Light Deep Learning based Edge Safety Surveillance

Yimo Lou*, Wengang Cao, Zhimin He, and Guan Gui*

College of Telecommunication and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing, 210003 China

*Email: 1219012829@njupt.edu.cn, guiguan@njupt.edu.cn

Abstract. Safety is considered as the first important factor in many industries such as construction sites. Hence, artificial intelligence based safety surveillance techniques have been received strong attentions in recent years. Conventional surveillance systems for monitoring whether the workers wearing helmets are not easy to install and carry, and the largest trouble is that the system needs considerable computation, which is not that simple to satisfy the requirement of hardware. Considering the characteristic about construction sites, in this paper, we proposed a new system based on CenterNet with MobileNet-V2 as backbone. It has a video camera, a marginal device embedded with Jetson TX2 and wireless communication routers to ensure real-time transmission about live-scene about construction sites. After inspection, the light-weight network we proposed can be run in portable marginal device smoothly and stably with slight loss of average precision.

Keywords: Intelligence safety surveillance, CenterNet, MobileNet-V2, marginal devices

1 Introduction

With the development of society, to improve environment of every aspect about human beings, new buildings spring up like mushrooms. Meanwhile, accidents happen in construction sites become a serious problem that threatens the progress of whole project. Moreover, consequences caused by accidents lead to inestimable loss to both enterprises and families of staff who are injured or dead. Thus ensuring safety of staff is significant in construction activities. From the article, we found about construction accidents in china during 2014 to 2018, 3014 accidents happened, causing nearly 4000 mortalities. More than half of these accidents (52.2%) due to falling from high place, other reasons are hitting by objects (15.2%), crane injury (7.5%), collapse (7.3%), mechanical injury (5.9%) and others (11.9%). In common sense, most of them could be avoided or less influential by wearing safety equipment, such as helmets. Hence, supervision about whether workers wear helmets is a meaningful direction to protect staff's lives.

Traditional measures to deal with this problem cannot be deployed without human, which are waste of manpower and money. Besides, supervision by humankind has flaws that may not have solution to make up. Since majority of people are just flesh and blood without super power, the most serious problem in artificial supervision is that mortal horizon is limited. There are blind spots everywhere in construction sites, not mentioned the field operation carried out vertically, arranging supervisors at every area or height is unrealistic and a total waste. If used, not only does this method cost a large amount of money, as the number of workers increasing for supervision, but also rise the probability of potential danger.

Therefore, automatic supervision about whether workers wearing helmets turns into a meaningful topic. Bolstered by rapid development of artificial intelligence (AI), we can use this method to deal with

EAI MOBIMEDIA 2020, August 27-28, Harbin, People's Republic of China

DOI 10.4108/eai.27-8-2020.2295046

problem above. However, advanced method via AI generally need considerable computation, which is not easy to realize by Simple portable equipment that can be used at construction sites. Besides, although the network detects helmets by frame, what is shows on screen is video composed by frames, the frequency of it must be ensured, not mentioned that it is a real-time detection. What's more, owing to instability of sites, equipment with low cost to avoid extra loss due to damage is more suitable. Transferring the whole network to a marginal device with tiny size needs to be considered either. To sum up, a light-weight network is needed.

In this paper, we propose a new method about helmet detecting with lighter construction. We use CenterNet [1] which has MobileNet-V2 [2] as backbone. Meanwhile we transfer this network into marginal devices. By putting external device, cameras, at place we want to monitor, accent color boxes will appear in screen to show whether workers wearing helmets.

2 Related Works

Nowadays, machine learning has integrated into every aspect of our life. G. Gui, M. Liu et al. proposed A series of method based on aviation big data and machine learning [3-4]. For some research about communication on channels and module classification, H. Huang et al. [5-12] use deep learning to improve original solution to get better result, not to mention the most popular topic, 5G/6G [13].

The whole method can be transformed into an object detection problem based on deep learning. One algorithm tried the most direct method [14], C. Wojek et al. proposed an algorithm using sliding windows to detect object from left to right then up to down by classification. In order to detect different class from different distance, it utilizes disparate length-width ratio. Unlike sliding window, Region Convolutional Neural Network (R-CNN) proposed by R. Girshick et al. [15] obtain region of interest (ROI) by region proposals. It considers each pixel as a group, and then calculate their texture and combine two regions, which are most close to each other, until all of them combined. To upgrade the accuracy, R-CNN needs a fat lot of candidate region, large part of them are overlap. Thus Fast Region Convolutional Neural Network (Fast R-CNN) [16] proposed by the same team, which uses CNN to extract features of whole image rather than extract multiple times on every image block. Upgraded version of Fast R-CNN is Faster R-CNN [17], which let the network learn which candidate region of image are by itself instead of using regular algorithm, it has higher efficiency. Above are two-stage object detection algorithm. Liu et al. proposed Single-Shot MultiBox Detector (SSD) [18], which uses deeper layers in convolutional network to detect object. Due to decrease of Spatial dimensions and resolution caused by convolutional layer, regular structure of it can detect larger object only while execute independent detection from multiple feature maps could solve this problem. You only look once (YOLO) proposed by J. Redmon et al. [19-21] detects features by using DarkNet after convolution layer. It smooths feature maps and then unites with another feature map with lower resolution.

Departed object detection algorithm, which can be used in this condition, are insufficient to satisfy the need of light-weight we want. Thus we choose CenterNet with MobileNet-V2 as our ideal solution, which we will introduce in next part.

3 The Proposed Method

The whole system is based on Internet of Things (IoT). The data acquisition layer is a camera

connected to marginal embedded device, power provided by Jetson TX2, and transmission layer utilizes router to data acquisition layer with server. The application layer is a screen to show the result.

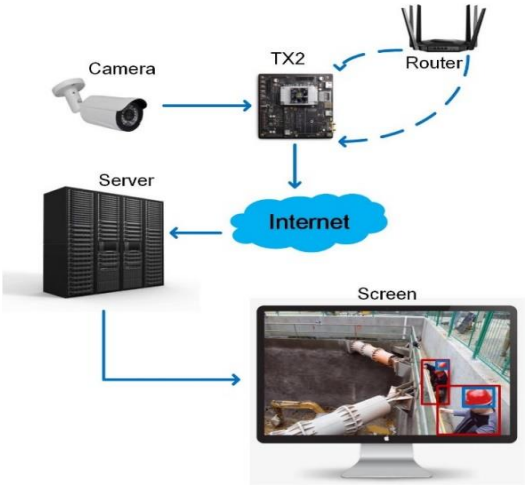

Fig. 1. Basic framework of the proposed edge safety surveillance.

3.1 CenterNet

Object detection needs to make a rectangular box containing target with minimal size. A successful detector will list a huge amount of candidate boxes and classify them, which is low effective and a total waste, subsequent disposal cannot be avoided neither. CenterNet proposed by X. Zhou et al. has a brand new thought that we only need to locate the object as a point, which is the center point of detection box. By using heat map, detector will find its center point and then regress other attribute of this object, such as size, coordinate of 3D location, direction, posture etc.

CenterNet is an end-to-end fast object detector. We use the center points of human and helmet to represent them, then the only anchor boxes shaped as size of objective are regressed directly from the peak in heat map (one anchor for one objective, cancel Non-maximum suppression (NMS), a quiet time-consuming process). Therefore, the detection and alignment of these two objective become a standard key point estimation.

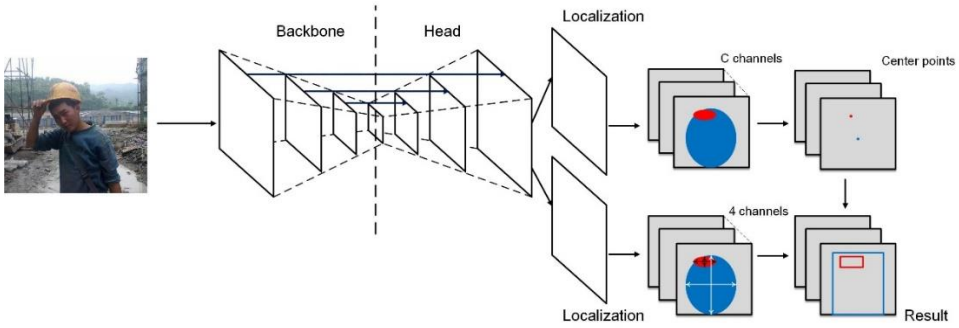

Fig. 2. The structure of CenterNet used in our proposed method, we take MobileNet-V2 as backbone.

Owing to adopting fully convolutional network, CenterNet could get heat map with larger resolution (output stride of 4, a fourfold reduction compared to traditional ones) without setting anchor boxes in advance. Channels of heat map has the same quantity with classifications of objects need to detect. The

net takes first 100 peaks of heat map as center points and set a threshold to get final point.

There is deformable convolution before every upsampling to make receptive field of network become more precise instead of being limited into rectangular convolutional box with size of 3×3. Meanwhile, feature maps under downsampling, which has output stride of four possesses higher resolution than normal network.

Intersection over Union (IoU) is a criterion to measuring accuracy of standard anchor based detection, such as CenterNet. There is a threshold value, when IoU bigger than that, the anchor counts as positive to any object.

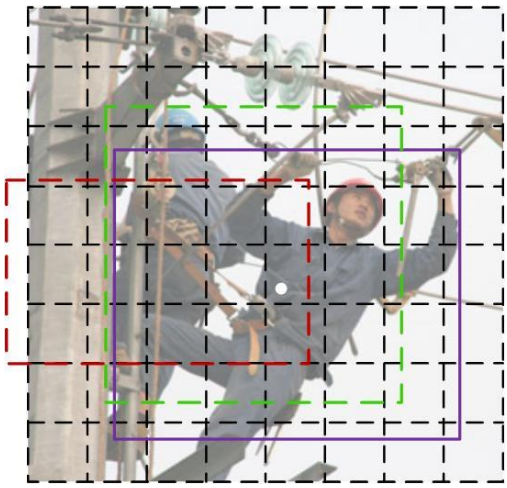

Fig. 3. Sketch map of IoU, where purple frame is the ground truth, when $\mathrm{IoU} = 0.7$, anchor green is positive, while anchor red is negative.

the improved loss function of classification in CenterNet in our method is as follows:

$$L_k = \frac{-1}{N} \sum_{xyc} \begin{cases} L_c \left(1 - \hat{Y}_{xyc}\right)^{\alpha} \log\left(\hat{Y}_{xyc}\right) & if \;\; Y_{xyc} = 1 \\ L_c \left(1 - Y_{xyc}\right)^{\beta} \left(\hat{Y}_{xyc}\right)^{\alpha} \log\left(1 - \hat{Y}_{xyc}\right) & otherwise \end{cases} \tag{1}$$

Due to images in datasets are not been fully labeled, we set a new parameter L_c, which represents whether the class has been labeled, $L_c = 1$ means labeled already while $L_c = 0$ means unlabeled. N is number of center points in image I, α and β are hyper parameters of Focal Loss, $\hat{Y}_{xyc}$ is predictive value of object.

3.2 MobileNet-V2

MobileNet proposed by M. Sandler et al. is a lightweight CNN network focused on mobile terminal or embedded device. Consider that operating in portable marginal devices, we take MobileNet-V2 as backbone and Feature Pyramid Network (FPN) [22] as neck of CenterNet. FPN can make a tradeoff between speed and accuracy to obtain more robust and semantic information.

MobileNet-V2 improves depthwise separable convolution [23], which separates the two steps to depthwise convolution and pointwise convolution. First, do bitwise multiply by channels, while keep quantity of channels the same; second, do traditional convolution with kernel, which size is 1×1, quantity of channels can be changed. Now computation reduce to

$$D_K * D_K * M * D_F * D_F + 1 * 1 * M * N * D_F * D_F \quad (2)$$

where D_F is the size of feature map, D_K is the size of convolutional kernel, M is quantity of input channels, N is the quantity of output channel. We set kernel size to 3, reducing computation by 8 to 9 times with slight accuracy loss compared with traditional convolution.

The batch normalization layer and ReLU6 after convolution, increasing the nonlinear variation and enhancing the generalization ability of model. However, ReLU6 consumes convolutional kernel in depthwise convolution, thus MonileNet-V2 replace ReLU6 to Linear activation function. Meanwhile, the accuracy is higher than former one.

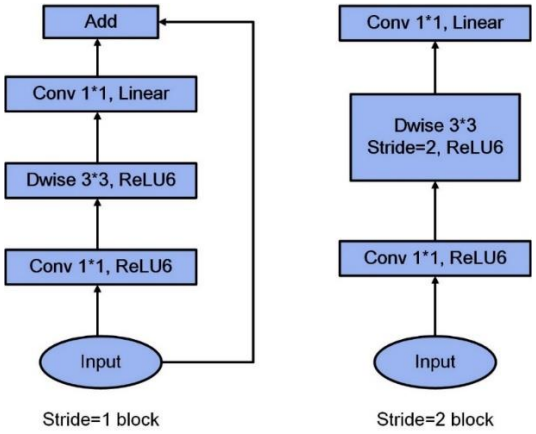

Fig. 4. The basic convolution unit structure of MobileNet-V2.

3.3 Other parts

Above are related algorithm used on detection, now we introduce other parts of this system. Video camera connected with marginal embedded AI device powered by NVIDIA Jetson TX2, provides real-time scene of construction sites to monitor. Jetson TX2 adopts 256-core NVIDIA Pascal GPU and 8 GB memory, providing higher computing speed and stronger inferential capability. Router provided the function of wireless long-distance data transmission. We use OpenCV to play the result on screen.

4 Experiment

4.1 Dataset

Data collection in our project includes three kinds: 1) pedestrian pictures from pedestrian labeled pictures in coco2017 (pedestrians labeled only, more than 60000); 2) photos taken by ourselves (both pedestrians and helmets labeled, 9000); 3) helmet pictures from Internet (helmets labeled only, more than 10000).

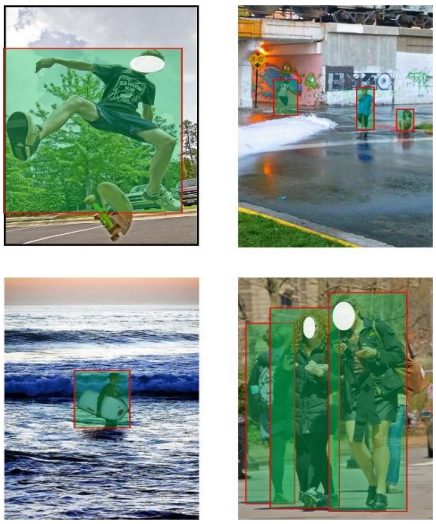

Fig. 5. Four samples of Dataset 1.

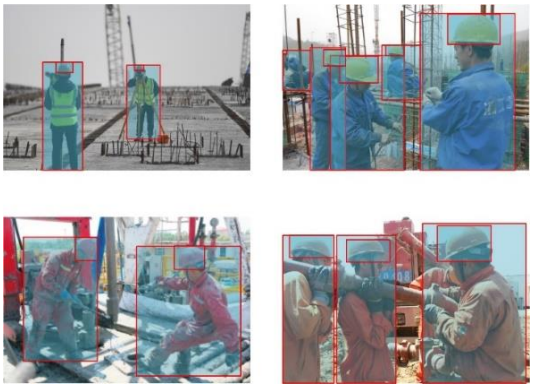

Fig. 6. Four samples of Dataset 2.

Fig. 7. Four samples of Dataset 3.

4.2 Training

We rescale the size of images into 512×512. The learning rate was initialized at $1e-3$, and was reduced by a factor of 10 at the end of 80, 120, 103 epochs.

4.3 Result

4.3.1 Different epochs and IoUs

We test our model with different epochs and IoU, try to find best combination. The results about average accuracy are as follow:

Table1. Average precision under epoch = 44

	IoU=0.30	IoU=0.50	IoU=0.70	IoU=0.75
Helmet	91.44	77.44	41.34	29.86
Pedestrian	78.88	70.62	43.66	33.16

Table2. Average precision under epoch = 55

	IoU=0.30	IoU=0.50	IoU=0.70
Helmet	93.04	79.42	42.44
Pedestrian	79.91	71.46	45.05

Table3. Average precision under epoch = 140

	IoU=0.30	IoU=0.50	IoU=0.70
Helmet	92.04	79.40	44.86
Pedestrian	81.42	73.92	49.80

We can see from the result that when epoch up to 140, the average accuracy becomes lower, and it is a total waste of time. If we set the IoU too high, we are goanna miss some information which lead accidents happen. To sum up, we choose to set $\text{IoU} = 0.5$ and $\text{epoch} = 55$.

4.3.2 Different backbones

To show the high efficiency of MobileNet-V2, we compared our method with other algorithms about CenterNet that have other network as backbone (DLA-34, DLA34-V0, ResNet-18). All the tests are conducted under TensorRT infer optimizer. (In this section, fp32 means 32-bit floating point, fp16 means 16-bit floating point)

DLA [24]: Deep Layer Aggregation (DLA), an image classification network with hierarchical skip connections, makes the model become more precise and have smaller quantity of parameters, providing a measure for generalization for deep visualization framework and effective expansion of application.

ResNet [25]: Residual Network (ResNet) is a feature-extracting network based on CNN, which can simplify the training of networks that have deeper structure than those we used previously.

Table 4. Inference Time tested on NVIDIA GeForce GTX 1080Ti

Backbone	Mode (precision)	Inference Time (ms)
MobileNet-V2	fp32	3.798
DLA-34	fp32	24
DLA34-V0	fp32	12.6
ResNet-18	fp32	5.81
MobileNet-V2	int8	1.75
DLA-34	int8	19.6
DLA34-V0	int8	6.76
ResNet-18	int8	3.63

Table 5. Inference Time tested on NVIDIA Jetson TX2

Backbone	Mode(precision)	Inference Time (ms)
MobileNet-V2	fp16	22
DLA-34	fp16	186
DLA34-V0	fp16	80
ResNet-18	fp16	41

We can know from the table that no matter what condition, MobileNet-V2 always has the best performance, which means it has the highest frame frequency.

4.3.3 Final test on real scenes

The final test is on videos we find about construction sites, which shows the result of detection directly.

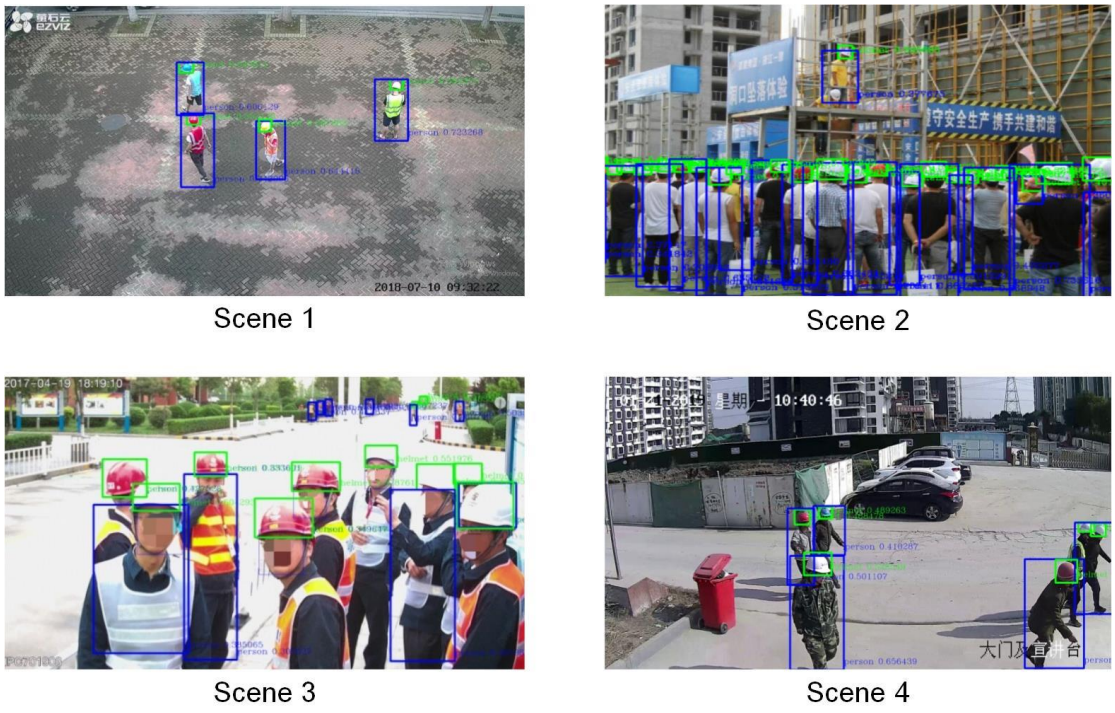

Fig. 8. Detection result on different scenes.

5 Conclusions

In this paper, we proposed a light-weight supervision system for construction sites which can be run on marginal devices. This method used MobileNet-V2 as backbone embedded into CenterNet to realize a faster and more precise detection. With Jetson TX2, the system upgrades in stability and convenience. In the future, we plan to do further research on improving average precision, because it still has slight decrease about that than those normal weight networks. Also, we are going to add new function to supervise other accident may happen in sites, to make this system more complete.

References

[1] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points", CoRR, vol. abs/1904.07850, 2019.

[2] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks" in 2018 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 2018, pp. 4510-4520

[3] G. Gui, F. Liu, J. Sun, J. Yang, Z. Zhou, and D. Zhao, "Flight delay prediction based on aviation big data and machine learning," IEEE *Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1065–1069, 2020.

[4] G. Gui, Z. Zhou, J. Wang, F. Liu, and J. Sun, "Machine learning aided air traffic flow analysis based on aviation big data," *IEEE Trans. Veh. Technol.*, in press, doi: 10.1109/TVT.2020.2981959

[5] H. Huang, Y. Peng, J. Yang, W. Xia, and G. Gui, "Fast beamforming design via deep learning," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1065–1069, 2020.

[6] Y. Wang, J. Yang, M. Liu, and G. Gui, "LightAMC: Lightweight automatic modulation classification using deep learning and compressive sensing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3491–3495, 2020.

[7] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sept. 2018.

[8] Y.Wang, J. Gui, Y. Yin, J.Wang, J. Sun, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "Automatic modulation classification for MIMO systems via deep learning and zero-forcing equalization," *IEEE Trans. Veh. Technol.*, doi: 10.1109/TVT.2020.2981995

[9] Y. Wang, J. Wang, W. Zhang, J. Yang and G. Gui, "Deep learning-based cooperative automatic modulation classification method for MIMO systems," *IEEE Trans. Veh. Technol.*, doi: 10.1109/TVT.2020.2976942

[10] H. Huang, Y. Song, J. Yang, and G. Gui, "Deep-learning-based millimeter-wave massive MIMO for hybrid precoding," *IEEE Trans. Veh. Technol.*, vol. 68, no. 3, pp. 3027–3032, Mar. 2019.

[11] Y. Wang, M. Liu, J. Yang, and G. Gui, "Data-driven deep learning for automatic modulation recognition in cognitive radios," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 4074–4077, Apr. 2019.

[12] H. Huang, J. Yang, H. Huang, Y. Song and G. Gui, "Deep learning for super-resolution channel estimation and DOA estimation based massive MIMO system," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8549–8560, 2018.

[13] H. Huang, S. Guo, G. Gui, Z. Yang, J. Zhang, H. Sari and F. Adachi, "Deep learning for physical-layer 5G wireless techniques: Opportunities, challenges and solutions," IEEE Wireless Commun. Mag., vol. 27, no. 1, pp. 214–222, 2020.

[14] C. Wojek, G. Dorkó, A. Schulz, B. Schiele, "Sliding-windows for rapid object class localization: A parallel technique," *DAGM 2008 on Pattern Recognition,* vol. 5096, pp. 71–81, 2008

[15] R. Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation", in 2014 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 580-587

[16] R. Girshick, "Fast R-CNN," in 2015 *IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440-1448.

[17] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[18] Liu, Wei, et al., "SSD: Single shot multibox detector," in 2016 *European conference on computer vision*. 2016, pp. 21-37.

[19] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time dbject detection," in 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-

788.

[20] J. Redmon, and A. Farhadi, “YOLO9000: Better, Faster, Stronger,” in 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517-6525.

[21] J. Redmon, and A. Farhadi, “Yolov3: An incremental improvement,” CoRR, vol. abs/1804.02767, 2018.

[22] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie. *Feature pyramid networks for object detection*, in 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[23] A. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, “MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications,” CoRR, vol. abs/1704.04861, 2017.

[24] F. Yu, D. Wang, E. Shelhamer, and T. Darrell. “Deep layer aggregation,” in 2018 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 2403-2412.

[25] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition,” in 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770-778.