.com

Abstract—In the field of small object detection, Yolov4-Tiny is inadequate in feature extraction and does not make best of multi-scale features. In this paper, an improved BiFPN framework is proposed based on Yolov4-Tiny to increase object detection precision. Moreover, the Yolov4-Tiny is taken as the backbone network and introduce spatial pyramid pooling (SPP) to connect and merge multi-scale regions. Finally, our method can achieve 79.53% map on Pascal VOC dataset, which is 2.12% higher than the original Yolov4-Tiny model.

1. Introduction

Object detection has become the foundation for solving more complex and advanced visual tasks, such as instance segmentation[1], scene understanding[2], target tracking[3] and so on. However, object detection methods are not only locating the different object in an image, also for classifying object category. Recently, these methods of object detection based on deep neuro network have been widely taken in pedestrian detection, worker's helmet detection, medical diagnosis et.al.

Early stage, object detection had low precision and less scope of application, so it is tough to achieve landing in practical application. The traditional object detection methods used many candidate regions by sliding window and extracting the features from the candidate regions, such as Haar-like feature[4], HOG[5], LBP[6] and DPM[7](Deformable Part Models). Various target detection methods based on handcrafted features had been proposed, and performed well on Pascal VOC dataset.[8] However, the above methods determine the candidate regions by window sliding that will lead to increase computational complexity.

More efficient methods begin to emerge in the field of image vision with the development of computer technology, especially the emergence of deep learning. Krizhevsky et al. proposed a convolutional neural networks (CNN) called AlexNet[9] in 2012 which got the best image classification accuracy in ILSRVC, It is proved that using CNN to extract features can classify images more accurately than traditional machine learning feature extraction methods. Meanwhile, in terms of target detection, AlexNet opens up a new method.

Since then, more and more researchers have begun to pay attention to deep learning, and people have gradually begun to try to use CNN to solve the problems in target detection[10]. In the field of target detection, traditional methods have been replaced, and convolutional neural networks have been rapidly developed and applied. Generally speaking, according to whether the network generates candidate frames, two mainstream methods have gradually formed in the field of target detection. One of them is a stage, and the other is one-stage. The main object detection methods based on two-stage are R-CNN [11], SPPnet [12], Fast R-CNN[13], and Faster R-CNN[14]. By using the selective search algorithm to

generate 2000 candidate boxes on each image, and sending these 2000 candidate boxes into CNN to extract features and SVM for classification in R-CNN. Using these candidate boxes does improve the performance of detection, but it also slows down the running speed of the network, especially when the data set is relatively large. In response to the question of the slow detection speed of R-CNN, SPPnet and Fast R-CNN were proposed. SPPnet is a spatial pyramid pooling at the top of R-CNN last convolution layer for a fixed-length feature map. Although, the SPPnet detection speed has improved, the training speed is still slow. Fast R-CNN can not only improve training speed and detection speed but also improve accuracy via sharing convolutional computing regions. Faster R-CNN propose a region proposal network (RPN) to generate candidate regions to replace the traditional selective search algorithm, and the speed is improved. Object detection methods based on one-stage include Yolo [15] and SSD [16], etc. The core idea of Yolo is that target detection is regarded as regression problem. The size of the input image is cropped to the set size, and then separated into S×S grid units. Each grid cell predicts B bounding boxes, and the position information, confidence and category of bounding boxes are obtained by regression in the output layer. For the one-stage target detection method SSD, selecting VGG16 as the backbone network when extracting features, borrow the anchor idea in Faster R-CNN to detect in different receptive fields, and uses Yolo regression idea to output boundary boxes and categories.

Multiple versions have evolved due to the Yolo and its variants [17][18]being highly representative and the potential for improvement in a single-stage inspection framework. Among them, Yolov4[19] and Yolov4-Tiny have excellent performance in accuracy and speed respectively and have a large space for improvement.

The multi-feature map fusion method improves the detection accuracy by fusing the feature maps of different scales and the receptive field information to form a new feature map. Therefore, CNN models don't need to fix the size of the input images, but also have robustness to detect multi-scale objects by fusing multi-scale features. Therefore, the finer layer of the network can take advantage of the context information learned from the coarser layer of the network. Lin et al. Proposed FPN (feature pyramid network)[20] method based on the inherent multi-scale pyramid structure of deep convolution neural network which a top-down transverse connection structure was designed to improve the accuracy of multi-scale detection. In recent years, FPN has been widely cited in the image field by using the advantages of fusing multi-scale features to obtain global information. PANet[21], NAS-FPN[22] and BiFPN[23]More network structures have been developed for cross-scale feature fusion to obtain more discriminative information which aiming to fuse feature outputs from different network layers, to enhance feature information and improve target detection accuracy.

Yolov4-Tiny target detection framework cannot make best of multi-scale features and insufficient feature extraction ability inspired by these classic research methods. This paper uses Yolov4-Tiny as the backbone network, then, introduces SPP and improved BiFPN feature extraction network to enhance the feature extraction ability. This paper aims to maintain the fast detection speed and small network volume as the premise to improve the accuracy of target detection. The major work and contributions are as follows:

1) We introduce the SPP module. The SPP module is mainly composed of three maximum pooling layers of different sizes. The pooling layers of different sizes extract features from different angles to form feature maps of receptive fields of different sizes.

2) We embed an improved BiFPN into the structure of Yolov4-Tiny that can fuse the features between different scales.

3) The improvements above are introduced in Yolov4-Tiny, which effectively alleviate the problems of insufficient feature extraction ability and inadequate utilization of multi-scale features.

2. Yolov4-Tiny network architecture

As a simplified version of Yolov4-Tiny, which has a relatively simple network structure, a small amount of computing, and a lower requirement on hardware, so it is easy to run in mobile terminal or device terminal. The Yolov4-Tiny network architecture (Figure 1) consists of the backbone network, neck, and

the head.

The main network of Yolov4-Tiny uses the network structure of CSPDarknet53-tiny, consisting mainly of two normal convolution processes and three CSP structure. The common convolution process adopts CBL structure, which is composed of the convolution layer, BatchNorm layer and LeakyRelu; BatchNorm layer can reduce the difference of range between different samples, speed up the training and convergence speed of the model on the computer, and prevent gradient explosion and gradient disappearance problems during the back propagation process. For the part of input less than zero, LeakyReLU can calculate the gradient, and avoid the phenomenon of neuron inactivation. This structure can reduce model parameters to reduce computational complexity while ensuring the accuracy of the network. CSPBlock compares the structural features of CSPNet and divides the features of base layer into two parts. The main function of the first part is to directly form the residual edge, and the final output can be obtained by convolution using the splicing of the second part. This network structure can reduce computational complexity while maintaining network performance.

In Neck part, Yolov4-Tiny adopts FPN structure, which can integrate features of different scales for implementing rich semantic information of deep network and geometric details of shallow network for strengthening the ability of extracting features. Yolo head used the feature map obtained by FPN to make the final prediction, and finally formed two prediction scales of 26×26 and 13×13.

In Yolov4-Tiny, the input image size is 416×416, and then multi-scale detection is carried out by fusing FPN idea, and two detection layers are output. The two detection layers divide the image into 26×26 and 13×13 respectively. Then each grid cell has predicted three bounding box, each boundary box contains (x, y, w, h, Ci) five elements. Among them, the x and y are referred to the width and height after normalization based on bounding box of the entire image corresponding to the cell w and h of the offset. Ci means the bounding box confidence score, reflects a box contains the possibility of detecting target, and the accuracy of the test box. In order to better predict bounding boxes, Yolov4-Tiny generates six anchor boxes through k-means clustering algorithm, which serves as suggestion boxes during detection. In addition, the loss function of Yolov4-Tiny consists of three parts: including bounding box regression loss, confidence loss and classification loss. The bounding box regression loss function adopts CIOU function, and the latter two use the same cross entropy function as Yolov3 [18].

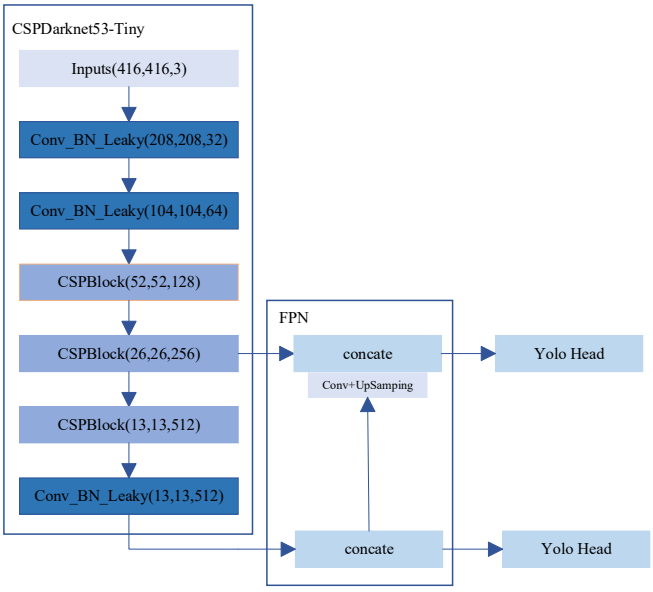

Figure 1 Network architecture of Yolov4-Tiny

3. methods

In this work, an improved target detection method based on Yolov4-Tiny is proposed. (hereinafter referred to as Yolov4-Tiny-Bi). The backbone network adopts CSPDarknet53-tiny, and the head uses a

spatial pyramid pool and the improved BiFPN is introduced to make most use of the multi-scale information between different layers. Figure 2 shows the network structure.

3.1 Improved BiFPN

Yolov4-Tiny contains two prediction scales only, 13×13 and 26×26, two feature maps are output from the deeper network, while the deeper network is easy to lose the shallow edge information. The semantics of features can change low dimension to high dimension with the deepening of the network level. Each layer of the network will cause the loss of features to a certain extent. It is necessary to integrate features at different levels to enrich the semantic information of features. BiFPN make the most of the features between shallow network and deep network, and carries out bidirectional fusion of network features from bottom-up and from top-down. Therefore, the improved BiFPN is used to replace the original FPN.

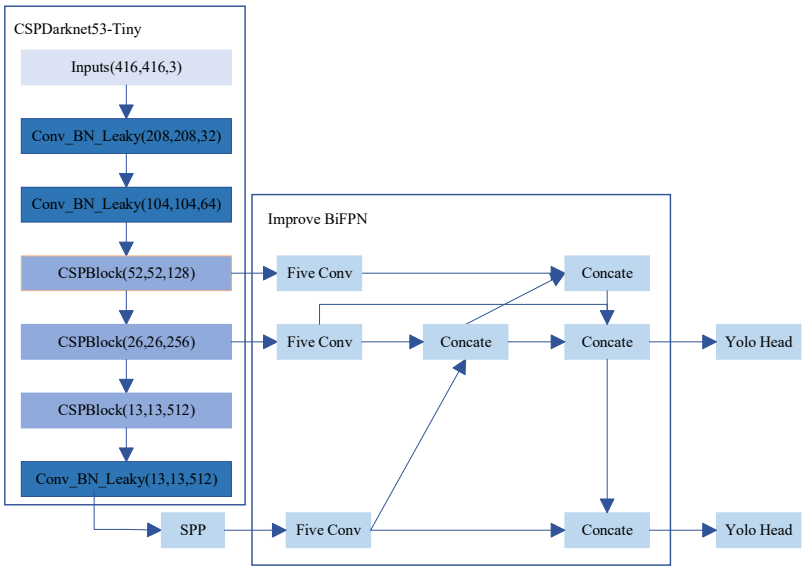

Figure 2 Network architecture of Yolov4-Tiny-Bi

BiFPN (Figure 3) use the idea of bidirectional fusion to construct a bottom-up and top-down bidirectional channel to carry out bidirectional fusion of information from diverse layers of the trunk network, so as to alleviate the loss of feature information caused by too many network layers.

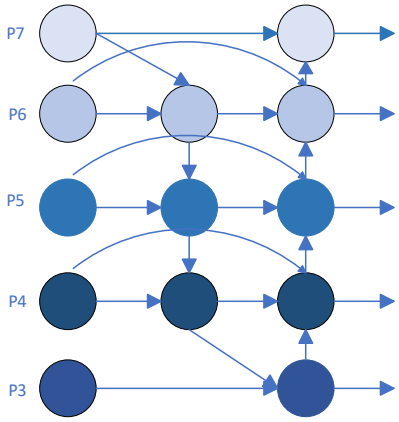

Figure 3 Network architecture of BiFPN

BiFPN is an improvement of PANet which removes the node with only one input edge, adding an edge from the primordial input to the last output node. Each bidirectional path is considered as a characteristic network layer and can be repeated several times. Bidirectional cross scale connections and the fast normalized fusion is integrated in our final BiFPN, the feature fusion method is shown in Equation (1):

$$P_5^{td} = conv\left(\frac{\omega_1 * P_5^{in} + \omega_2 * Resize(P_6^{td})}{\omega_1 + \omega_2 + \epsilon}\right) \quad (1)$$

$$P_5^{out} = conv\left(\frac{\omega_1' * P_5^{in} + \omega_2' * P_5^{td} + \omega_3' * Resize(P_4^{out})}{\omega_1' + \omega_2' + \omega_3' + \epsilon}\right) \quad (2)$$

In this formula P_5^{td} and P_5^{out} is the intellectual feature and the output feature at level 4, respectively.

The improved BiFPN is introduced between the backbone network and the head in Figure 4. We extract features of three different scales from the backbone network, we first carry out five convolution for these three feature layers of different scales, which consists of two 1×1 convolution layers and three depth separable convolution layers. We integrate the three different scale feature layers after five convolution into the improved BiFPN network, fusing the multi-scale features and setting the prediction branches of 13×13, 26×26 feature resolutions. In order to fully integrate the shallow features, we change the BiFPN bi-directional channel top down and bottom up. The improved feature fusion method is shown in Equation (2):

$$P_4^{td} = conv\left(P_4^{in} + Resize\left(P_5^{in}\right)\right) \quad (3)$$

$$P_4^{out} = conv\left(P_4^{in} + P_4^{td} + Resize(P_3^{out})\right) \quad (4)$$

In this formula P_4^{td} and P_4^{out} is the intellectual feature and the output feature at level 4, respectively.

Different from the BiFPN proposed by Google et al. our new BiFPN in Yolov4-Tiny-Bi does not integrate the network features from bottom-up and top-down. Instead, we first convolved layers P3, P4 and P5 for five times, then carried out two-way fusion of deep and shallow layer features from bottom-up and top-down, and output only two prediction scales of 13×13, 26×26 at the end.

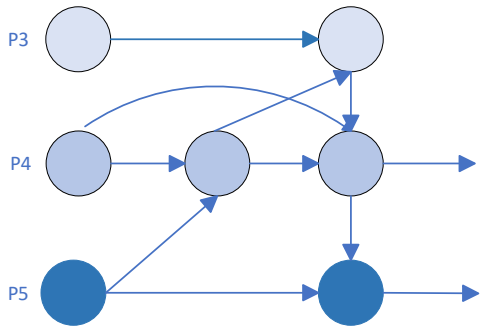

Figure 4 Network architecture of improved BiFPN

3.2 SPP

Spatial pyramid pooling (SPP) was proposed by He et al. SPP uses three sliding Windows of different sizes to maximize the pooling of the input feature graphs, the feature map is assembled by sliding windows whose size is the same as that of the plane elements, and one-dimensional feature vectors are obtained. Therefore, it is not feasible to apply it in full convolutional networks. Therefore, Redmon and Farhadi improve the SPP module to a series of maximum pool outputs with the kernel size of K × K, where k = {1, 5, 9, 13} and the span is equal to 1 in the design of Yolov3[18]. The receptive field of the backbone features is effectively increased by a relatively large K × K maximum pool. We add the SPP block to P5 because it effectively increases the receptive field, isolating above all contextual features that results in almost no reduction in network speed.

4. Experimental Analysis

4.1 Dataset

The PASCAL VOC [24] dataset is used in this article. This data includes 20 categories of objects. The 20 categories include planes, trains, people, potted plants, birds and so on. Yolov4-tiny is one of the most popular convolutional neural networks for embedded computing devices, its model size is small and their computational complexity is low, so they are used as a reference. The VOC2007 training set was used to train Yolov4-Tiny-Bi network, and average accuracy was calculated on the test set to evaluate the detection accuracy.

4.2 Model training

The detailed hardware configuration in this experiment is: the CPU of the computer is Intel(R) Core(TM) i7-10750H, frequency is 2.60GHz, the memory is 16GB, and the graphics card model is NVIDIA GeForce GTX 3060. The operating system for 64-bit Windows10 system, deep learning development platform uses Pytorch1.8. The parallel computing framework is CUDA10.0, and the deep neural network accelerator library is CUDNN V11.1.

When Pascal VOC2007 is trained, the initial learning rate is set to 0.001 and the batch size is set to 32.

4.3 Experimental results and analysis

The precision of Yolov4-Tiny-Bi is greatly improved due to its lightweight network structure by comparison with the Yolov4-Tiny. Compared with Yolov4-Tiny, MAP is 2.12 percentage points higher due to the introduction of improved BiFPN structure to enhance the feature fusion between different levels and make full use of shallow features. Meantime, the detection speed of the improved BiFPN structure has little change because the calculation amount is small. Finally, the map of Yolov4-Tiny-Bi algorithm using 416×416 resolution, SPP structure and improved BiFPN structure reached 79.53%, which was higher than that of Yolov4-Tiny, and achieved the expected effect.

Table 1. Comparison of the accuracy on the VOC2007 dataset

method	dataset	MAP/%
Yolov4-Tiny	VOC2007	77.41
Yolov4-Tiny-Bi	VOC2007	79.53

Figure 5 shows the comparison of detection results between Yolov4-Tiny and Yolov4-Tiny-Bi, indicating that Yolov4-Tiny has poor detection effect on occlusive objects and the problem of missing detection, while Yolov4-Tiny-Bi proposed in this paper detect most of the targets.

5. conclusion

Yolov4-Tiny algorithm has insufficient feature extraction ability in small target detection and can't take advantage of multi-scale features. In this paper, a bidirectional feature pyramid network target detection algorithm based on Yolov4-Tiny is proposed. An improved BiFPN is proposed and spatial pyramid pooling is introduced to improve the generalization ability of the model in actual detection scenes and make the model more geared for object detection tasks. Experimental results show that compared with the original algorithm Yolov4-Tiny, MAP improves by 2.12% on Pascal VOC dataset, and the detection rate reaches 72 frame/s.

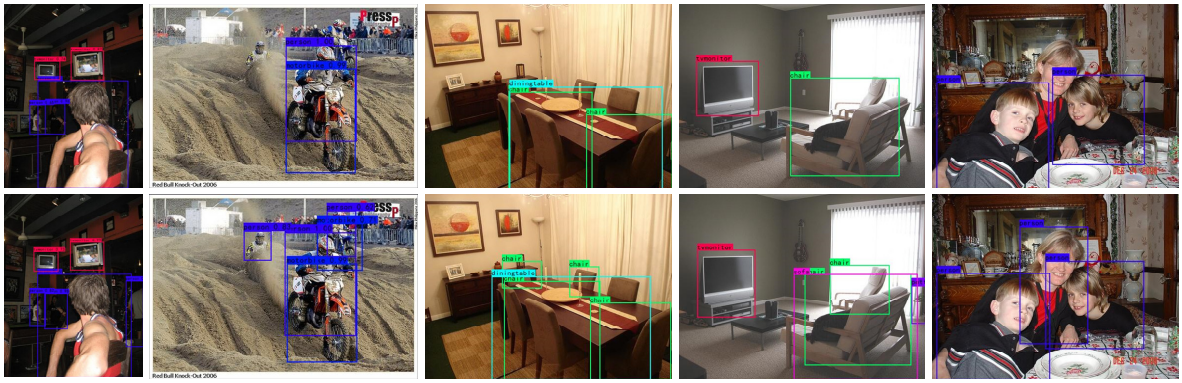

Figure 5 Yolov4-Tiny and Yolov4-Tiny-Bi comparison of prediction results

Experimental results show the Yolov4-Tiny network model combining the improved BiFPN and SPP structure can meet the timeliness and accuracy of object detection, and has a good engineering application. The accuracy of the Yolov4-Tiny-Bi target detection algorithm in this article is less than that of some large-scale detection networks. In future research work, we will further study and optimize Yolov4-Tiny-Bi to make the model more suitable for more detection scenarios.

Acknowledgement
This work was supported by the National Natural Science Foundation of China (62061019, 61866016), General Project of Jiangxi Natural Science Foundation(20202BABL202014), the General Project of Jiangxi Education Department(GJJ190587), and the Youth Top Talent Foundation of Jiangxi Science and Technology Normal University (2018QNBJRC002).

REFERENCES

[1] Peng S., Jiang W., Pi H., Li X., Bao H., & Zhou X. (2020). Deep snake for real-time instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8533-8542.
[2] Pan X., Shi J., Luo, P., Wang, X., & Tang, X. (2018). Spatial as deep: Spatial cnn for traffic scene understanding. In Thirty-Second AAAI Conference on Artificial Intelligence.
[3] Bertinetto L., Valmadre J., Henriques J. F., Vedaldi A., & Torr P. H. (2016). Fully-convolutional siamese networks for object tracking. In European conference on computer vision. pp. 850-865.
[4] Papageorgiou C. P., Oren M., & Poggio T. (1998). A general framework for object detection. In Sixth International Conference on Computer Vision. pp. 555-562.
[5] Dalal N., & Triggs B. (2005). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition. pp. 886-893.
[6] Ojala T., Pietikainen M., & Maenpaa T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on pattern analysis and machine intelligence, 24(7): 971-987.
[7] Felzenszwalb P. F., Girshick R. B., McAllester D., & Ramanan D. (2009). Object detection with discriminatively trained part-based models. IEEE transactions on pattern analysis and machine intelligence, 32(9): 1627-1645.
[8] Zhang J., Huang K., Yu Y., & Tan T. (2011). Boosted local structured hog-lbp for object localization. In CVPR 2011.pp. 1393-1400.

[9] Krizhevsky A., Sutskever I., &Hinton G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25: 1097-1105.
[10] Girshick R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision. pp. 1440-1448.
[11] Girshick R., Donahue J., Darrell T., & Malik, J. (2014). 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.
[12] He K., Zhang X., Ren S., & Sun J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 37(9):1904-1916.
[13] Girshick R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision. pp. 1440-1448.
[14] Ren S., He K., Girshick R., & Sun J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, 28:91-99.
[15] Redmon J., Divvala S., Girshick R., & Farhadi A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779-788.
[16] Liu W., Anguelov D., Erhan D., Szegedy C., Reed, S., Fu C. Y., & Berg A. C. (2016). SSD: Single shot multibox detector. In European conference on computer vision. pp. 21-37.
[17] Redmon J., & Farhadi A. (2017). YOLO9000: better, faster, stronger. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7263-7271.
[18] Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767.
[19] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.
[20] Lin T. Y., Dollár P., Girshick R., He K., Hariharan B., & Belongie S. (2017). Feature pyramid networks for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2117-2125.
[21] Liu S., Qi L., Qin H., Shi J., & Jia J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8759-8768.
[22] Ghiasi G., Lin T. Y., & Le Q. V. (2019). Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7036-7045.
[23] Tan M., Pang R., & Le Q. V. (2020). Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10781-10790.
[24] Everingham M., Van Gool L., Williams C. K., Winn J., & Zisserman A. (2010). The pascal visual object classes (VOC) challenge. International journal of computer vision, 88(2): 303-338.