The data conundrum: compression of automotive imaging data and deep neural network based perception

Do not include any identifying information; CIC is a double-blind review process.

Abstract

Video compression in automated vehicles and advanced driving assistance systems is of utmost importance to deal with the challenge of transmitting and processing the vast amount of video data generated per second by the sensor suite which is needed to support robust situational awareness. The objective of this paper is to demonstrate that video compression can be optimised based on the perception system that will utilise the data. We have considered the deployment of deep neural networks to implement object (i.e. vehicle) detection based on compressed video camera data extracted from the KITTI MoSeg dataset. Preliminary results indicate that re-training the neural network with M-JPEG compressed videos can improve the detection performance with compressed and uncompressed transmitted data, improving recalls and precision by up to 4% with respect to re-training with uncompressed data.

Introduction

Advanced driving assistance systems (ADASs) and automated vehicles (AVs) constitute an emerging technology which has the potential to revolutionise the transport sector "by making everyday journeys greener, safer, more flexible and more reliable" [1]. In fact, safety is of foremost importance, and automated functions can play a pivotal role in dramatically reducing accidents, considering that 94% of serious car accidents occur because of human negligence, according to NHTSA [2].

A way of defining the automation capability integrated in a vehicle is by using the SAE taxonomy, called Levels of Driving Automation [3]. These levels are used to describe the gradual transition of responsibility for driving tasks from the human to the machine, see Table I. Level 0 to Level 2 (L0-L2) encompass automated functions that are meant to support the driver, i.e. ADASs, whereas Level 3 to Level 5 cover AVs, including fully autonomous vehicles (L5) in which the existence of a driver is unnecessary.

Perceiving and understanding the environment and its actors (e.g. pedestrians, bicycles, cars, lorries, etc.) around a vehicle are complex tasks, which in ADASs and AVs are implemented using data generated by a sensor suite, namely an array of perception sensors (i.e. camera, LiDAR, Radar and Ultrasonic) [4]. Each of these heterogeneous sensor technologies has its own limitations, and a more complete perception of the situation may be achieved by combining information from all the sensors. This process is known as *sensor fusion* [5]. Sensor fusion leads to more accurate and higher resolution measurements, extends the temporal and spatial coverage and increases reliability [6]. However, the amount of data collected by perception sensors can be estimated at 3Gbit/s-40Gbit/s, and current automotive data networks have difficulty supporting these data rates [7-9]. One possible solution is to reduce or compress the amount of sensor data to be transmitted; in this work we will focus on video camera data compression. The novelty of this work is to consider video compression in combination with the perception task, namely with object detection implemented via deep neural data networks (NN) in ADASs and AVs. We demonstrate that carefully tuning a NN with compressed data can improve the accuracy of object detection with respect to the same NN trained only with uncompressed data.

Table I: Overview of the SAE Levels of Driving Automation from [3]; L0-L2 are related to ADASs, L3-L5 are related to AVs. Abbreviations: envir., environment; accel., acceleration; decel., deceleration.; ODD, Operational Design Domain, DDT, Dynamic Driving Task.

SAE	Human	Vehicle System	ODD	Examples of Functions
L0	All driving tasks	Audio/ visual/ haptic warnings to the driver	Limited	Blind Spot Monitoring; Lane Departure Warning
L1	Monitoring envir.; steering or accel./decel.; DDT in case of fallback.	Steering OR acceleration/ deceleration; feedback/ warnings to the driver	Limited	Adaptive Cruise Control; Lane Keep Assist; Collision Avoidance
L2	Monitoring envir.; DDT in case of fallback.	Steering AND acceleration/ deceleration; feedback/ warnings to the driver	Limited	Traffic Jam Assist
L3	Take back control of DDT in case of failure/exit from ODD	Monitoring environment; steering AND accel./ decel.; feedback/ warnings to the driver	Limited	Highway Autopilot; Emergency Driver Assistance
L4	No driving tasks	As L3 + fallback.	Limited	Geo-fenced autonomous driving
L5	No driving tasks	As L4	Un-limited	End-to-end autonomous driving

Background

Among the available perception sensors which may be deployed in a vehicle sensor suite, visible light cameras are an affordable solution with a good angular resolution and optionally providing colour information [4]. However the amount of data produced by the sensor suite, even by a modest 2 Mpixel camera

on its own can be too onerous to be transmitted by traditional vehicle data networks, so *ad hoc* expensive and heavy weight connectors and wiring are required. Alongside the need for multiple sensors to cover the entire 360° around the ego-vehicle and the required redundancy when considering higher levels of autonomy (L3 and above), the data produced by all the perception sensors cannot be supported by current standard data networks [7-9]. Hence, efficient compression techniques or saliency selection techniques need to be applied to perception sensor data to transmit data in a timely, reliable and affordable (in terms of cost, weight and required power) fashion. This paper's focus is automotive video compression, but we expect that similar challenges will apply to the other perception sensor technologies and their data.

Video compression can be categorized into lossy and lossless compression methods. *Lossless compression* allows the received data to be decoded into a received picture which is a bit-accurate copy of the original data. There is a limit to lossless compression which is based on the entropy of the raw data. Due to this limit, the vehicle's data network bandwidth can still be overloaded, leading to data loss. *Lossy compression* provides a higher degree of compression but usually results in a degradation of the received picture.

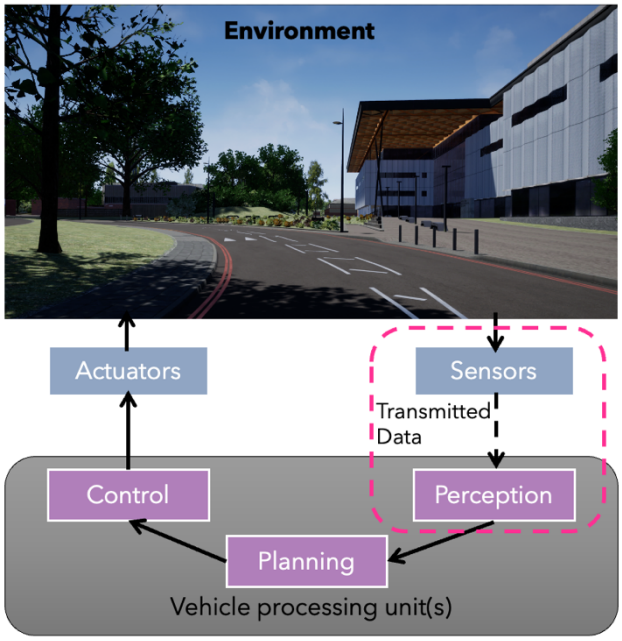

Figure 1. *Sensors and vehicle control loop, modified from [10]. Our focus is analysing the effect on the perception step (based on deep NN) when the transmitted data are compressed (highlighted in figure by the shape with the dashed contour).*

However, most of the work on lossy digital video compression has been designed to maintain the perceived video quality for human vision (leveraging our visual perceptive system weaknesses). Here instead we focus on compression for video consumed by the vehicle processing unit(s) of AVs or ADASs, Fig 1, and in particular we will focus on the perception step. This step entails the extraction of multiple types of information, e.g. localisation, identification of traffic objects (signs, lights, road markings, etc.) and also identification and tracking of surrounding hazardous objects [10]. This exploratory work is focused on vehicle detection, since vehicles are key road actors and an array of ADAS functions are determined from the detection of vehicles [11].

Camera data can be processed using traditional computer vision (CV) algorithms or through neural networks (NN). There is a trend in automotive engineering to use machine learning for object detection as it can provide high accuracy and higher flexibility to possible real-life variations (e.g. car models, colours, positions, brands, denting, etc.) with respect to traditional CV techniques [12]. Although NNs can be re-trained and tuned for their use in automated vehicles, they still have some drawbacks and currently it is impossible to achieve a detection accuracy of 100%, which poses a problem when AVs will be relying on these NNs. The inclusion of lossy compression in AV to ensure effective video data transmission can result in diminished video quality (e.g. loss of some details/objects), and/or introduction of artefacts. In this paper we evaluate the combined effect of compression and perception based on NNs, considering quality of videos for training, size of transmitted data and implication on the ADAS or AV perception stage.

Methodology

Our methodology aims to study the effect on object detection accuracy using an automotive camera video stream under different rates of lossy compression. We also consider the relationship between accuracy and compression rate for re-training the NN-based object detector. A schematic view of the applied methodology is presented in Fig.2, and its detailed description is provided in the following subsections. This structure aims to mimic an AV architecture wherein perception sensor data are compressed by or close to each sensor, then transmitted by the vehicle data networks to the vehicle processing unit/units, where data are consumed (i.e. we focus on the region delimited by the shape with dashed contour in Fig.1).

Deep NN based detection and training

In our experiments we used a pre-trained Faster RCNN (Faster Region proposed deep convolutional NN) architecture, ResNet50. ResNet50 operates on RGB input images, it has a depth of 50 convolutional layers, 25.6 M parameters and a size of about 96 MB; it provides a good compromise between detection accuracy and prediction speed with a relatively small size of the NN [13]. The training of the selected NN has been improved using different training sets, one at a time, to produce several different re-trained or *compression trained* Faster RCNN. These different training sets have themselves been created by applying different rates of compression to the chosen KITTI MoSeg dataset (described in the following subsection), from uncompressed videos to videos with a size of 15%, 7%, 4% and 2% with respect to the original dataset. Therefore, the number of compression trained NNs we have used is 5, with one of them re-trained with the original uncompressed MoSeg dataset. Note that the NN hyperparameters have been optimised for the NN re-trained with the uncompressed dataset, and kept constant for the compression trained NNs. The specific compression trained NN, used in each experiment to infer the prediction accuracy depending on neural network training and transmitted data compression rate, is represented by the light blue block in Fig. 2.

KITTI MoSeg dataset

In the last few years, whilst virtual verification has been gaining traction, datasets for training and testing of AV and ADAS functions have been proliferating [14-16]. These datasets usually contain data collected from several automotive environmental perception sensors mounted on test vehicles

driving in different regions of the world.

Amongst the above mentioned datasets, the KITTI dataset is a well-established benchmarking tool for the plethora of functions (from object detection to trajectory prediction, from segmentation to identification, etc.) that can be created for automated driving tasks and for computer vision in the automotive context [17]. KITTI offers data collected using one 360^{O} LIDAR scanner, four cameras and one inertial navigation system. The KITTI MoSeg dataset (a subset of the KITTI dataset) has been chosen in this paper as it provides time correlated, individual video frames with vehicle labels in each frame. It contains 1950 frames in total, with 2383 and 5997 annotated static and moving vehicles respectively [18-19].

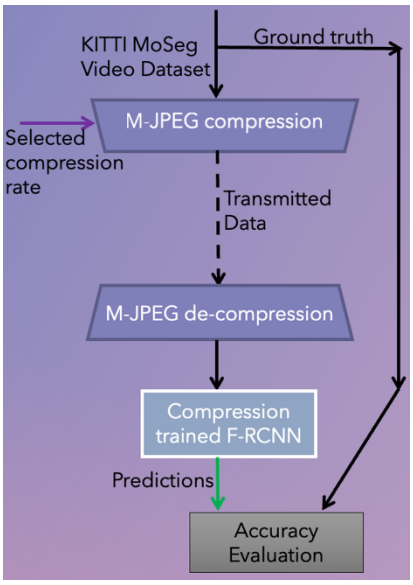

Figure 2. *Schematic view of the experimental methodology. The labelled video data from KITTI MoSeg dataset are used as the inputs and their labels are used as ground truth for evaluating NN accuracy. Data are compressed with different compression rates and then used as inputs of a compression trained Faster-RCNN (F-RCNN) employed to detect vehicles in the videos.*

Video Compression

Motion JPEG or M-JPEG is an intraframe-only compression method, i.e. the compressed image quality depends on the spatial complexity of the single frames and the method does not take into account movement and variation over time. M-JPEG exploits JPEG to compress each frame in the video stream, and it is a common format used in applications with low-latency requirements [20]. Due to time sensitivity of automotive application, we therefore decided to use M-JPEG as the tool to compress the video data to be transmitted.

In our experiment we have used the KITTI MoSeg dataset to create 5 different datasets by applying different M-JPEG compression rates (from uncompressed to a size of 2% with respect to the uncompressed videos). These 5 datasets correspond to the transmitted data in Fig.1-2, and they have also been used to generate the compression trained Faster RCNN.

Figure 3 a) and b) show a selected area from one frame of the KITTI MoSeg dataset with no compression and with maximum compression respectively. Some details have been zoomed and it is possible to observe the degraded quality of the compressed picture (right sides in Fig. 3 c)-f)), with the traffic light in the distance no longer visible, the degraded intensity of brake lights, the P sign characters and arrow distorted, the blocky artifacts arising in different parts of the frame.

Accuracy Evaluation

There are several performance metrics computed over different criteria that can be used to evaluate the performance of an object detector [21-22]. In the case of the KITTI MoSeg dataset the ground truth is in the form of the vehicle bounding boxes associated with each frame. We compare these to the bounding boxes provided as an output by the compression trained Faster RCNN. The comparison is implemented by calculating intersection over unit (IoU) of predicted versus ground truth bounding boxes; for our experiment the criterion for a successful match is that an object is detected when IoU>0.5. Based on this definition, for each compression trained NN and each set of transmitted data (25 combinations) we have evaluated recall (R) and average precision (AP) for the NN results [21-22].

Figure 3. *A selected area from one frame of KITTI MoSeg dataset, (a) uncompressed, (b) maximum compression; the rectangles in a) and b) highlight the details which are zoomed: c) traffic light (not visible on the right), d) brake lights (intensity degraded on the right), e) P sign characters (distorted on the right), f) parking slot (blocky artifacts on the right). In c), d), e) and f) the details from the frame uncompressed and with maximum compression are on the left and right sides respectively.*

Results

As a first attempt, we re-trained and fine-tuned the hyperparameters of the ResNet50 network using the uncompressed KITTI MoSeg dataset (so the light blue block in Fig.2 was re-trained with the original videos). Then we observed the effect on the NN output when data are transmitted over the vehicle data networks with different compression rates. Figure 4 a) shows the recall values (dashed line) and average precision (continuous line) for the NN output (i.e. bounding boxes of detected vehicles in the videos) as a function of the size of the transmitted video data (the size of the compressed dataset is normalised versus the size of the original dataset and expressed as a percentage, i.e. a size of 100% represent the uncompressed data). The figure shows that this NN has good performance for increasing compression rates, however R and AP start to degrade when the size of the transmitted data is below the 10% of the original dataset.

Fig. 4 b) shows again R and AP for a compression trained NN, namely the network re-trained with a dataset size of the 7% with respect to the original KITTI MoSeg dateset. Notably, the compression tuned neural network outperforms the one re-trained with the original dataset for all the sizes of transmitted data.

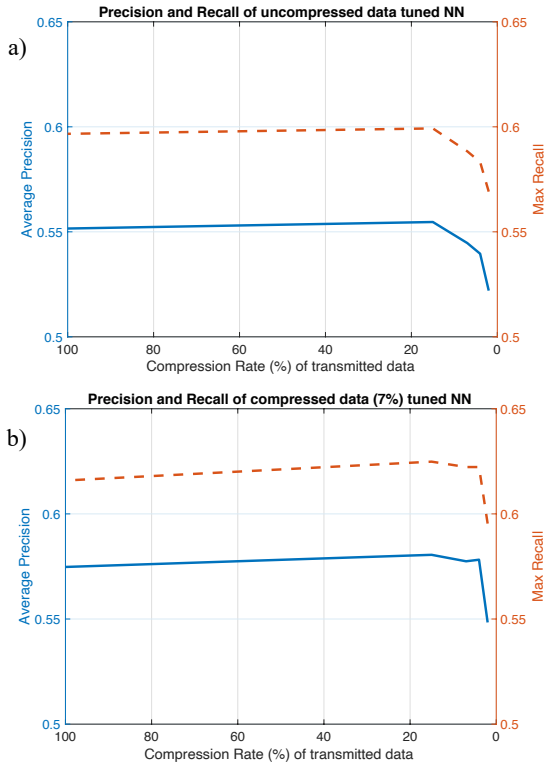

Figure 4. *Recall values (dashed line) and average precision (continuous line) versus transmitted video data size (a) for the NN trained on uncompressed data, and (b) for the NN trained with compressed training data (7% size with respect to the original dataset).*

Figures 5 a) and b) summarise the results achieved with the compression tuned NNs. For almost all the possible combinations the compression tuned NNs outperform the NN re-trained with uncompressed data or re-trained with datasets with a size above the 10% of the original dataset. However, when the size of the transmitted compressed dataset is the smallest (2% of the original dataset) both the recall values and average precision are below the other transmitted datasets, demonstrating an increase in the false positives (FP) and false negatives (FN) generated by the NN. This increase can be due to artifacts manifesting at higher compression rates, and further investigation of the frames and detections is required to understand if it possible to further tune the NN hyperparameters to counteract this issue. Furthermore, in these extreme cases, the compression tuned NNs are more able to counteract the effect of compression on transmitted data, and the tuning is particularly beneficial in decreasing false negative with respect to the NN re-tuned on uncompressed data. It is possible that compression trained NN are more robust to distortions/variations in the targets or that M-JPEG quantisation and high frequency filtering causes a reduction of noise in the transmitted data. However compression artifacts can still cause high false positives.

It can also be noted from all the figures that recalls are always higher than that the average precision, indicating that the number of FPs is higher than the number of FNs. It is of outmost importance that FNs are suppressed in automotive, as they are targets (namely, in our study, vehicles) which are not detected and can therefore be the cause of an accident. Therefore, fine-tuning needs to target a further reduction in the number of FNs.

Conclusion and Further work

The results presented in this work suggest not only that compression can be successfully used with deep neural network based object detection in videos, but that the detection accuracy can be improved using an appropriately retrained network. In fact, retraining the NN with compressed data can provide results that exceed the detection performance of a system not employing compression, until the compression reaches a level that it starts to cause severe performance degradation. More work is in progress to further optimise the hyperparameters of compression trained NN, and also to consider bigger training datasets, raw sensor data and other NN architectures. In this paper we have used motion JPEG to compress the dataset, as it is a widespread approach in low latency application, however M-JPEG is best suited for videos with low-complexity, and achievable compression rates are lower with respect to other methods (e.g. MPEG). Therefore, we are also currently evaluating other compression techniques and their suitability/adaptability for highly dynamic and crowded automotive scenarios.

This work aims to pave the way to a holistic approach to the perception sensor data conundrum in automated vehicles, considering vehicle sensor fusion architecture and where it is convenient to process or pre-process data. From one side there is the need to collect the information of the vehicle surrounding environment with rich details and enough redundancy to support the higher levels of autonomy, and from the other side there is the need to transmit these data with low latency, without dramatically and impractically increasing the cost and power needed for storage, transmission and elaboration. Furthermore, this paper is focused specifically on data produced by sensors in AVs and ADAS functions, however big data transmission, storage and compression is becoming a troublesome issue in several fields, e.g. robotics, manufacturing, biostatistics, etc., and we expect that this work will influence the approaches to data pre-processing in numerous fields.

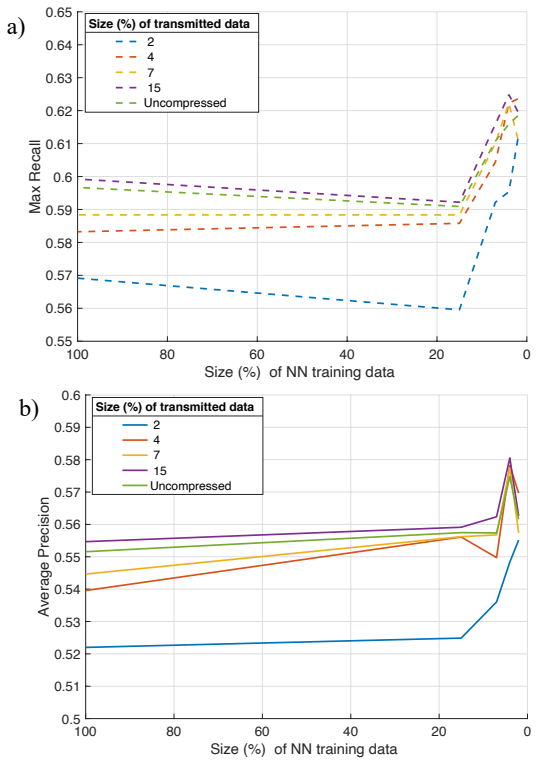

Figure 5. *(a) Recall values (dashed lines) and (b) average precision (continuous lines) as a function of the size of the dataset used to re-train the NN for different sizes of the transmitted data (different colours).*

References

[1] Centre for Connected and Autonomous Vehicles, About us [Accessed online 2 April 2021] Available at: https://www.gov.uk/government/organisations/centre-for-connected-and-autonomous-vehicles/about

[2] NHTSA, Automated Vehicles for Safety, [Accessed online 2 April 2021] Available at: https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

[3] "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," SAE, Warrendale, PA, USA, 2018.

[4] C.-P. Hsu et al., "A review and perspective on optical phased array for automotive LiDAR," *IEEE J. Sel. Topics Quantum Electron.*, vol. 27, no. 1, pp. 1–16, Jan./Feb. 2021

[5] J. Kocić, N. Jovičić, V. Drndarević, "Sensors and Sensor Fusion in Autonomous Vehicles," *2018 26th Telecommunications Forum (TELFOR)*, Belgrade, Serbia, pg. 420-425 (2018).

[6] S. Gogineni, "Multi-Sensor Fusion and Sensor Calibration for Autonomous Vehicles," *International Research Journal of Engineering and Technology (IRJET),* Volume 07, Issue 07 July 2020.

[7] S. Heinrich. "Flash memory in the emerging age of autonomy.", *Proceedings of the Flash Memory Summit, Santa Clara, CA,* USA (2017): 7-10.

[8] S. Tuohy, M. Glavin, C. Hughes, E. Jones, M. Trivedi, and L. Kilmartin. "Intra-vehicle networks: A review." *IEEE Transactions on Intelligent Transportation Systems* 16, no. 2 (2014): 534-545.

[9] L. van Dijk, and G. Sporer. "Functional safety for automotive ethernet networks." *Journal of Traffic and Transportation Engineering* 6, no. 4 (2018): 176-182.

[10] S.D. Pendleton, H. Andersen, X. Du, X. Shen, M. Meghjani, Y.H. Eng, D. Rus, M.H. Ang. "Perception, planning, control, and coordination for autonomous vehicles," *Machines*. 2017 Mar;5(1):6.

[11] X. Hu et al., "SINet: A Scale-Insensitive Convolutional Neural Network for Fast Vehicle Detection," in IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 3, pp. 1010-1019, March 2019

[12] X. Zhao, P. Sun, Z. Xu, H. Min and H. Yu, "Fusion of 3D LIDAR and Camera Data for Object Detection in Autonomous Vehicle Applications," in *IEEE Sensors Journal*, vol. 20, no. 9, pp. 4901-4913, 1 May1, 2020.

[13] MATLAB 2020b, "Pretrained Deep Neural Networks," 2020, [Online]. MathWorks. Available: https://www.mathworks.com/, Accessed Apr. 21, 2021.

[14] J. P. Espineira, J. Robinson, J. Groenewald, P. H. Chan and V. Donzella, "Realistic LiDAR With Noise Model for Real-Time Testing of Automated Vehicles in a Virtual Environment," in *IEEE Sensors Journal,* vol. 21, no. 8, pp. 9919-9926, Apr. 15, 2021,

[15] J. Geyer *et al*. "A2d2: Audi autonomous driving dataset," arXiv preprint arXiv:2004.06320. 2020 Apr 14.

[16] Y. Kang, H. Yin and C. Berger, "Test Your Self-Driving Algorithm: An Overview of Publicly Available Driving Datasets and Virtual Testing Environments," in *IEEE Transactions on Intelligent Vehicles,* vol. 4, no. 2, pp. 171-185, June 2019,

[17] C. Geiger, P. Lenz and R. Urtasun, " Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite," *Conference on Computer Vision and Pattern Recognition (CVPR),* 2012, pp. 1-8.

[18] Siam et.al, "MODNet: Moving Object Detection Network with Motion Appearance for Autonomous Driving,", arXiv preprint arXiv 1709.04821, 2017.

[19] A. Geiger, P. Lenz, C. Stiller and R Urtasun, "Vision meets Robotics: The KITTI Dataset," *International Journal of Robotics Research (IJRR),* 2013.

[20] E. Belyaev, L. Bie and J. Korhonen, "Motion JPEG Decoding via Iterative Thresholding and Motion-Compensated Deflickering," *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP),* 2020, pp. 1-6,

[21] R. Padilla, S. L. Netto and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," *2020 International Conference on Systems, Signals and Image Processing (IWSSIP),* 2020, pp. 237-242,

[22] N. K. Ragesh and R. Rajesh, "Pedestrian Detection in Automotive Safety: Understanding State-of-the-Art," in *IEEE Access*, vol. 7, pp. 47864-47890, 2019,