Development of an AI-based System for Automatic Detection and Recognition of Weapons in Surveillance Videos

Shenghao Xu
School of Science and Technology
The Open University of Hong Kong
Hong Kong, China
s1211896@ouhk.edu.hk

Kevin Hung
School of Science and Technology
The Open University of Hong Kong
Hong Kong, China
khung@ouhk.edu.hk

Abstract – Security cameras and video surveillance systems have become important infrastructures for ensuring safety and security of the general public. However, the detection of high-risk situations through these systems are still performed manually in many cities. The lack of manpower in the security sector and limited performance of human may result in undetected dangers or delay in detecting threats, posing risks for the public. In response, various parties have developed real-time and automated solutions for identifying risks based on surveillance videos. The aim of this work is to develop a low-cost, efficient, and artificial intelligence-based solution for the real-time detection and recognition of weapons in surveillance videos under different scenarios. The system was developed based on Tensorflow and preliminarily tested with a 294-second video which showed 7 weapons within 5 categories, including handgun, shotgun, automatic rifle, sniper rifle, and submachine gun. At the intersection over union (IoU) value of 0.50 and 0.75, the system achieved a precision of 0.8524 and 0.7006, respectively.

Keywords – surveillance video, security camera, artificial intelligence, weapon detection, TensorFlow, Single Shot MultiBox Detector

I. Introduction

Nowadays, video surveillance systems have become an integral part our daily living, ensuring safety and security of the general public. Security cameras already have a high penetration rate in Hong Kong. They have been installed in public places such as strategic locations in transport infrastructure, public housing estates, shopping areas, immigration and customs areas at the borders, and entrances to public facilities and lift cabins [1]. Despite the widespread use of such systems, their routine operations and detection of high-risk situations are still performed manually by security personnel. Having 294,000 qualified security services personnel, Hong Kong is still facing the challenge of 5% manpower shortage in the sector [2]. The above situations will intensify, which may lead to increasing cases of undetected or delay in detection of threats in public areas. Noting these challenges, various parties have developed automated and artificial intelligence (AI)-based solutions with the focus in detection and recognition of dangerous items in surveillance videos. Unlike humans who are affected by emotions and subjective judgement, AI has the advantage of higher accuracy, consistency, and speed in the detection of threats [3]. It will also provide alleviate the burden of manpower shortage and high operating cost.

The aim of this work is to develop a low-cost, efficient, and AI-based solution for the real-time detection and recognition of weapons in surveillance videos under different scenarios. The system was developed based on TensorFlow, which is an open-source platform for machine learning; the Single Shot MultiBox Detector (SSD), a popular object detection algorithm; and MobileNet, which is a convolution neural network (CNN) for producing high-level features. This was a collaborative project with a security solution provider in Hong Kong.

II. Literature Review

A. Review of Existing Products

Athena Security AI System is designed for detecting mass shooting threats. With computer vision, it can detect and recognize 900 types of firearms within three seconds. It has been evaluated to have achieved an accuracy rate ranging from 88.9% to 100%, depending on the scenario [4]. Once threat is detected, the system automatically issues a warning to the suspected perpetrator, and sends the real-time video feed to security staff. The system developed by ZeroEyes is compatible with legacy security cameras. It also automatic performs facial redaction to ensure privacy [5]. Its algorithm learns adaptively, so that its performance would enhance over time. GunLockers and SafeSchool are guns and threat detection systems from Virtual eForce. It makes use of both video and sound sources for detection, and additionally provides tracking and perimeter lockdown functions [6]. The abovementioned solutions are AI-based products in the market. In the current work, the aim is to develop a system which is low-cost, open-source, and easily reconfigurable.

TABLE 1. Comparison between TensorFlow and similar platforms.

Platform	Language	GPU	Pre-trained?	Developer
TensorFlow	Python, JAVA, C++	CUDA	Yes	Google
Caffe	C++, Python	CUDA, OpenCL	Yes	BVLC
Keras	Python	CUDA	Yes	F.chollet
Torch7	Lua	CUDA,	Yes	Facebook
Theano	Python	CUDA, OpenCL	Yes	MILA

B. Review of TensorFlow

TensorFlow is an open-course software library and framework developed by the Google Brain team for machine learning applications. Based on dataflow and differentiable programming, it is widely used in many applications [7]. TensorFlow has been chosen for the development in the current work due to its advantages of i.) expandability with self-developed superstructures and libraries; ii.) portability with support of a wide range of platforms and processors; iii.)

978-1-7281-5033-8/20/$31.00 ©2020 IEEE

support for visualization and multiple programming languages, including Python, C++, and C; and iv.) comprehensive libraries for deep learning. Table 1 gives a comparison between TensorFlow and other similar systems [8].

III. System Overview

Figure 1 shows the functional blocks of the proposed system. After the video is captured by the surveillance camera, it is passed to the key frame extraction subsystem, which reduces data size by selecting key frames for feasible real-time running of the subsequence steps. The extracted frames are then input into the weapon detection algorithm. The detected weapons are classified and labeled. Figure 2 illustrates details of the system's operation flow.

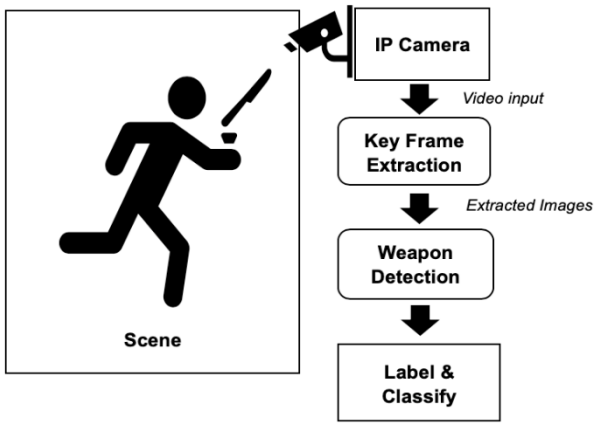

Fig. 1. Functional blocks of the proposed system.

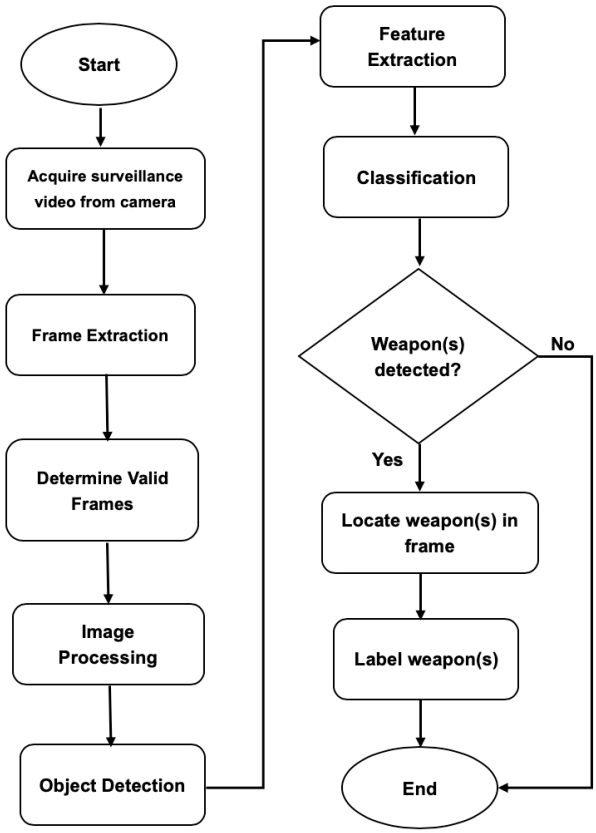

Fig. 2. Flow of operations in the proposed system.

The image processing is mainly RGB to grayscale conversion. The processed images were input into the SSD-MobileNet model, which consisted of a Single Shot MultiBox Detector (SSD) and MobileNet light weight deep neural networks. The CNN in MobileNet performs feature extraction at different scales in the convolutional layers of different scales. With the extracted features, the SSD algorithm was used to obtain the feature information to locate and classify the detected object(s) [9]. After classification, the system finally labels the detected weapons within the frames. Table 2 summarizes the software environment used for developing the system's functions.

TABLE 2. Software environment for system development

Operating System	Microsoft Windows 10 (64-bit)
Python	3.7.3
TensorFlow	1.14.0
Graphics Drive	Radeon Software Adrenalin 19.20
Conda	4.7.12

IV. Methodologies

A. Key Frame Extraction

Within a typical surveillance video, most of the frames are identical due to the fixed location and background. Therefore, it would be inefficient use of computing resources if all frames are processed. As such, the proposed system extracts key frames from the video based on the ceoncept of interframe differences as follows [10]. Consider frames k and k-1 of a video sequence as f_k and f_{k-1}, The grayscale pixel values of the corresponding frames at location (x,y) are denoted as $f_k(x,y)$ and $f_{k-1}(x,y)$. Then the difference image $D_k(x,y)$ is obtained by taking the absolute value of the difference between the two pixel values, as shown in Equation (1). Then each D_k is then quantized to a binarized foreground image pixel $R_k(x,y)$ according to a predefined threshold T_1, as shown in Equation (2). If the D_k element is less than T_1, it corresponds to the background, otherwise it is considered as foreground, i.e., a moving object.

$$D_K(x,y) = |f_k(x,y) - f_{k-1}(x,y)| \quad (1)$$

$$R_k(x,y) = \begin{cases} 0, D_K(x,y) < T_1 \\ 1, D_K(x,y) \geq T_1 \end{cases} \quad (2)$$

In the proposed system, smoothing average followed by a local maximum algorithm based on interframe difference are used to extract the key frames. The frames which correspond to local maxima of average interframe difference are considered as key frames.

B. Weapons Detection and Recognition

The neural network in the system was trained via supervised classification learning in two steps: i.) a weapons dataset containing 1218 machine gun images was extracted from the COCO dataset; and ii.) using the weapons dataset, the weapons detector was trained by fine-tuning the pretrained model. The COCO dataset project is a large visual database for visual object recognition software research [11]. The fine-tuning resulted in a more efficient model that could detect more weapons efficiently. The fine-tuning processes involved the preparation work which included i.) key frame extraction;

ii.) manually labelling the training images; iii.) generating XML, CSV, and TFRecord files for describing the objects in the images, iv.) setting configuration file for the model, as shown in Table 3, and v) training the model with the TFRecord files and key frames.

TABLE 3. Initial Configuration Parameter Settings

Pretrained model	ssd_moblienet_v1_coco
Num_classes	5
Batch_size	5
Matched_threshold	0.5
Max_detections_per_class	100
initial_learning_rate	0.004

V. Experiment and Results

Processing in the system was performed at a desktop computer with Intel Core i3-9100F 3.60GHz 4 Core processor and GPU of AMD Radeon RX 560 4GB. For key frame extraction, trials with window sizes of 10, 20, and 30 were performed. Table 4 shows the execution time of key frame extraction for the different settings.

TABLE 4. Execution Time for Different Window Size

Window Size (Hanning Window)	**Execution Time** (sec)
len_window = 10	48.61696243286133
len_window = 20	46.70806813240051
len_window = 30	46.64224433898926

Figures 3a to 3c show the mean inter-frame difference intensity achieved for different window sizes. It can be seen that smaller window size results in stronger and less change in inter-frame difference intensity. This would correspond to the cases of having similar frames extracted as key frames. A larger window size would result in less key frames. The larger the number of key fames, the slower the longer the processing time. Based on the experiments, window size of 30 resulted in the most efficient key frame extraction. After key frame extraction, the images were used to train the model via supervised classification. As shown in Figures 4 to 6, as the number of training steps increased, the total loss, classification loss, and localization loss gradually decreased. Localization and classification became prominently more accurate at training step of 10k and beyond. Precision and recall were used to measure the model's performance [12].

Figure 7 shows the mean average precision (mAP) after 50,000 training iterations. As the number of iterations increased, the model's mAP increased gradually with small fluctuation, and achieved the highest value of 0.5431. Figures 8 and 9 show the change of average precision of the model when the intersection over union (IoU) was 0.5 and IoU is 0.75. When IoU = 0.5, the AP (average precision) reached 0.8524 and in strict metric condition when IoU = 0.75 the AP was 0.7006. Figure 10 shows how the recall value increased with number of iterations. It reached 0.6216 after model training was completed. When the model completed the training, it was able to detect and recognize weapons from the images, as shown in the examples in Figures 11 and 12.

VI. Conclusion

A low-cost, efficient, and artificial intelligence-based solution for the real-time detection and recognition of weapons in surveillance videos has been developed based on Tensorflow. The system was preliminarily tested with a 294-second video which showed 7 weapons within 5 categories, including handgun, shotgun, automatic rifle, sniper rifle, and submachine gun. At the intersection over union (IoU) value of 0.50 and 0.75, the system achieved a precision of 0.8524 and 0.7006, respectively.

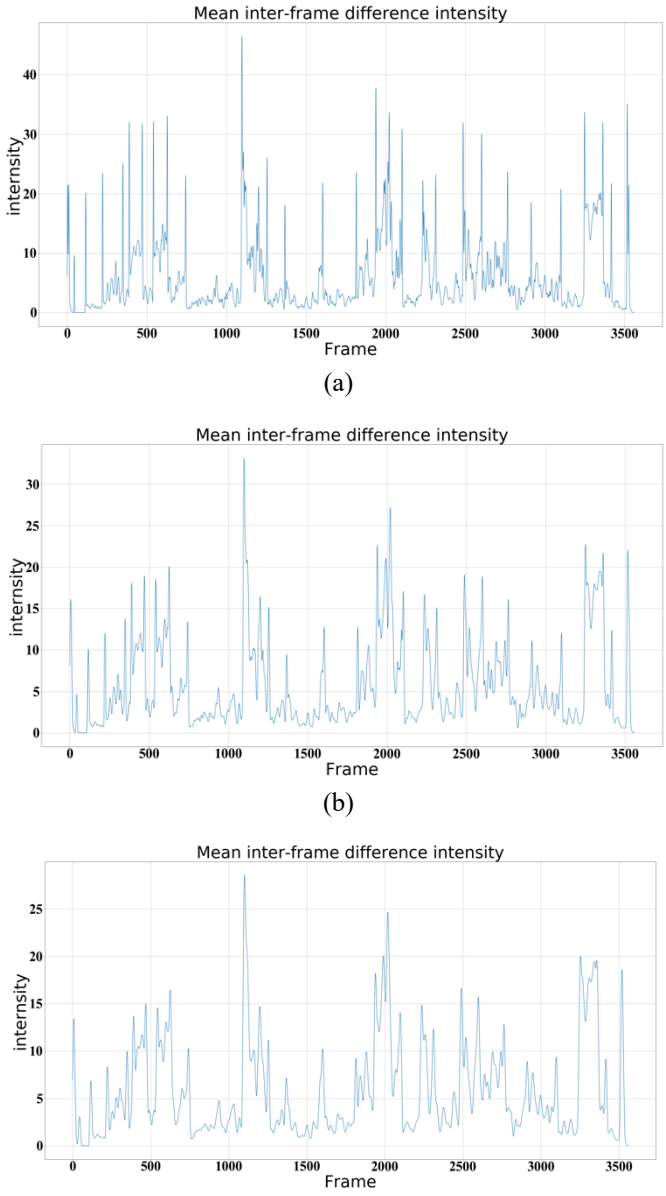

Fig. 3. Mean inter-frame difference intensity for setting the window length at (a) 10, (a) 20 and (c) 30.

Future research will focus on two aspects. First, more training datasets and test dataset will be used to improve the system's detection performance. 1275 images of automatic firearm from ImageNet will be used for that purpose. System performance will then be compared with that trained with the COCO dataset [13]. Second, two modes of operation will be developed: energy-saving mode and a high-performance mode. In high-performance mode, system will detect videos in real time and frame-by-frame. This mode will be used in densely populated or high-risk venues, providing a higher level of security. In energy-saving mode, the system only processes extracted key frames that correspond to changes. It will be suitable for scenarios with low human traffic.

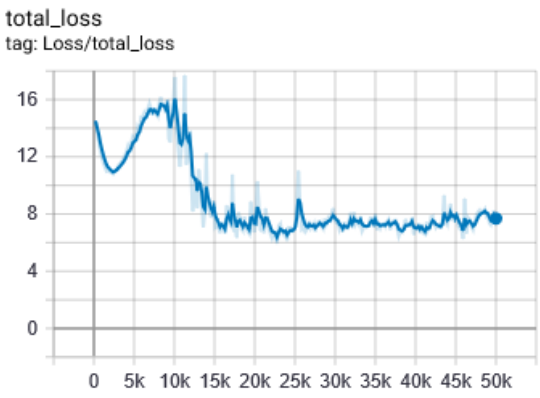

Fig. 4. Total loss function of the model vs. training steps.

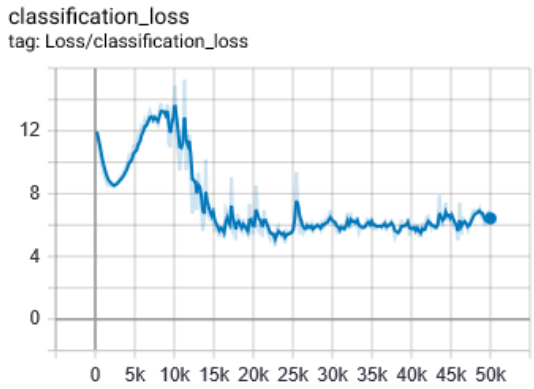

Fig. 5. Classification loss of the model vs. training steps.

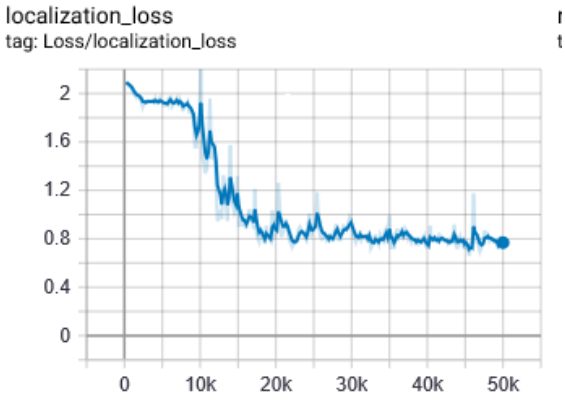

Fig. 6. Localization loss of the model vs. training steps.

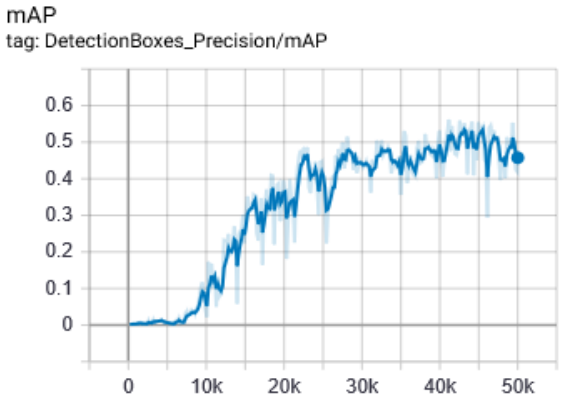

Fig. 7. mAP (mean Average Precision) of the model vs. training steps.

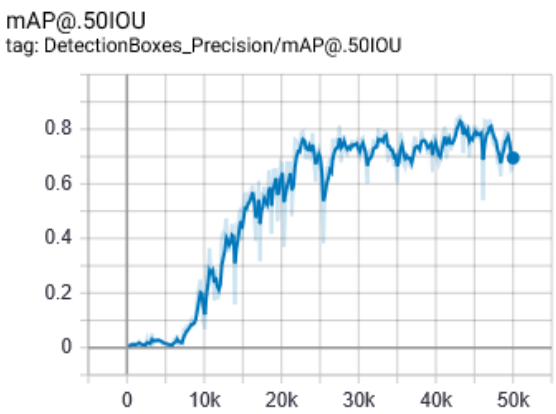

Fig. 8. mAP of the model at IoU=0.5.

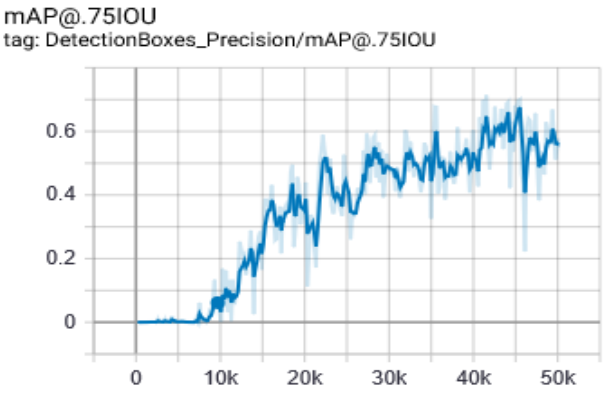

Fig. 9. mAP of the model at IoU=0.75

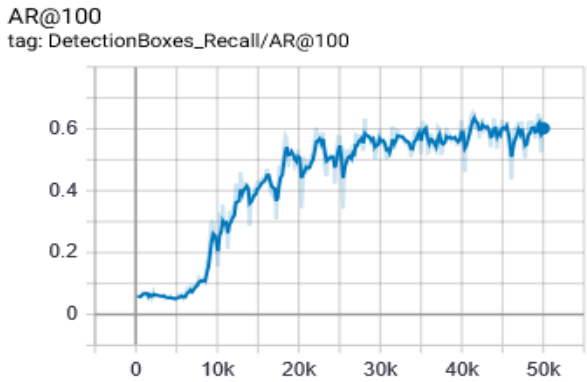

Fig. 10. AR (average recall) given 100 detections per image

Fig. 11. Output of the weapons detection when more than one weapon is embedded in the frame.

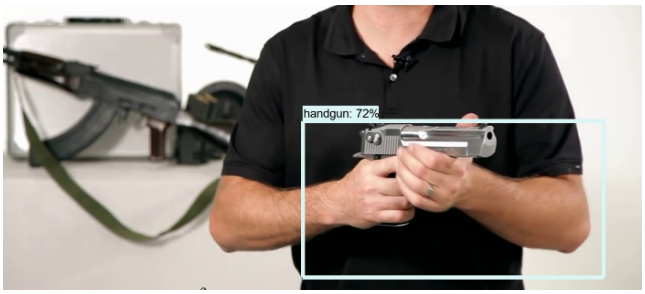

Fig. 12. Output of the weapons detection when only one weapon is embedded in the frame.

ACKNOWLEDGMENT

The authors would like to thank Integrated Corporation and their R&D team for providing the technical advices and collaboration opportunity in this project.

REFERENCES

[1] Secretary for Security, *Replies to LegCo Questions LCQ20: Locsed Circuit Television Cameras*, Hong Kong Special Administrative Region Government, Feb. 3, 2010. [Online]. Available: https://www.info.gov.hk/gia/general/201002/03/P201002030201.htm. [Accessed: Feb. 10, 2020].

[2] Security Services Industry Training Advsiory Committee, *Security Services Industry: Specification of Competency Standards – Version 1*, Hong Kong Special Administrative Region Governemnt, Dec. 2017. [Online]. http://www.hkqf.gov.hk. [Accessed; Feb. 10, 2020]

[3] D. Schuller and B.W. Schuller, "The Age of Artificial Emotional Intelligence," *IEEE Computer*, vol.51, iss 9, pp. 38-46, September 2018.

[4] Athena Security, Inc. *Athena Security Corp White Paper: A Study of Gun Detection Accuracy in the Athena Security Artificial Intelligence System.* [Online]. Available: https://athena-security.com/athena-security-gun-detection-white-paper.pdf. [Accessed: Feb 10, 2020].

[5] ZeroEyes. *School Security* [Online]. Available: https://zeroeyes.com/school-security. [Accessed: Feb. 10, 2020].

[6] Virtual eForce. *AI Based Gun Detection, Shooter Tracking & Mass Notification*. [Online]. Available: https://www.virtualeforce.com/. [Accessed: Feb. 10, 2020].

[7] T. Mikolov, M Karafiat, L Burget, J Cernocky, S. Khudanpur, "Recurrent Neural Network Based Language Model", in Proc. 11th Annual Conference of the International Speech Communication Association, INTERSPEECH 2010, Makuhari, Chiba, Japan, September 26-20, 2010. pp. 1045-1048.

[8] Z. Zeng, Q. Gong and J. Zhang, "CNN Model Design of Gesture Recognition Based on Tensorflow Framework," 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chengdu, China, 2019, pp. 1062-1067.

[9] A. Heredia and G. Barros-Gavilanes, "Video processing inside embedded devices using SSD-Mobilenet to count mobility actors," 2019 IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI), Barranquilla, Colombia, 2019, pp. 1-6.

[10] L. He and and L. Ge, "CamShift Target Tracking Based on the Combination of Inter-frame Difference and Background Difference," in Proc. 2018 37th Chinese Control Conference, CCC, Wuhan, China, July 25-27, 2018. pp. 9461-9465.

[11] COCO Dataset. COCO Explorer [Online]. Available: http://cocodataset.org/#explore. [Accessed: Feb 10, 2020].

[12] S. Chauhan, *Understanding Mean Average Precision for Object Detection (with Python Code)*, Medium. Jun. 28, 2019. [Online]. Available: https://medium.com/@jonathan_hui/map-mean-average-precision-for-object-detection-45c121a31173. [Accessed: Feb. 10, 2020].

[13] K. He, R. Girshick and P. Dollar, "Rethinking ImageNet Pre-Training," 2019 IEEE CVF International Conference on Computer Vision (ICCV), Seoul, Korea (South), 2019, pp. 4917-4926.