# An AI Powered Computer Vision Application for Airport CCTV Users

Mehmet Cemal Atlıoğlu[1,*], Gökhan Koç[2]
[1,2] *R&D Department of TAV Technologies, Istanbul, Turkey*

***Abstract*— Investments in aviation were experiencing difficult times due to the Covid-19 pandemic, and poverty drives the industry to generate value with existing products. Therefore, technology providers modernize legacy systems with AI add-ons like the usage of old CCTV cameras for securities operations even if they are not designed for these purposes [1]. In this study, the detection of objects such as people, luggage, and vehicles are executed and tested for the aviation ecosystem with a real-time computer vision application built on existing CCTV cameras. Also, the detection performance measurements and achievements of the application are shared.**

***Keywords*— artificial intelligence, machine learning, deep learning, human detection, airport, security, CNN, CCTV**

## I. INTRODUCTION

With the onset of the Covid-19 epidemic, airports and airlines quickly began to conduct various researches on safe travel methods. As a result of these researches, measuring the airport passengers' social distance with existing CCTV cameras and preventing violations is considered an important solution by many airports [1]. In this study, following this trend, a computer vision application has been developed to work on the airport CCTV camera images in real-time by using open source libraries defined by literature review and enables the detection of objects such as people, luggage, and vehicles. Also, the detection performance measurements and achievements of the application are shared.

According to the research, the number of published articles related to object detection increased more than 20 times in 2018 compared to 1998 [2]. When the 20-year development period of Computer Vision-based object detection is examined, it is seen that traditional methods like Viola-Jones Detector, Histogram of Oriented Gradients, and Deformable Part-based Model have been developed and used during the first 10 years [3-6]. However, with the rebirth of the convolutional neural network in the last 8-10 years, the widespread use of deep learning techniques has increased [7][8]. Object detection, classification, and tracking on videos have become extremely easy today, with numerous algorithms born and developed with this wind of revolution. After the literature review, it is seen that in the past 10 years, many successful algorithms, such as CNN, RCNN, SSPNet, Fast RCNN, Faster RCNN, YOLO, SSD, FPN, Retina-Net, RefineDet, TridentNet, have been developed and used by deep learning developers and experts [9-16]. Also, open-source frameworks like TensorFlow, PyTorch, and Deeplearning4j have a great impact on the rapid implementation of these developments from the academic world to the business.

Manuscript received June 28, 2021; accepted November 25, 2021.
*Corresponding author: mehmetcemal.atlioglu@tav.aero

Deep learning techniques, which meet the complex business requirements at a high level in detection and classification issues, naturally require very high computational power levels [17], but this is not the only challenge architects struggle with. They are also struggling with difficulties such as using a neural network for different purposes, dividing and managing the GPU power for different blocks, performing distributed training with the multi-GPU to shorten the training process, and eliminating CPU / GPU incompatibilities [18]. In addition to dealing with architectural difficulties in projects, it is also necessary to work on business-related functions that have not yet been resolved strongly in the literature. For example, abnormal behavior detections and classifications of crowds [19][20], abandoned luggage detection [21], depth-related calculations on RGB cameras are some of the common business-level challenges that developers are facing [22][23].

In this study, details of the application development processes, algorithms, business-level test results, library comparisons, and configurations of the developed application are shared in other sections.

## II. METHODOLOGY

In the evolution of the state-of-the-art object detectors, 3 main architectures were mostly used. As the earliest approach, the “Classical Detectors” emerged as basic FIR filters, which applies a predefined function kernel, which is used to extract the target object features over image pixels where the operation is basically called a convolution process. The most known CNN architecture “LeNet” was proposed by LeCun et al. [24] to solve the handwritten digit recognition problem during his researches at Bell labs.

The successors of the classical detectors have outperformed the convolution-based detection approach with the rise of deep learning techniques, which lead to the birth of the “Two-State Detectors”. As the main difference from the classical approach, the two-state detectors operate on region-based proposals. Region proposal methods focus on a sparse set of candidate proposal regions where the target objects may lie. Thus they eliminate negative object locations to simplify the search space. As the next step, the classifiers are operated on these sparse regions to detect the objects. The (R-CNN) algorithm [25] can be regarded as one

of the most used two-stage detector algorithms where more generalized extensions such as (Faster RCNN) [12] based variants are still used in low-scale edge devices for object detection tasks.

However, another latest object detection algorithm family operates very similarly to human eye biology by executing only a single detection stage step. The algorithms that fall into this family are called “Single State Detectors”. Today three main single state detection algorithms can be listed as SSD (single shot detector) [14], YOLO (you only look once) [13], and RetinaNet [15]. Among three algorithms, YOLO, which is implemented on the Darknet framework [26], is the fastest with acceptable near accuracy compared to (Faster RCNN) [27][28]. The YOLO algorithm's main advantage is that a single neural network evaluates the whole image rather than the proposed regions where the two-stage detection algorithm does.

The CNN algorithm considered in this paper has been built on an open-source platform complied from the Darknet framework [26] which is written in C, with CUDA, CuDNN and OpenCV bindings that implement the YOLO architecture of the 4th version. Tthe DLL of the C libraries has ben encapsulated by the managed .Net wrappers so that algorithm can be called from the main application on-premise manner. Video capturing and CCTV stream capturing features has been implemented on managed OpenCV wrapper EmguCV libraries. The camera enumeration process has been implemented with the opensource Directshow library.

## III. TEST AND RESULTS

With the application developed, detection and classification performance measurements were carried out on many different types of videos. Some of the videos had very distant images, so the videos' classification range limit was defined and performance analyses were calculated for only defined areas. Precision, Recall, F1 score, and IOU (Intersection over Union) calculations were made for the videos' model outputs. Also, all images used in the study have 720p resolution.

a) Test Video 1 - Road

b) Test Video 2 - Road

c) Test Video 3 - Road

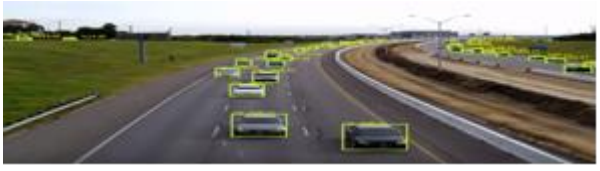

Fig. 1. Test videos for road

The first tests were carried out on 3 different road videos, which can also be seen in Fig. 1. In these tests, the model was expected to detect and classify vehicle types (truck, car, motorbike) and people. When the videos are examined, it is seen that the camera also displays very distant vehicles in Test Video 1 and Test Video 3. In classification analysis, performances within a 100-meter range were measured in these videos, and instantaneous faulty detections were ignored.

As seen in Table I, the model could not detect 2 humans in Test Video 1 because they were 90 meters far from the camera. Also, for the same video, vehicle classification failures of model happened with the vehicles which are too far from the camera. Moreover, another finding in parallel with this result is it has been seen that in the same video that the confidence level for nearby vehicles is mostly exceeding 90%, but for far vehicles, it could be lower than 60%.

TABLE. I. CLASSIFICATION PERFORMANCE OF MODEL ON DIFFERENT ROAD VIDEOS

| **Classification Performances** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Cars | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| | Test Video 1 | 33 | 0 | 1 | 3 | 0,97 | 0,92 | 0,94 |
| | Test Video 2 | 53 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| | Test Video 3 | 111 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| | Sum of All | 197 | 0 | 1 | 3 | 0,99 | 0,99 | 0,99 |
| Trucks | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| | Test Video 1 | 5 | 0 | 0 | 1 | 1,00 | 0,83 | 0,91 |
| | Test Video 2 | 3 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| | Test Video 3 | 2 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| | Sum of All | 10 | 0 | 0 | 1 | 1,00 | 0,91 | 0,95 |
| Motorbike | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| | Test Video 1 | 2 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| | Test Video 2 | 3 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| | Test Video 3 | 3 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| | Sum of All | 8 | 0 | 0 | 0 | 1,00 | 1,00 | 1,00 |
| Humans | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| | Test Video 1 | 1 | 0 | 2 | 0 | 0,33 | 1,00 | 0,50 |
| | Test Video 2 | 71 | 0 | 6 | 3 | 0,92 | 0,96 | 0,94 |
| | Test Video 3 | 3 | 0 | 0 | 1 | 1,00 | 0,75 | 0,86 |
| | Sum of All | 75 | 0 | 8 | 4 | 0,90 | 0,95 | 0,93 |

When the results of Test Video 2 were examined, it is seen that all buses were classified correctly, but there was no bus in other videos, so this finding is not included in comparisons. The range and camera position enabled people sitting in the front seats of the vehicles to be detected. However, the confidence level for human detection was mostly measured below 60%, while the confidence level for vehicle detection was above 90%. The reasons are that model could only see half of the person's bodies, and there are distorting effect of the light reflections on the vehicle windows.

In Test Video 3, only the people on the motorcycles were detected because the people inside the vehicles were not visible. The analysis was carried out according to that condition.

When the detection and classification performance results shared in Table I for 3 videos of the model are examined, it is seen that the model is successful in vehicle classification with a 100-meter viewing range and 720p camera resolution. However, it is successful in detecting people at shorter distances. In the other tests of this study, as seen in Fig. 2, more detailed inferences were obtained from the videos specific to human classification.

Confidence levels dropped to 30% in Test Video 4. The classification of people who are away from the camera proved to be much weaker. In Test Video 5, the confidence level was generally seen above 90% due to the camera's distance to people and people walking on a clear background. In the last video about Human Walking, the confidence level was mostly calculated as 97% and above in the classification of those passing between the camera and the parked vehicle.

a) Test Video 4 – Human Walking

b) Test Video 5 – Human Walking

c) Test Video 6 – Human Walking

Fig. 2. Test videos for human walking

As shown in Table II, when the model's joint performance in all 3 videos is evaluated together, the F1 score value becomes 0.83. According to the range of the camera and resolution, this value can be up to 0.98.

TABLE. II. CLASSIFICATION PERFORMANCE OF MODEL ON DIFFERENT HUMAN WALKING VIDEOS

| **Classification Performances** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| Humans | Test Video 4 | 192 | 0 | 97 | 16 | 0,66 | 0,92 | 0,77 |
| | Test Video 5 | 56 | 0 | 2 | 0 | 0,97 | 1,00 | 0,98 |
| | Test Video 6 | 34 | 0 | 3 | 0 | 0,92 | 1,00 | 0,96 |
| | Sum of All | 282 | 0 | 102 | 16 | 0,73 | 0,95 | 0,83 |

Within the scope of another test, the videos displayed in Fig. 3 and the classification tests of luggage such as suitcase, backpack, and handbag were performed.

When the results in Table III are examined, it is seen that the model sometimes detects backpacks as handbags in Test Video 7.

In Test Video 8, it was determined that the model could not classify the suitcases with covers on them. Besides, since there are only suitcases in this video, the classification result related to other objects is not shared.

There was a higher success in classifying covered suitcases in Test Video 9 than in Test Video 8, but the confidence level ranges from 20% to 30% for such suitcases. Also, there were no handbags in the video, so the results of classification for that object not shared in Table III.

a) Test Video 7 - Luggages

b) Test Video 8 - Luggages

c) Test Video 9 - Luggages

d) Test Video 10 - Luggages

Fig. 3. Test videos for luggage classification

The model sometimes classifies backpacks as a suitcase while backpacks are on the carousel in the Test Video 9. In Test Video 10, detection and classification performances were calculated only for luggage close to the camera.

TABLE. III. CLASSIFICATION PERFORMANCE OF MODEL ON DIFFERENT LUGGAGE VIDEOS

| **Classification Performances** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Handbag | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| | Test Video 7 | 12 | 0 | 7 | 4 | 0,63 | 0,75 | 0,6 |
| | Test Video 10 | 6 | 0 | 8 | 0 | 0,43 | 1,00 | 0,6 |
| | Sum of All | 18 | 0 | 15 | 4 | 0,55 | 0,82 | 0,6 |
| Suitcase | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| | Test Video 7 | 13 | 0 | 7 | 0 | 0,65 | 1,00 | 0,7 |
| | Test Video 8 | 27 | 0 | 2 | 2 | 0,93 | 0,93 | 0,9 |
| | Test Video 9 | 60 | 0 | 2 | 8 | 0,97 | 0,88 | 0,9 |
| | Test Video 10 | 5 | 0 | 3 | 0 | 0,63 | 1,00 | 0,7 |
| | Sum of All | 10 | 0 | 14 | 10 | 0,88 | 0,91 | 0,9 |
| Backpack | *Videos* | *TP* | *TN* | *FP* | *FN* | *Precision* | *Recall* | *F1* |
| | Test Video 7 | 6 | 0 | 7 | 0 | 0,46 | 1,00 | 0,6 |
| | Test Video 9 | 7 | 0 | 6 | 1 | 0,54 | 0,88 | 0,6 |
| | Test Video 10 | 7 | 0 | 2 | 1 | 0,78 | 0,88 | 0,8 |
| | Sum of All | 20 | 0 | 15 | 2 | 0,57 | 0,91 | 0,7 |

When the results of the 4 tests shared in Table III are examined together, the model's classification success does not seem as high as vehicle or human classification, although the camera is close to the related objects. It can only be said that suitcases are a little easier to classify than backpacks and handbags.

a) Test Video 11 – Aircrafts

b) Test Video 12 – Aircrafts

Fig. 4. Test videos for aircraft classification

TABLE. IV. IOU PERFORMANCES OF MODEL ON DIFFERENT TEST VIDEOS

| **IoU Performances of Model** | | |
|---|---|---|
| *Videos* | *Types* | *Average IoU* |
| Test Video 1 | Road | 0,89 |
| Test Video 2 | Road | 0,93 |
| Test Video 3 | Road | 0,94 |
| Test Video 4 | Human Walking | 0,81 |
| Test Video 5 | Human Walking | 0,92 |
| Test Video 6 | Human Walking | 0,95 |
| Test Video 7 | Luggages | 0,95 |
| Test Video 8 | Luggages | 0,96 |
| Test Video 9 | Luggages | 0,93 |
| Test Video 10 | Luggages | 0,92 |
| Test Video 11 | Aircrafts | 0,97 |
| Test Video 12 | Aircrafts | 0,64 |

The model was able to classify all 124 aircraft displayed in Test Video 11 and 12 shown in Fig. 1, in cases where they did not pass consecutively. For example, among the aircraft in line in Test Video 12, only the aircraft seen in front can be classified.

As important as measuring classification performance is the measurement of Intersection over Union (IOU) metrics for the model outputs. The calculated averages of IOU for each test video are shared in Table IV, but some points where the minimum intersection is seen are also shared in Fig. 5.

It is seen that the larger the size of the objects in the pictures, the better the IOU ratio. Besides, it was determined that the IOU ratio for the objects positioned one after the other easily decreased. However, the model easily rose up above 0.9 in many test videos.

a) Test Video 6 – Human Walking

b) Test Video 2 - Road


c) Test Video 9 - Luggages

d) Test Video 12 – Aircrafts

e) Test Video 11 – Aircrafts

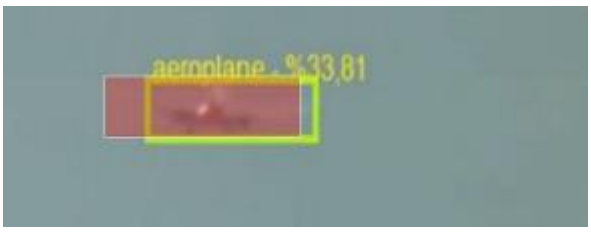


Fig. 5. Some of segmentation errors

## IV. CONCLUSION

Various findings have been reached due to the tests performed on 13 different 720p resolution videos focused on more than 700 objects.The vehicle classification success of the model within the range of 100 meters seems high. If the camera is at a sufficient distance from the passing vehicles, the people inside the vehicles and motorbikes can also be classified with approximately 60% confidence level. However, it has also been observed that the reflections on the windows of the vehicles can affect these results.

In the classification of people, model performance has not been as high as the model's vehicle classification performance. While the distance between the camera and humans is less than about 20 meters and under defined conditions, the F1 score varies between 0.77 and 1 in human classification. Also, it is seen that model fails to classify humans when they are standing behind each other. Besides, if the ground on which people are walking is homogeneous, model classification performance increases. The classification performance scores calculated for the luggage class were smaller than the other object class classifications mentioned above. It has been observed that the model can mix backpack and handbag if they are on walking people. Also, suitcases and backpacks can be mixed on the carousels. Another finding is that sometimes model fails in the classification of covered suitcases. When the suitcase, backpack, and handbag are compared among themselves, it is seen that the model is more successful in the suitcase classifications.

Finally, the model produces highly successful results in aircraft classification, but if the aircraft is positioned one after the other in front of the camera, the model cannot detect the aircraft behind and draws the segmentation border incorrectly. When the IuO values are examined in general, if the model's objects are not sequenced consecutively, the model achieves similar success in segmentation for all objects. However, as the objects move far away from the camera, segmentation errors increase. In other words, the confidence level of classified objects increases as the objects get closer to the camera.

This study's results, which determine the model's success and failure points, show that the relevant model can be evaluated in many business scenarios. Various possible studies can be done to increase the performance of the model. In addition to the transfer learning, with the images taken from the field where the model will be used, the model's business-specific training can increase the performance of the model. It will also be helpful to increase the number of nodes in the model. In addition to the model modifications, the camera equipment positioning will also produce useful results. It would be beneficial to position the cameras as close as possible to the objects. Also, the camera's height should be considered carefully to prevent displaying objects in front of each other.

## ACKNOWLEDGMENT

Part of this work was presented orally at the IV. International Conference on Data Science and Applications 2021 (ICONDATA'21).

## REFERENCES

[1] Milne, R. J., Delcea, C., Cotfas, L. A., & Ioanăș, C. (2020). Evaluation of boarding methods adapted for social distancing when using apron buses. IEEE Access, 8, 151650-151667.

[2] Zou, Z., Shi, Z., Guo, Y., & Ye, J. (2019). Object detection in 20 years: A survey. arXiv preprint arXiv:1905.05055.

[3] Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008, June). A discriminatively trained, multiscale, deformable part model. In 2008 IEEE conference on computer vision and pattern recognition (pp. 1-8). IEEE.

[4] Viola, P., & Jones, M. (2001, December). 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. CVPR 2001 (Vol. 1, pp. I-I). IEEE.

[5] Viola, P., & Jones, M. J. (2004). Robust real-time face detection. International journal of computer vision, 57(2), 137-154.

[6] Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In 2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05) (Vol. 1, pp. 886-893). Ieee.

[7] Girshick, R. B. (2012). From rigid templates to grammars: Object detection with structured models. Chicago, IL, USA: University of Chicago, Division of the Physical Sciences, Department of Computer Science.

[8] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[9] 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).

[10] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2015). Region-based convolutional networks for accurate object detection and segmentation. IEEE transactions on pattern analysis and machine intelligence, 38(1), 142-158.

[11] 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.

[12] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. arXiv preprint arXiv:1506.01497.

[13] 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).

[14] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). Ssd: Single shot multibox detector. In European conference on computer vision (pp. 21-37). Springer, Cham.

[15] Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

[16] Zhang, S., Wen, L., Bian, X., Lei, Z., & Li, S. Z. (2018). Single-shot refinement neural network for object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 4203-4212).

[17] Chen, X. W., & Lin, X. (2014). Big data deep learning: challenges and perspectives. IEEE access, 2, 514-525.

[18] Zhang, T., Gao, C., Ma, L., Lyu, M., & Kim, M. (2019, October). An empirical study of common challenges in developing deep learning applications. In 2019 IEEE 30th International Symposium on Software Reliability Engineering (ISSRE) (pp. 104-115). IEEE.

[19] Junior, J. C. S. J., Musse, S. R., & Jung, C. R. (2010). Crowd analysis using computer vision techniques. IEEE Signal Processing Magazine, 27(5), 66-77.

[20] Mousavi, H., Mohammadi, S., Perina, A., Chellali, R., & Murino, V. (2015, January). Analyzing tracklets for the detection of abnormal crowd behavior. In 2015 IEEE Winter Conference on Applications of Computer Vision (pp. 148-155). IEEE.

[21] Smeureanu, S., & Ionescu, R. T. (2018, September). Real-time deep learning method for abandoned luggage detection in video. In 2018

26th European Signal Processing Conference (EUSIPCO) (pp. 1775-1779). IEEE.

[22] Liu, C., Gu, J., Kim, K., Narasimhan, S. G., & Kautz, J. (2019). Neural rgb (r) d sensing: Depth and uncertainty from a video camera. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10986-10995).

[23] Zhang, Y., & Funkhouser, T. (2018). Deep depth completion of a single rgb-d image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 175-185).

[24] LeCun, Y., Jackel, L. D., Bottou, L., Brunot, A., Cortes, C., Denker, J. S., ... & Vapnik, V. (1995, October). Comparison of learning algorithms for handwritten digit recognition. In International conference on artificial neural networks (Vol. 60, pp. 53-60).

[25] Girshick, R. (2015). Fast r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 1440-1448).

[26] Darknet: Open source neural networks in c. Available at: https://github.com/AlexeyAB/darknet

[27] Rajasegarar, S., Leckie, C., Palaniswami, M., & Bezdek, J. C. (2007, June). Quarter sphere based distributed anomaly detection in wireless sensor networks. In 2007 IEEE International Conference on Communications (pp. 3864-3869). IEEE.

[28] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934