On the Multi-scale Real-Time Object Detection Using ResNet

Zhengyao Bai(✉) and Dong Jiang

School of Information Science and Engineering, Yunnan University, Kunming 650500, China
baizhy@ynu.edu.cn, l012833912@qq.com

Abstract. Real-time target detection and location has important values in video surveillance. Aimed at the low accuracy of existing real-time object detection algorithms, this paper proposes a multi-scale real-time target detection algorithm based on residual convolution neural network. Firstly, the residual convolutional neural network is introduced into the YOLOv3-Tiny algorithm. The jump connection of the low-level and high-level networks forms the residual module in the YOLOv3-Tiny algorithm to effectively prevent network degradation while increasing the depth of the neural network. Secondly, a new prediction layer is added to the neural network to improve the results of small target detection. Finally, the trained model is tested on the Pascal VOC public dataset. The experimental results show that the proposed algorithm achieves 64.6% accuracy on the validation dataset, and the speed of 60FPS in the video detection. The detection accuracy is improved to a higher level at a small cost of a little lower speed still meeting the real-time detection requirements, and small targets in the image can be effectively detected. The algorithm is effective and robust.

Keywords: Real-time object detection · Residual convolutional neural network · YOLOv3-Tiny

1 Introduction

The task of object detection is to find the objects in the image or video, determine their position, size and category. Therefore, fast and accurate object detection has always been one of the research hotspots in the field of computer vision. In early stage, Viola [1] train the neural network to achieve face target detection based on the AdaBoost algorithm, although the method can process images quickly and it has high accuracy, but category is single and can't be widely used. With the development of deep learning, the convolutional neural network has achieved breakthrough results in image classification and object detection. Object detection algorithms based on deep learning can be divided into two categories, one is based on candidate regions, the other one is based on regression.

The typical algorithm based on regression is the YOLO [2] (You Only Look Once) series of algorithms. The YOLOv1 algorithm directly uses a convolution network to complete the classification and localization of targets. Due to the reduced amount of repetitive calculation, the YOLOv1 algorithm is fast. However, the detection accuracy

Z. Lin et al. (Eds.): PRCV 2019, LNCS 11857, pp. 63–73, 2019.
https://doi.org/10.1007/978-3-030-31654-9_6

is slightly lower. In the YOLOv2 [3] algorithm, Redmon used the anchor box strategy in Faster RCNN, but unlike the empirically driven anchor box in Faster RCNN [4], YOLOv2 uses the clustering method to analyze the anchor box, and the final object detection result is better. In order to further improve the detection effect of the network on small targets, Redmon borrowed the idea of Feature Pyramid Networks [5] (FPN) to perform object detection on multiple scales, and proposed the YOLOv3 [6] algorithm, which makes the detection effect of small target better.

The faster detection speed in the YOLO series is the YOLOv3-Tiny algorithm, but its detection accuracy is lower. In view of this problem, this paper combine the residual convolution network with YOLOV3-Tiny network to improve the processing of image features in convolutional networks and improve the detection accuracy of the algorithm. On the basis of the fusion of the convolutional neural network, a layer of prediction is added to improve the detection effect on small targets. The experimental results show that the combination network improves the detection accuracy and improves the detection effect on small targets, while the detection speed can still meet the requirements of real-time detection.

2 YOLOv3-Tiny Network Modifications

2.1 Detection Procedure in YOLO

The YOLO series algorithm does not need to generate candidate boxes for the classification and detection, it can directly output the target categories and positions simultaneously after the neural network training. First, the YOLO series algorithm extracts the input image through the convolutional neural network, and obtains a fixed-size feature map through a series of convolution processing and down sampling operations. For example, for image size is 416 × 416, through the YOLO-Tiny network processing, a feature map with a size of 13 × 13 can be obtained. According to the obtained feature map, the YOLO-Tiny network divides the input image into several grids. When the center coordinates of the target are within a certain grid, the grid is responsible for predicting it. When predicting, each grid predicts a fixed number of bounding boxes. The detection step of YOLO-Tiny algorithm is shown in Table 1.

Table 1. YOLO-Tiny algorithm

1: Input an image and resize to 416x416
2: Divides the image into an S×S grid
3: Whether the target center is in some grid
if target's center in some grid:
this grid is responsible for predict the target, it predicts B bound boxes, confidence for those boxes, and C class probabilities.
end if
4: repeat step 3.
end.

2.2 Multi-scale Prediction

In the YOLO algorithms, the number of bounding boxes predicted by each grid in the YOLOv1 network is two. In YOLOv2 network it is increased to five, and YOLOv2 predict 845 bounding boxes on a feature map of size 13 × 13. While in the YOLOv3 network the number of boxes is three, but it is predicted on three different size feature maps. Therefore, the actual number of bounding boxes is 10,647, which is more than ten times of YOLOv2. The number of bounding boxes increases, which makes the algorithms detect object more precise. Since the small-scale feature map extracted by the low-level network in the convolutional neural network contains more position information of the image, such as detecting on the feature map of size 52 × 52, more position information about the small target can be obtained, and the feature map extracted by the high-level network contains more detailed information of the image, which is beneficial to the recognition of the target category. Therefore, the feature map extracted by different layers can be detected separately to achieve multi-scale prediction of specific targets. In order to make the bounding box better predict the target, the author uses the k-means clustering method to cluster the labeled real positioning boxes in the coco dataset [7], and obtains nine different sizes of anchor boxes, as shown in Table 2.

Table 2. Anchor box for different size feature maps

Feature map	Anchor box
13 × 13	(116 × 90), (156 × 198), (373 × 326)
26 × 26	(30 × 61), (62 × 45), (59 × 119)
52 × 52	(10 × 13), (16 × 30), (33 × 23)

As can be seen from Table 2, because the feature map of size 13 × 13 has the largest receptive field and contains the most abundant feature information, it is suitable for detecting the large target existing in the image. Therefore, the three largest anchor boxes are used for detection this layer. The 26 × 26 feature map uses a medium-sized anchor box to detect medium-sized targets in the image. The 52 × 52 feature map receptive field contains more position information of the target, so it is suitable for detecting smaller targets in the image. Therefore, the prediction is performed using the smallest three anchor boxes.

2.3 Modifying YOLO-Tiny Using ResNet

YOLOv3-Tiny is a simplified version of YOLOv3 with fast detection speed but low accuracy. The YOLOv3-Tiny network only uses two scale feature maps to predict the object. The YOLOv3-Tiny network consists of only the convolutional layer and the pooling layer. The network layer is shallow and the calculation is low, but the feature extraction and processing of the input image is not sufficient. In general, the deeper the network level is, the more feature information is extracted from the image, but simply increasing the network level will lead to gradient dispersion or gradient explosion.

The solution to this problem is to introduce Batch Normalization [8] into the neural network. However, it is still easy to cause new problems, resulting in the saturation or decline of the accuracy of the network during the training process and the degradation of the neural network. In order to make the neural network fully extract the image feature information and avoid the above problems, this paper introduces the idea of residual convolutional neural network [9] (ResNet) to YOLO-Tiny network. The ResNet was proposed by He Kaiming et al in 2015. It has added direct channels to the lower and upper layers, which allows shallow input information to be directly connected to the deep network, as shown in Fig. 1. If the input information is x, the output F(x) is obtained after convolution processing, and the output node is connected with the input information. This not only can protect the integrity of the input information, but also can adjust and optimize the weight of the network. According to the input and output residuals of the network in the training process, which simplifies the learning objectives and computational complexity of the network. ResNet is widely used in image detection, recognition and segmentation because it can effectively extract image features, restrain network degradation and reduce error rate.

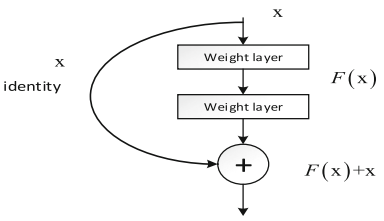


Fig. 1. Residual block

Inspired by residual convolution neural network, aimed at the problems of shallow YOLOv3-Tiny network and insufficient feature extraction, this paper introduces residual module structure to extract image features in YOLOv3-Tiny network. In addition, this paper adds another prediction layer to the network to improve the detection effect of small targets. Those networks shown in Fig. 2.

The Res-Tiny network use the same size feature map to perform object detection like YOLv3-Tiny. The experimental result shows that the detection accuracy of the Res-Tiny network is higher than YOLOv3-Tiny network 6%. In the convolutional neural network, because the feature map with larger size contains less feature details, but the location information is more abundant, this paper adds a predication layer based on the Res-Tiny network. Object detection is carried out to improve the detection effect on small targets. It is verified that the detection accuracy of the Res-Tiny-3 network for small targets is increased by more than 11% compared with the YOLOv3-Tiny network.

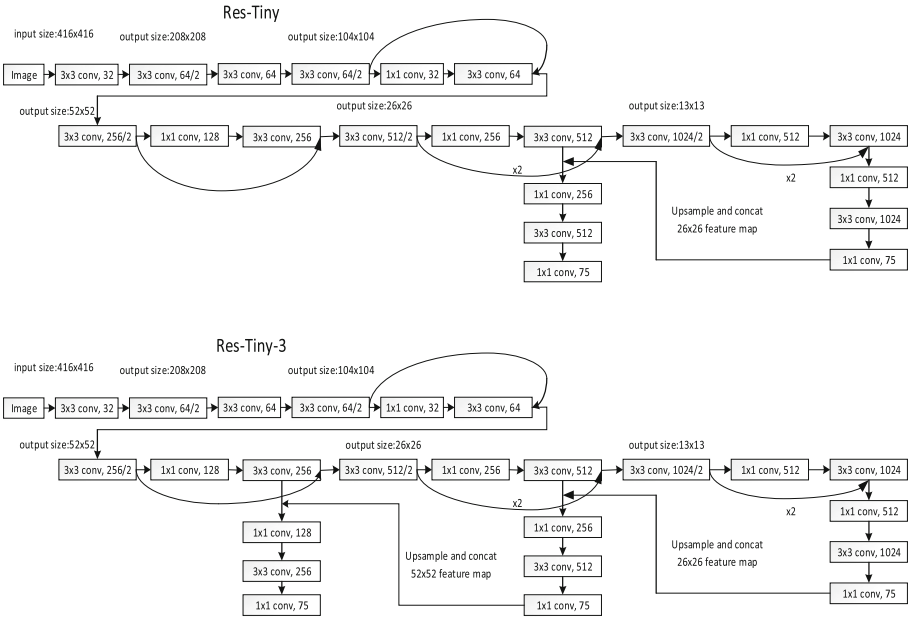


Fig. 2. Res-Tiny network and Res-Tiny-3 network

3 Evaluation Indexes

3.1 Confusion Matrix

At present, the evaluating indexes commonly used in object detection include confusion matrix, recall, precision, F1-score, Average Precision (AP), mean Average Precision (mAP) and so on. Among them, the definition of the confusion matrix is shown in Table 3.

Table 3. Confusion matrix

Truth	Prediction results	
	Positive example	Negative example
Positive example	True positives (TP)	False positives (FP)
Negative example	False negatives (FN)	True negatives (TN)

In Table 3, TP represents the number of samples that are actually positive samples and are correctly identified as positive. TN represents the number of samples that are actually negative samples and is correctly identified as negative samples. FP represents the negative sample is recognized as the number of positive samples. FN represents the number of positive samples identified as negative samples. Through the confusion

matrix, two indicators, recall rate R and precision rate P can be calculated to evaluate the effect of target detection. They are defined as follows:

$$R = \frac{TP}{TP + FN}, P = \frac{TP}{TP + FP} \tag{1}$$

The recall rate and precision rate are contradictory measures. If the recall rate is high and accuracy rate remains high, the performance of the classifier is better. In order to balance the contradiction between recall rate and precision rate, a certain value is used to evaluate the performance of the classifier, and the F1-score is derived. Usually, the bigger the value of F1-score, the better the performance of the classifier. F1 is defined as follows:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \tag{2}$$

In addition, in object detection, each image often contains multiple targets of different categories. Therefore, the mAP is often used to evaluate the detection performance of the network. The AP is the area under the PR (precision-recall) curve and mAP is the sum of the Average Precision of each type of target divided by the number of categories.

3.2 Intersection Over Union (IoU)

The intersection over union represents the overlap degree between the detected and real object positioning box in the image. The optimal intersection over union ratio is 1, which is a complete overlap. The calculation method is shown in Fig. 3.

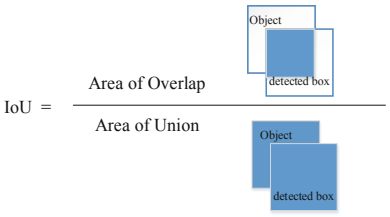


Fig. 3. Computation of intersection over union

4 Experimental Results and Analysis

First, we compare Res-Tiny and Res-Tiny-3 with other real-time detection systems on PASCAL VOC 2007. To understand the differences between Res-Tiny, Res-Tiny-3 and YOLOv3-Tiny, we explore the confusion matrix on VOC 2007 made by Res-Tiny, Res-Tiny-3 and YOLOv3-Tiny. We also present AP of each class on VOC 2007 test dataset. Finally, we show PR curve to analyze the difference between Res-Tiny-3 and YOLOv3-Tiny on small target detection.

4.1 Comparisons with Other Detection Algorithms

Many object detection algorithm focus on making detection speed fast. However, only Sadeghi [10] design algorithm that could runs in 30 frames per second (FPS). While the other efforts don't meet the real-time requirement we also compare their relative mAP and speed. YOLOv3-Tiny is the fastest object detection method on PASCAL with mAP 54.07%, it is more than twice as accurate as prior work on real-time detection. Res-Tiny pushes mAP to 60.92% while detection speed still can reach 63FPS. After add a prediction layer on Res-Tiny, the mAP increase about 4%, but it speed slower than Res-Tiny 3FPS.

Fastest DPM [11] effectively speeds up DPM without sacrificing much mAP but it still can't meet the real-time detection. It also is limited by DPM's relatively low accuracy on detection compared to neural network approaches. R-CNN minus R [12] use static bounding box proposals to locate target. Its mAP higher but slower.

Fast R-CNN [13] make the class classification faster but it still relies on selective search which can take around 2 s per image to generate bounding box proposals. Thus it has high mAP but it can't meet real-time detection. The Faster R-CNN [4] replaces selective search with region proposal networks. It most speed model achieves 18 FPS while less accurate. The Faster R-CNN VGG16 is 9 mAP higher but is also 9 times slower than Res-Tiny-3. The Zeiler Fergus Faster R-CNN [4] is 3 times slower than Res-Tiny-3 and less accurate.

Table 4. Detection accuracy and detection speed of each algorithm

Detection algorithm	Training data	mAP (%)	FPS (Frames Per Second)
100 Hz DPM [10]	2007	16.0	100
30 Hz DPM [10]	2007	26.1	30
Fastest DPM [11]	2007	30.4	15
R-CNN Minus R [12]	2007	53.5	6
Fast RCNN [13]	2007 + 2012	70.0	0.5
Faster RCNN VGG16 [4]	2007 + 2012	73.2	7
Faster RCNN ZF [4]	2007 + 2012	62.1	18
YOLO [2]	2007 + 2012	63.4	45
YOLOv2 [3]	2007 + 2012	76.8	36
YOLOv3-Tiny	2007 + 2012	**54.0**	**67**
Res-Tiny	2007 + 2012	**60.9**	**63**
Res-Tiny-3	2007 + 2012	**64.6**	**60**

From the data in Table 4, it can be seen that the previous YOLO and RCNN algorithms can't keep the detection accuracy and speed at a relatively ideal level at the same time. The rise of any one of mAP and FPS will lead to the decline of another, so it is difficult to apply in situations where the detection speed and accuracy are required to be higher. We compared with the YOLO series and RCNN series detection algorithms, the Res-Tiny-3 achieves a better balance between detection accuracy and detection speed.

4.2 Analysis of Confusion Matrix and IoU

To further illustrate the effectiveness of the proposed method, Table 5 shows the true positives (TP), false positives (FP), false negatives (FN), precision, recall, F1-score, and IoU.

Table 5. Confusion matrix parameters and IoU

Algorithm	TP	FP	FN	Precision	Recall	F1-score	IoU
YOLOv3-Tiny	5124	3229	6908	0.61	0.43	0.50	0.45
Res-Tiny	6508	3403	5524	0.66	0.54	0.59	0.50
Res-Tiny-3	7277	5466	4755	0.57	0.60	0.59	0.43

It can be seen from the data in Table 5 that the method can improve the recognition of true positives and reduce the number of detections of false negatives. After improving the basic network, the precision is increased by 5%, the recall is increased by 11%, and F1-score is also higher and more robust than the YOLOv3-Tiny network. Because the number of targets detected by the test set is huge, this paper calculates the average IoU to analyze the positioning accuracy of the algorithms. It can be seen from Table 5 that the average IoU of the Tiny algorithms is 45%. After use the residual network structure, the average IoU is increased by 5%, and the target location is more accurate. By adding the prediction layer based on the residual network, the number of detected true positives increased, but the number of false positives also increased, resulting in a 4% reduction in precision, but 17% increase in recall.

4.3 Analysis of Small Target Detection

In this paper, a layer of prediction layer is added to Res-Tiny-3 to improve the detection effect of small targets. In order to better analyze it, this paper calculates YOLOv3-Tiny network and Res-Tiny-3 AP value for each type of target is shown in Table 6.

Table 6. AP value of each type of object

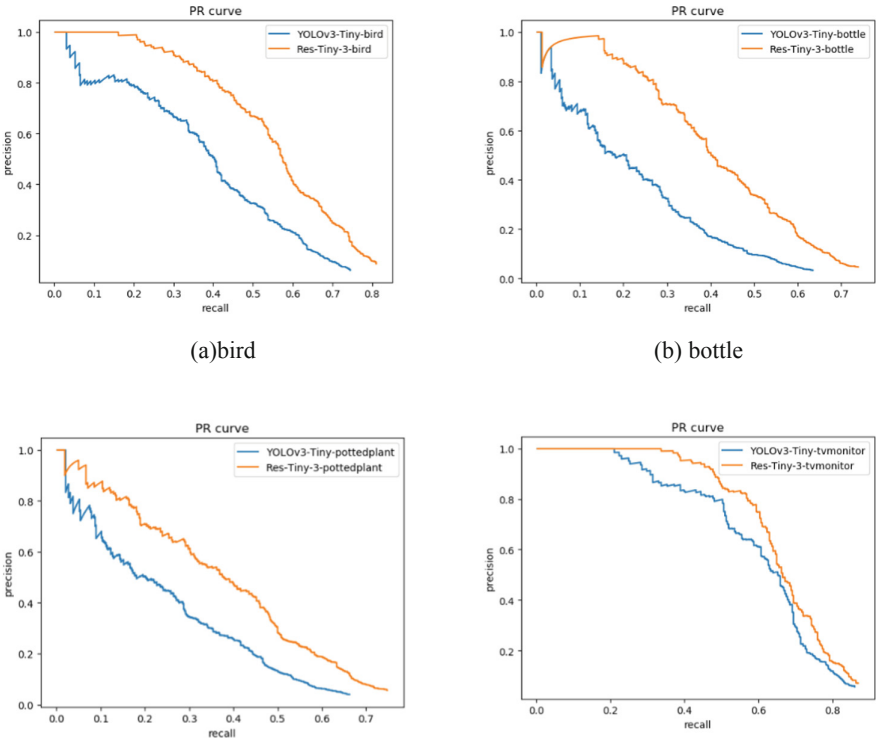
Category	Algorithm	
	YOLOv3-Tiny	Res-Tiny-3
Aeroplane	61.83%	70.90%
Bicycle	66.64%	75.17%
Bird	**40.27%**	**55.79%**
Boat	39.77%	52.92%
Bottle	**25.29%**	**41.66%**
Bus	66.80%	72.48%
Car	71.26%	79.99%
Cat	59.61%	70.84%
Chair	32.46%	46.40%
Cow	55.80%	66.49%
Table	48.07%	59.64%
Dog	53.71%	67.02%
Horse	67.41%	76.97%

(*continued*)

Table 6. (*continued*)

Category	Algorithm	
	YOLOv3-Tiny	Res-Tiny-3
Motorbike	68.76%	73.13%
Person	63.98%	75.74%
Potted-plant	**26.38%**	**37.81%**
Sheep	56.05%	65.17%
Sofa	48.43%	62.69%
Train	70.02%	77.23%
TV monitor	**58.77%**	**64.10%**

The AP data in Table 6 show that the Res-Tiny-3 network has greatly improved the accuracy of 20 categories in the Pascal VOC dataset, especially for the four types of bird, bottle, potted-plant and tv monitor. The accuracy of four small target is improved by more than 11%. It is verified that the feature map extracted by the low-level network has more position information and use this feature map make the detection of small targets more accurate. Therefore, target detection on the feature map with small receptive field can improve the detection effect on small targets. The PR curve of four categories is shown in Fig. 4.



(a)bird (b) bottle

(c) potted-plant (d) tv monitor

Fig. 4. PR curves of bird, bottle, potted-plant and tv monitor

As can be seen from Fig. 4, as the recall increases, the accuracy of detection of objects by the YOLOv3-Tiny and Res-Tiny-3 networks decreases. Among them, the YOLOv3-Tiny network curve declines faster, and in order to improve the recall rate, more precision is lost, while the Res-Tiny-3 network's PR curve declines much slower than the YOLO-Tiny network, and the recall is improved. At the same time, it can still maintain a high precision, so the Res-Tiny-3 network is better than the YOLOv3-Tiny network for detecting small objects. The detection effect of the two algorithms is shown in Fig. 5.

(a) YOLOv3-Tiny algorithm detection result

(b) Res-Tiny-3 algorithm detection result

Fig. 5. Detection results

5 Conclusion

Based on the real-time target detection algorithm YOLOv3-Tiny, this paper introduces the idea of residual convolutional neural network for extraction of image features, which can effectively prevent network degradation and gradient dispersion while deepening the network level, and enhance the network's ability to understand and analyze image feature information. In this paper, the improved neural network is trained and tested on Pascal VOC dataset. Compared with YOLOv3-Tiny network, the improved neural network has higher detection accuracy, and better detection effect on small targets. The improved algorithm is more effective and robust. The algorithm improves the detection accuracy in real-time detection, but the detection speed still has a slight decrease during the experiment. The future work includes the network structure optimization, speeding the detection.

References

1. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 511–518. IEEE, Kauai (2001)
2. Redmon, J., Divvala, S., Girshick, R.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
3. Redmon, J., Farhadi, A.: YOLO9000: better, faster, stronger. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7263–7271. IEEE, Honolulu (2017)
4. Ren, S., He, K., Girshick, R.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. **39**(6), 1137–1149 (2015)
5. Lin, T., Dollár, P., Girshick, R.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
6. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
7. Veit, A., Matera, T., Neumann, L.: Coco-text: dataset and benchmark for text detection and recognition in natural images. arXiv preprint arXiv:1601.07140 (2016)
8. Ioffe, S., Szegedy, C.: Batch normalization: accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning, pp. 448–456. JMLR.org, Lille (2001)
9. He, K., Zhang, X., Ren, S.: Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778. IEEE, Las Vegas (2016)
10. Sadeghi, M.A., Forsyth, D.: 30 Hz object detection with DPM V5. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 65–79. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10590-1_5
11. Yan, J., Lei, Z., Wen, L.: The fastest deformable part model for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2497–2504. IEEE, Columbus (2014)
12. Lenc, K., Vedaldi, A.: R-CNN minus R. arXiv preprint arXiv:1506.06981 (2015)
13. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448. IEEE (2015)