2019 IEEE 5th International Conference on Computer and Communications

Feature Pyramid SSD: Outdoor Object Detection Algorithm for Blind People

Zhigong Zhou[1 2], Xiaosong Lan[2], Shuxiao Li[2], Chengfei Zhu[2], Hongxing Chang[2*]

[1]University of Chinese Academy of Science, Beijing 100049, China

[2]Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China

{zhouzhigong2018, lanxiaosong2012, shuxiao.li, chengfei.zhu, hongxing.chang}@ia.ac.cn

Abstract—**To benefit the blind by using advanced deep learning techniques, we establish a new outdoor object detection dataset, BLIND. Different from PASCAL VOC, the scales of objects in our dataset are quite various because of different distances between objects and the camera. The characteristic of the BLIND dataset requires a high ability of scale invariance for the object detector, which classical SSD isn't adequate. We propose a novel object detector named Feature Pyramid SSD (FPSSD) focusing on BLIND, applying feature fusion strategies to classical SSD. FPSSD achieves 75.4% mean Average Precision (mAP) on BLIND, surpassing classical SSD by 1.7%. Extensive experimental results and analyses demonstrate the necessity to establish the BLIND dataset and validate the effectiveness of the proposed FPSSD object detection algorithm for the blind people.**

Keywords-deep learing; object detection; blind assistance

I. Introduction

According to a study, the number of blind people in the world is around 36 million and is set to triple by 2050 [1]. There are many restrictions for the blind or visually impaired people on outdoor activities as they can't see things clearly. It is a significant social problem. We are aiming to help them live independently.

Walking outdoors is one of the most expectations for the blind or visually impaired persons. Helping them recognize objects in front on the streets is the first problem must be solved. The complex scene and various objects make the task challenging.

Since Alex *et al* [2] achieved state-of-art on ImageNet [3] in 2012, deep learning has been widely used in computer vision. Compared to typical hand-crafted methods, deep convolutional neural networks can extract features with strong representation automatically. The performances of CNN-based object detectors on public datasets such as PASCAL VOC [4] and MS COCO [5] are much higher than traditional ones.

DNN-based models are driven by big data. In the scene of outdoor activities of the blind, the scales of objects are quite various because of different distances between objects and the camera. Datasets like PASCAL VOC [4] are dominated by large and medium objects and are not suitable for the scene we need. Thus, we establish a new outdoor object detection dataset named BLIND by imitating blind people walking outdoors.

We apply the structure of SSD [6] to our method. Previous version of SSD uses original feature maps for detection and isn't adequate for strong scale invariance. To enhance scale invariance, we propose a novel object detector named Feature Pyramid SSD (FPSSD) focusing on BLIND by applying feature fusion to classical SSD.

To summarizes, our contributions are listed as follows:

- We establish a new outdoor object detection dataset, BLIND, by imitating blind people walking outdoors.
- We propose a novel outdoor object detection method, Feature Fusion SSD, focusing on the variable scales of objects in outdoor scene.
- We validate the effectiveness of our method on BLIND and PASCAL VOC dataset with ablation studies.

The rest of the paper is organized as follows. Section II introduces traditional hand-crafted and related CNN-based object detection approaches. The proposed FPSSD network is described in section III in detail. Section IV gives experimental results and analyses. The last section is the conclusion.

II. Related Work

A. Classical Object Detectors

Early object detectors firstly generate candidate regions with different scales by applying sliding windows on an original image. Then hand-crafted features are used to represent candidates. Finally, classifiers such as SVM [7] are applied to identify categories of candidates. Classical hand-crafted features include Haar feature [8] based on Haar basis functions, SIFT [9] and HOG [10] based on block-wise orientation histograms, etc. DPM [11] mixes multi-scale deformable part models to represent highly variable object classes. However, with the development of deep learning, CNN-based detectors quickly maintain the top results on most public datasets. These approaches are roughly divided into two families, the two-stage approach with relatively high accuracy and the one-stage approach with relatively high speed.

B. Two-stage Approach

The two-stage approaches contain two steps. The first step generates a series of category-agnostic object proposals, and then recognizes object category and refines bounding boxes. R-CNN [12] uses CNN to extract features from region proposals generated by Selective Search [13], then uses SVM for classification and linear regression for localization. Fast R-CNN [14] uses neural networks not only for feature extraction but also for classification and localization. To achieve higher efficiency, Region Proposal Network (RPN) [15] replaces Selective Search [13] and shares convolutional features with the detection network. Following R-CNN family, Deformable Convolutional Networks [16], SPP [17], R-FCN [18] are proposed to further improve the performance.

978-1-7281-4743-7/19/$31.00 ©2019 IEEE

C. One-stage Approach

Compare to the two-stage approaches, the one-stage approaches pay more attention to speed. Redmon *et al* [19] present a real-time detector, called YOLO (You Only Look Once), which frames object detection as a regression problem and predicts the bounding boxes directly without region proposals. To achieve higher accuracy, YOLO9000 [20] improves YOLO by adding anchors, batch normalization and removing fully connected layers. After that, YOLOv3 [21] adds more tricks and performs better. SSD [6] uses multiple feature maps from different layers to improve performance on small objects. DSSD [22] introduces additional context into SSD via deconvolution layers.

III. Method

A. SSD Method

SSD [6] is a typical one-stage detector, using slightly adjusted VGG [23] as the backbone to extract features from images. The structure of SSD is shown in Fig. 1. To ensure scale invariance, SSD initializes prediction outputs by a set of default boxes and then refines them. Finally, NMS is used to ignore redundant boxes. SSD predicts bounding boxes on 6 feature maps with different sizes from different layers. To each feature map for prediction, SSD assigns default boxes with corresponding aspect ratios and scales respectively. Without region proposals, SSD is easy to be trained and applied to the system needing an object detection component.

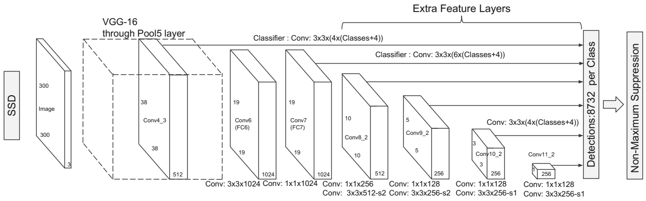

Figure 1. Structure of SSD.

B. Feature Pyramid SSD Method

Deep neural networks achieve a strong ability of representation. The deeper layer feature maps are extracted from, the larger receptive field each position in feature maps corresponds. Thus, deep features contain richer semantics than shallow ones, which benefits classification. However, distant objects are usually small in the outdoor scene. Just searching objects in original feature maps will hardly achieve satisfactory accuracy.

Shallow features keep high-resolution and are useful for object localization. Combining deep features and shallow features can generate higher-resolution and semantically strong features. It is important for keeping scale-invariant.

FPN [24] merges features from different layers by element-wise addition. Before merging, nearest neighbor interpolation and 1×1 convolutional layers are applied to reshape feature maps into the same dimension.

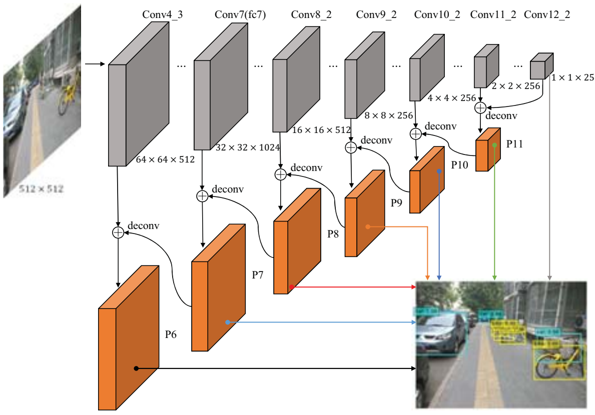

Figure 2. Illustration of the feature fusion module.

Classical SSD detects objects on multi-scale feature maps and alleviates scale variance problem to some extent. However, detecting objects on original feature maps is not enough for the high requirement for scale invariance in scenes like outdoor activities of the blind. To further enhance scale invariance, we apply FPN to SSD to predict bounding boxes on fused feature maps instead of original feature maps. We choose features from conv4_3, conv7, conv8_2, conv9_2, conv10_2, conv11_2, conv12_2 in backbone for feature fusion. We directly use conv12_2 as P12 fused feature without any operation. The structure of FPSSD is shown in Fig. 2. Note that we replace nearest neighbor interpolation by deconvolution for upsampling. More details in feature fusion are shown in Fig. 3.

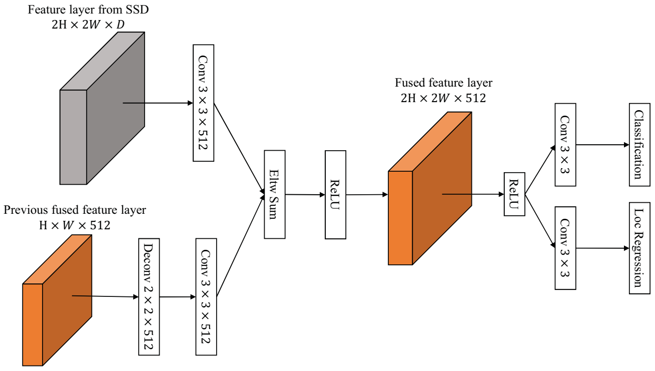

Figure 3. Illustration of the feature fusion module.

IV. Experiments and Analyses

A. Preliminaries

In the scene of outdoor activities of the blind, the scales of objects are quite various because of the law that the object is big when near and small when far. Dominated by large and medium objects, datasets like PASCAL VOC do not have such a variety of scales. Considering that there is currently no

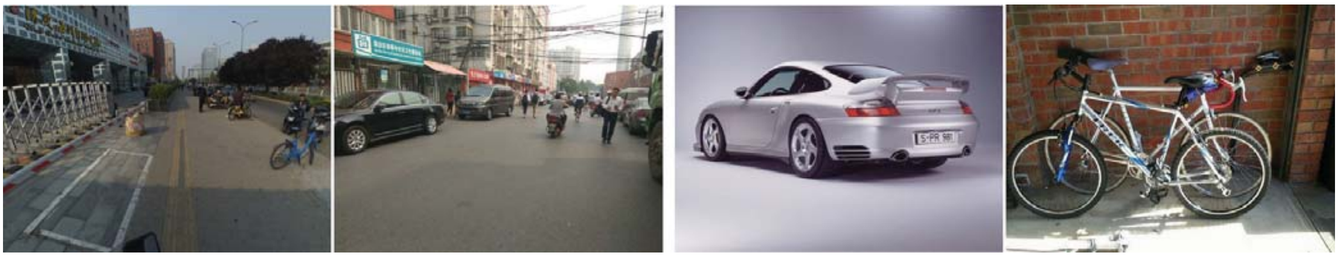

Figure 4. Data comparison between BLIND (the first and second column column) and PASCAL VOC (the third and forth column).

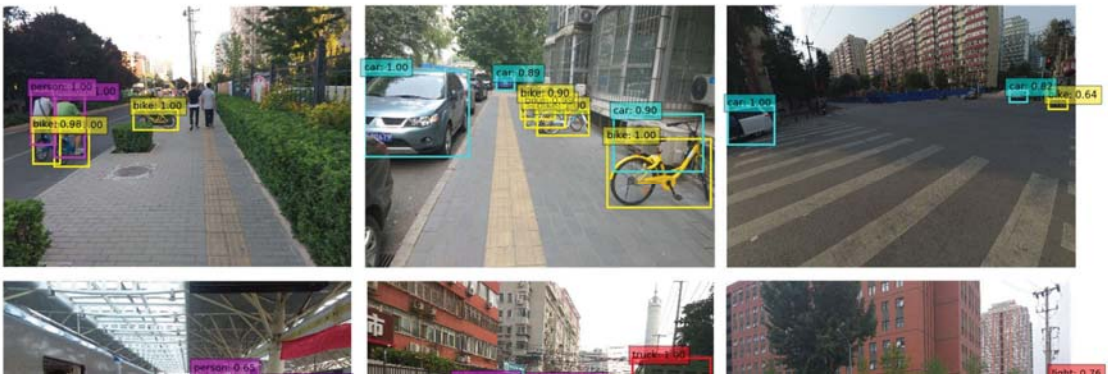

Figure 5. Test examples for Feature Pyramid SSD on BLIND.

public dataset available, we established a new image dataset named BLIND.

To collect photos, we fixed the camera on the head and walked along the blind roads. It is to recover the perspective of the blind as much as possible when outdoors. We sampled a photo every 4 seconds to avoid the contents in images are too similar, which will impact the generalization of the model trained on it. We also collected photos in traffic intersections and subway station. BLIND contains 1194 images. We select 900 images used for training and others for testing.

To satisfy the needs of blind people, seven categories of objects are labeled in images we collected, including bicycle, bus, car, motorcycle, person, truck and traffic light. Note that although most categories in BLIND also appear in PASCAL VOC, our experiment result demonstrates the data distribution in two datasets is very different and our BLIND dataset is useful and indispensable to help the blind going outside. Data comparison between two dataset are shown in Fig. 4.

We use LabelImg [25], a graphical image annotation tool, to label the bounding boxes for objects belong to the categories above. All annotations are saved as XML files in PASCAL VOC [4] format. There are 12488 target objects in BLIND totally and 1784 objects per each category in an average. All our experiments are performed on the Nvidia GTX-1080Ti GPU, Intel Core I7-6950K CPU with Caffe 1.0 [26].

B. Training details

Considering the difference in receptive fields for feature map positions, it's necessary to set the scales of default boxes specifically. The relative scales of the default boxes to each feature map Pk is calculated as:

$$s_k = s_{min} + \frac{s_{max} - s_{min}}{m-1}(k-1), \qquad k \in [1, m] \tag{1}$$

where s_{min} is 0.15 and s_{max} is 0.9. In Fig. 6, the statistic shows that the aspect ratio for target objects is roughly distributed between [0.3, 3] uniformly in BLIND. Thus we denote aspect ratio as $r \in \{1, 2, 3, 1/2, 1/3\}$ to contain all situations. In addition we also add a default box with scale $s_k = \sqrt{s_k s_{k+1}}$ and aspect ratio $r = 1$. The width and height of default boxes are computed as:

$$w_k = 512 \times s_k \times \sqrt{r} \tag{2}$$

$$h_k = 512 \times s_k \;/\; \sqrt{r} \tag{3}$$

The backbone VGG model is pretrained on the ILSVRC classification dataset [3]. We fine-tune the resulting model using SGD with 0.9 momentum, 0.0005 weight decay and batch size 16. The initial learning rate is 10^{-3} and then we decrease it by a factor of 10 after the 12 thousandth iteration and the 15 thousandth iteration. We stop training after the 16 thousandth iteration. The resolution of input images is set to 512 × 512. The same data argumentation strategies are applied as those in SSD [6]. The loss function consists of the classification confidence loss and localization loss:

$$L(x,c,l,g) = \frac{1}{N}\left(L_{conf}\left(x,c\right) + \alpha L_{loc}\left(x,l,g\right)\right) \quad (4)$$

Where N is the number of default boxes matched with the ground truth; x is an indicator for matching the default box to the ground truth box; c is the classification confidence; l is the prediction box and g is the ground truth box; the weight parameter α is set to 1.

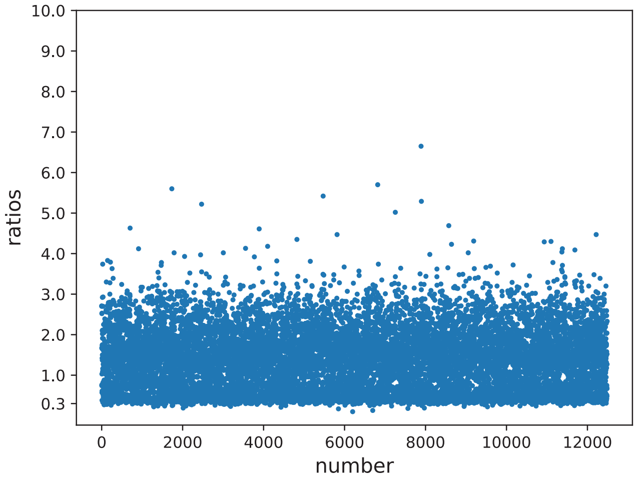

Figure 6. The statistic aspect ratio for target objects in BLIND.

C. Testing details

During the testing phase, we still set the input size to 512 × 512. Other settings are the same as SSD [6]. The bounding box predicted is correct if Intersection over union (IoU) ≥ 0.5. We use the average precision (AP) as the evaluation metrics. AP is the average precision value for recall value over 0 to 1. The mean average precision (mAP) is the mean of AP for each class. The precision value and the recall value are computed as:

$$Precision = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive} \quad (5)$$

$$Recall = \frac{True\ Positive}{True\ Positive\ +\ False\ Negative} \quad (6)$$

D. Experimental Results

To explore the similarity of data distribution between BLIND and PASCAL VOC, we test SSD trained on VOC directly on BLIND. The result is shown in Table II. The extremely poor performance demonstrates that there is a huge difference in data distribution between BLIND and PASCAL VOC.

We train FPSSD on BLIND. FPSSD achieves satisfactory performance, especially for objects closed to camera which blind people are most concerned about. Some test examples are shown in Fig. 5.

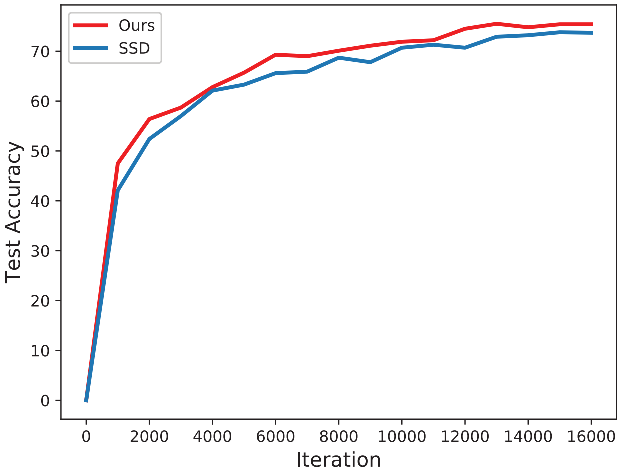

Figure 7. Test accuracy during training models on BLIND.

To verify the effectiveness of feature fusion, we compare the accuracy of FPSSD and classical SSD [6] on BLIND. Fig. 7 shows the change of test accuracy during training models. FPSSD achieves higher test accuracy than classical SSD throughout the training stage and finally s urpass classical SSD by 1.7% mAP. Table III shows the test detection result for each object category in BLIND. The result proves feature fusion enriches semantics while remaining resolution unchanged which can further enhance scale invariance and improves the accuracy.

To evaluate the generalization, we also train and test FPSSD and classical SSD on PASCAL VOC [4]. The train set is the union of VOC2007 and VOC2012 trainval. The test set is VOC2007 test. Both two models are trained 120 thousand iterations with the same hyperparameter. The initial learning rate 0.001 and is decreased by a factor of 10 at the 80 thousandth iteration and the 100 thousandth iteration. The batch size is 32.

The test detection accuracy is shown in Table I. FPSSD surpasses classical SSD by 1.2% mAP and is more accurate in most object categories. Our method has a satisfactory generalization and can be applied to other datasets.

V. Conclusion

In this paper, we have established a new outdoor object detection dataset, BLIND. Focusing on the variable scales of objects in outdoor scene, we propose an outdoor object detection method, Feature fusion SSD, aiming at helping blind people. Extensive experiments demonstrate the effectiveness of our method. Otherwise, we test our method on PASCAL VOC

TABLE I
Test Detection Result On Pascal VOC

Method	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
SSD	84.9	85.8	80.8	73.1	58.0	87.8	88.4	87.6	63.6	85.5	73.2	86.3	87.7	83.8	82.7	55.3	81.6	79.2	86.5	80.3	79.6
FPSSD	86.8	87.0	81.0	75.6	60.3	88.6	88.7	88.4	62.9	87.3	76.1	86.9	88.5	87.5	82.6	56.7	84.0	79.4	88.0	80.2	80.8

TABLE II
Test Detection Result On Blind Of SSD Trained By VOC

Dataset	bicycle	bus	car	motor cycle	person	truck	traffic light	mAP
VOC	42.1	13.6	57.7	33.8	38.5	-	-	26.5
BLIND	76.7	78.4	80.8	77.3	75.6	74.2	53.2	73.7

TABLE III
Test Detection Result On Blind

Method	bicycle	bus	car	motor cycle	person	truck	traffic light	mAP
SSD	76.7	78.4	80.8	77.3	75.6	74.2	53.2	73.7
FPSSD	79.6	79.6	82.9	79.1	77.6	75.8	53.0	75.4

and verify its generalization. In the future, we will expand our dataset and improve the speed of our method.

Acknowledgement

This work is supported by the National Science Foundation of China (NSFC) with granting No. 61573350.

References

[1] "The number of blind people in the world is set to triple by 2050," https://www.thejournal.ie/blindness-triple-2050-3527437-Aug2017/.

[2] I. Sutskever, G. E. Hinton, and A. Krizhevsky, "Imagenet classification with deep convolutional neural networks," *Advances in neural informa-tion processing systems*, pp. 1097–1105, 2012.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[4] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[5] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[7] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[8] P. Viola, M. Jones *et al.*, "Rapid object detection using a boosted cascade of simple features," *CVPR (1)* vol. 1, no. 511-518, p. 3, 2001.

[9] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[10] C. G. Harris, M. Stephens et al., "A combined corner and edge detector." in Alvey vision conference, vol. 15, no. 50. Citeseer, 1988, pp. 10–5244.

[11] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE transactions on pattern analysis and machine intelligence, vol. 32, no. 9, pp. 1627–1645, 2009.

[12] 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.

[13] J. R. R. Uijlings, K. E. A. V. D. Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[14] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[16] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 764–773.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[18] J. Dai, Y. Li, K. He, and J. Sun, "R-fcn: Object detection via region-based fully convolutional networks," in *Advances in neural information processing systems*, 2016, pp. 379–387.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[20] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[21] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[22] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "Dssd: Deconvolutional single shot detector," *arXiv preprint arXiv:1701.06659*, 2017.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] 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.

[25] "labelimg," https://github.com/tzutalin/labelImg.

[26] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.