Optimising Faster R-CNN training to enable video camera compression for assisted and automated driving systems

Dr Valentina Donzella†
WMG, University of Warwick
Coventry, United Kingdom
V.Donzella@warwick.ac.uk

Pak Hung Chan†
WMG, University of Warwick
Coventry, United Kingdom
Pak.Chan.1@warwick.ac.uk

Dr Anthony Huggett
ON Semiconductor
Bracknell, United Kingdom
Anthony.Huggett@onsemi.com

***Abstract*—Advanced driving assistance systems based on only one camera or one RADAR are evolving into the current assisted and automated driving functions delivering SAE Level 2 and above capabilities. A suite of environmental perception sensors is required to achieve safe and reliable planning and navigation in future vehicles equipped with these capabilities. The sensor suite, based on several cameras, LiDARs, RADARs and ultrasonic sensors, needs to be adequate to provide sufficient (and redundant, depending on the level of driving automation) spatial and temporal coverage of the environment around the vehicle. However, the data amount produced by the sensor suite can easily exceed a few tens of Gb/s, with a single 'average' automotive camera producing more than 3 Gb/s. It is therefore important to consider leveraging traditional video compression techniques as well as to investigate novel ones to reduce the amount of video camera data to be transmitted to the vehicle processing unit(s). In this paper, we demonstrate that lossy compression schemes, with high compression ratios (up to 1:1,000) can be applied safely to the camera video data stream when machine learning based object detection is used to consume the sensor data. We show that transfer learning can be used to re-train a deep neural network with H.264 and H.265 compliant compressed data, and it allows the network performance to be optimised based on the compression level of the generated sensor data. Moreover, this form of transfer learning improves the neural network performance when evaluating uncompressed data, increasing its robustness to real world variations of the data.**

***Index Terms*—perception sensors, unmanned vehicles, assisted and automated driving, neural networks, intelligent systems, autonomous systems**

I. INTRODUCTION

Vehicles equipped with assisted and automated driving (AAD) functions will provide many benefits from increased safety and reduction of accidents to allowing more efficient journeys and increased productivity of their drivers when disengaged from the driving task [1]. Many current vehicle manufacturers already offer models equipped with a range of advanced driver assistance systems that provide up to level 3 out of the six levels of automation (levels 0 to 5), as defined by the Society of Automotive Engineers [2]. Between levels 0-3, the lateral and longitudinal control of the vehicle can be increasingly delegated to vehicle systems. However, the driver is still required to be always alert and to take over control when the situation cannot be handled by the system. Up to level 2 the driver is still responsible for monitoring the environment, whereas from level 3 and above the system is in charge of this task. For levels 4 and 5, the vehicle can fully perform the driving task within the operational design domain, with the main difference that level 5 can accomplish this task under all the possible circumstances and everywhere.

Environmental perception sensors (i.e. camera, RADAR, LiDAR, etc.) are the interfacing element between the AAD system and the environment. They generate the information required by the computing units to plan the vehicle actions.With the number of sensors and the amount of data predicted to be generated (i.e. in excess of 40 Gb/s) for higher levels of autonomy (levels 3-5), current wired vehicle data networks are inadequate to reliably transmit the required data amount [3]–[5]. With the increased demand of automotive cameras providing high resolution (8-12 Mpixel) and the required high dynamic range (HDR) to cope with the luminosity variations when driving (e.g. bright sun in front of the sensor when travelling in a dark tunnel), cameras can significantly contribute to the amount of generated data by the sensor suite; moreover multiple cameras are required to provide 360°coverage of a vehicle's surroundings. Compression or reduction of sensor data will be key to allow a prompt and cost-effective transmission of the data produced by the multiple sensors in higher levels of autonomy. Lossless compression does not change the original data but allows a relatively low amount of compression. Lossy compression allows a more significant reduction in the amount of data, but the original data will not be exactly recoverable. Therefore, lossy compression needs to be thoroughly investigated to understand if the introduced changes to the sensor data will affect the vehicle safety.

This work investigates the effects of compression on one of the potential consumers of automotive camera data, i.e. an object detector based on a deep neural network (NN). Different levels of compression and different compression standards (namely AVC/H.264 and HEVC/H.265) have been analysed.

† These Authors contributed equally.

Dr Donzella acknowledges the support of the Royal Academy of Engineering via the Industrial Fellowships scheme. The Authors wish to acknowledge the support of ON Semiconductor and High Value Manufacturing CATAPULT.

More importantly, lossy compressed videos have been used to retrain the NN object detector, demonstrating an improvement of the average precision and recall when evaluating transmitted data. This improvement is noticeable not only for compressed data (up to a compression of about 1:1,000) but also for lossless data. These results help to pave the way to the use of lossy video compression in automotive applications.

II. Background

Video compression has been traditionally optimised for the Human Visual System, in the context of digital television and video streaming. However, there is an increasing deployment of NN to postprocess videos and frames to extract higher level information from them, as highlighted in [6]. In particular, in this work, we focus on the consumption of the video camera stream in AAD functions, used to build the vehicle situational awareness and support the *perception-planning-control* pipeline. Recent research in academia and industry has been focusing on the effects of data reduction and compression to decrease the amount of data that needs to be transmitted from the sensors to the processing units. The *caveat* is that also the sensor fusion strategy can have an impact on the need of data reduction, however, solutions like centralised or zonal architectures may need the transmission of raw sensor data and therefore a careful consideration of data reduction/compression techniques [7]. Nonetheless, AAD are safety critical systems, therefore any latency, data loss, or artefacts created when reducing/encoding the sensor data can be the cause of incorrect decisions and therefore need to be carefully analysed. In the following subsection we will review some of the recent works covering the specific challenges of video compression for AAD systems. It is worth noting that a considerable amount of research is developing also in the area of LiDAR data reduction, but it is outside the scope of this paper. A good coverage of some of the challenges associated with pointcloud reduction can be found in [8].

A. Automotive camera video compression

The necessity of considering the combination of sensor data compression with the specific automotive system requirements has been known for several years. For example, Foster *et al.* evaluated the combination of compression algorithms (based on JPEG, JPEG2000, and AVC/H.264) in combination with a semi-global matching algorithm for combining stereo camera images, and suggested to optimise the compression standard based on the required compression ratio [9]. Moreover, in the last few years, with the push towards vehicle automation and the discussed need of several bandwidth-hungry perception sensors, automotive video compression has been studied in combination with some potential solutions. These solutions have mainly two main approaches: one is to study traditional compression techniques in combination with the perception step (e.g. object detection, segmentation, tacking, etc.) and if possible to optimise their interaction [10], [11]; the other one is related to partially or completely(end-to-end) substituting traditional algorithms with machine learning (i.e. deep neural

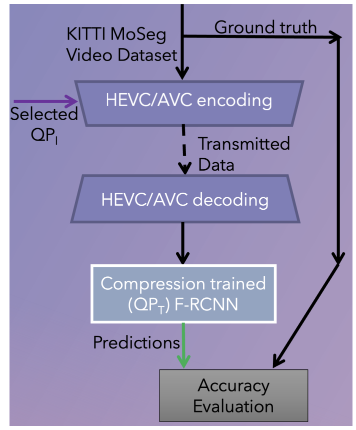


Fig. 1. Schematic view of the experimental methodology. The original KITTI MoSeg dataset and its labels are used as the baseline. The videos, compressed with different compression rates and the chosen compression standard (AVC or HEVC), are transmitted as the inputs of a compression trained Faster-RCNN (F-RCNN). This NN has been trained with a selected compressed dataset amongst the 14 generated.

TABLE I
Compression Ratio of the Generated Datasets with Respect to the Size of the Originial Kitti MoSeg Dataset (i.e. *circa* 16 MB)

QP	0	17	23	29	35	41	51
AVC	1:20	1:110	1:250	1:570	1:1280	1:2820	1:8040
HEVC	1:18	1:80	1:160	1:380	1:1040	1:3140	1:10220

networks) methods and compare the performance of these NN techniques with traditional compression methods. Our proposed approach belongs to the first category, and a first proof of concept with low compression ratio and M-JPEG has been given in [11].

Different works have been focusing on different perception steps as the consumers of the camera data, for example Beratoglu at al. have investigated H.265/HEVC compression in combination with a YOLO V3 tiny object detector for detection of vehicle plates. The authors demonstrated a benefit in terms of the decreased bandwidth requirements for the same level of inference accuracy when comparing compressed data to lossless. Furthermore, they discuss a decrease of the NN inference time, which can be key in low-latency time sensitive automotive applications [12]. Tanaka *et al.* carried out a similar investigation: compression based on HEVC and tracking based on the combination of YOLO V3 and SORT models [13]. The results showed a negligible variation of the tracking accuracy at low compression ratios (i.e. QP up to 30), however the authors acknowledge the need to investigate other tracking techniques. Aktar *et al.* have carried out a detailed study on H.265/HEVC compression and its relationship with a specific step implemented in NN, which is the identification of the intersection over union (IoU) value between the ground truth and the bounding boxes predicted by the network [14].

In terms of end-to-end NN approaches for compression, several network architectures have been proposed and evaluated specifically for automotive. Lohdefink *et al.* designed

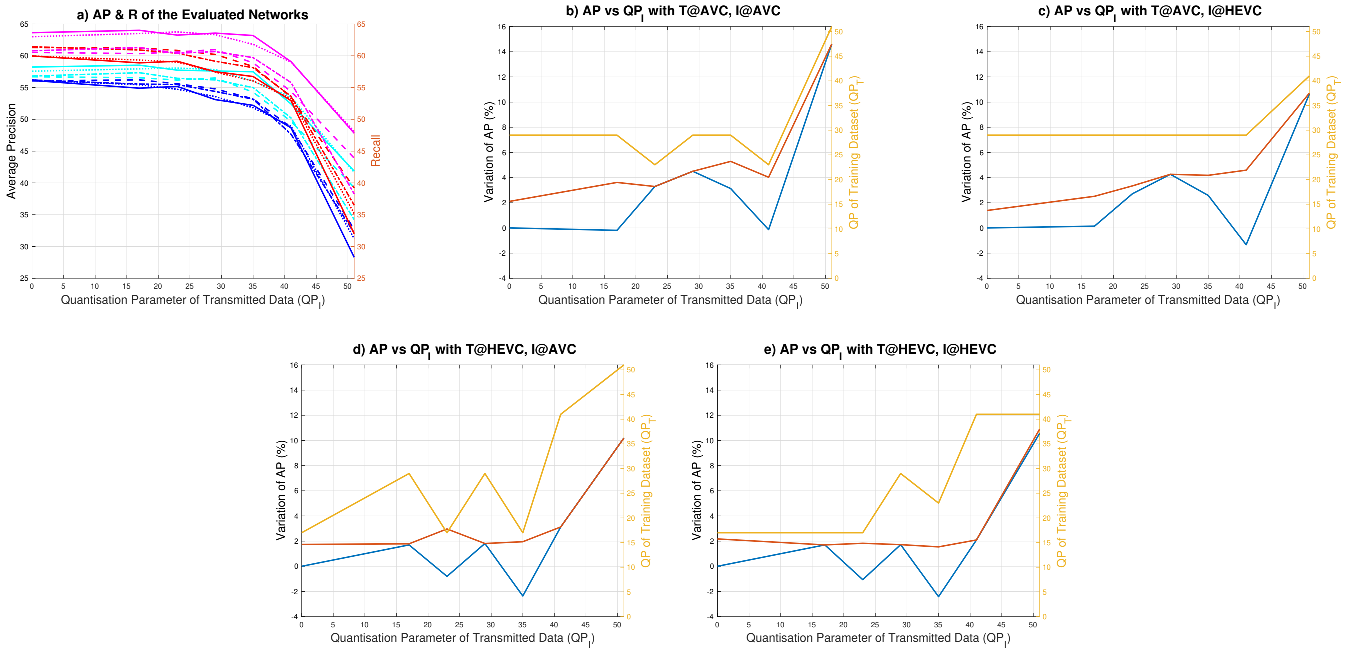


Fig. 2. a) AP (left axis, blue and cyan plots) and R (right axis, red and magenta plots) versus increasing compression rate of transmitted data for the NN trained with the datasets at QP_T=0 (blue and red plots) and QP_T=29 (cyan and magenta plots). Different linestyles represent the four different combinations of compression standards for transmitted and training data. b)-e) AP variation (left axis) with respect to baseline (QP_T=0) for NN with QP_T=QP_I (blue plots) and for the NN with the best performance for each QP_I (orange plots). The optimised QP_T value is shown on the right axis (yellow plots).

a generative adversarial network (GAN) to implement an encoder/decoder compression method and compared it with different traditional compression algorithms (namely JPEG, JPEG2000, and WebP). The interesting point discussed, linked with the evaluation approach adopted in our paper (i.e. using NN object detection), is that the traditional performance metrics used for image evaluation (i.e. peak signal-to-noise ratio, PSNR, structural similarity, SSIM) are inadequate to evaluate the implications on the NNs using the compressed videos. Therefore, the authors used also a semantic segmentation model to compare their GAN method with the traditional compression schemes and demonstrated an improvement in the segmentation despite the fact that the GAN method has lower performance in terms of PSNR [15]. Another approach based on deep NN has been presented by Lu *et al.*, however it has not been discussed specifically in the automotive context. The Authors used optical flow estimation leveraging a learning approach and combined it with motion estimation to reconstruct frames. The results outperform videos compressed with AVC and are comparable with HEVC compressed ones, but the Authors compare the results only based on SSIM and do not discuss latency nor the implication on the performance of the final consumer of the compressed data [16] . Another promising NN approach, but not end-to- end, is presented by Sankisa *et al.* [17]. Interestingly the Authors proposed to use a deep neural network (based on 3DConv and ConvLSTM layers) to decode compressed videos based on predictions using exclusively past frames; this NN can also conceal errors due to the transmission channel. The results are evaluated only based on PSNR and SSIM, but the proposed approach can certainly open new areas of investigation.

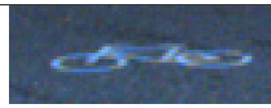
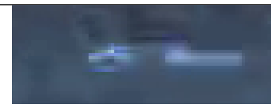
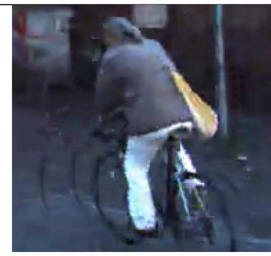
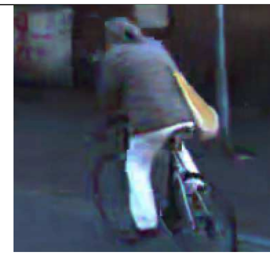
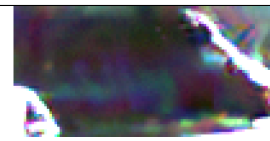
III. Methodology

This paper follows on our work with M-JPEG [11], and investigates the effect of compression using AVC and HEVC on NN based vehicle detection. Compared to M-JPEG, AVC and HEVC based compression techniques are more sophisticated and incorporate interframe interpolation. Taking advantage of temporal redundancy via interframe interpolation can provide a strong advantage in automotive applications, as it is highly probable that video frames will contain vehicles, pedestrians, bikes etc. moving with a certain speed and repeating in consecutive frames at different positions. The experimental process has been schematically illustrated in Fig. 1; the compression rate was controlled using constant Quantisation Parameter (QP) for both AVC and HEVC codecs. It is worth noting that we used compressed data for two different aims: one aim is to mimic the compressed data to be transmitted on the wired vehicle communication networks (transmitted dataset, QP_I); the other aim is that the compressed data have been used also to re-train the selected NN (training dataset, QP_T). In this work, we tested our experimental flow using different combinations of training and transmitted datasets; more details can be found in [18].

A. Compression of the KITTI MoSeg Dataset

We have used a constant QP to control the compression of the generated datasets using X264 and X265 ffmpeg libraries

TABLE II
COMPARISON OF ARTEFACTS RESULTING FROM COMPRESSION, IMAGES HAVE BEEN ENLARGED FROM THEIR ORIGINAL SIZE

Artefact (Identified Errors)	**Original**	**Compressed at 41 QP AVC**	**Compressed at 41 QP HEVC**
Loss of edge definition (deformation of shapes, blurring of bicycle)			
Removal of high frequency and blocking artefacts (blurring of coat pattern, ground and face of person are "blocked")			
Similar colour uniformised, loss of edge definition (vehicle blurring and blending into van)			

(compatible with AVC and HEVC respectively); this choice allows the resulting videos to have similar visual quality across their frames, with the downside that the bitrate can vary during transmission. In this work, we have chosen to compress the KITTI Moseg dataset at QP of 0, 17, 23, 29, 35, 41 and 51 in both AVC and HEVC to generate 14 datasets. The compression ratios achieved by the two standards are shown in Table I, with HEVC achieving slightly lower levels of compression at lower QPs and higher levels of compression for the extreme cases of QP =41,51. We will refer to the datasets generated with QP=0 as uncompressed dataset, whereas the other will be named compressed datasets.

B. Neural Network

Faster R-CNN, a two stage object detector, was chosen to be used in this work. Although two stage NNs do not provide the same speed of one stage object detectors (e.g. compared to the previously discussed YOLO V3 used in [12], [13]), they are often more accurate, making them advantageous for automotive applications [19]. Faster R-CNNs have been proposed as the fastest version of two stage NN, and we have therefore selected a pre-trained Faster R-CNN as a good compromise between accuracy and speed [20]. The backbone used is ResNet-101 which extracts the features within the proposed regions in the first stage of the network.

To carry out our experiments, we have generated a series of compression tuned Faster R-CNN re-training the original network with each one of the QP_T compressed datasets; re-training has been implemented using always the same hyper-parameters (optimised for the NN trained with lossless data). Each retrained NN has then been used to evaluate the datasets compressed through QP_I. Similarly to what we presented in [11], the average precision, AP, and recall, R, across the vehicle class (including cars, lorries, stationary and dynamic targets) are calculated for each combination of QP_T trained network and QP_I compressed evaluation dataset. Additionally, the compression tuned NNs have been used with training and evaluating datasets achieved with different compression standards (i.e. training with AVC and evaluating the HEVC compressed datasets – and *vice versa*). As a consequence, we have generated 196 evaluations of our compressed datasets combined with the compression tuned NNs.

IV. RESULTS AND DISCUSSION

From now on, we will consider as the baseline for our experiments the performance of uncompressed re-trained neural network, blue plots in Fig. 2a. The curves show the calculated AP when the compression of the transmitted data is increasing, and the four different linestyles represent the four combinations of the compression standards used for training and transmission (namely AVC-AVC, continuous, AVC-HEVC, dotted, HEVC-AVC, dash-dotted, HEVC-HEVC, dashed). Moreover, the light blue plots show the same 4 combinations but for the compression tuned NN, in this case tuned with a QP_T = 29. We then compared the baseline with two cases: one is assuming to transmit data with the same level of compression of the NN tuning (QP_T = QP_I); the other one is selecting for each QP_I the QP_T that yields to the best AP. In this case the optimized QP_T is plotted in yellow (right axes) in Fig. 2b-2e. The case QP_T = QP_I is represented by the blue lines (left axes) in Fig. 2b-2e (for the four combinations of AVC and HEVC) and the max achievable AP (when fixing QP_I) by the orange lines. On the left axes we have the variation of AP with respect to the baseline, when this value is positive means that the compression tuned NNs are performing better than the uncompressed tuned NNs. We can observe that transmitting data with the same level of compression of the compression tuned NN not always yields improved results with respect to the baseline, particularly in HEVC. However, if the compression rate of data used for re-training the NN is optimised (i.e. by optimising QP_T) the compression tuned NNs outperform the baseline for all the QP_I, including for uncompressed data. Interestingly, the orange plots show that it is always possible to achieve around 2% increase in performance when not training with uncompressed data, and even if different compression standards are used for training and transmitted data. In fact, the increase in AP is of at least 2% for lower values of QP_I (i.e. $QP_I \leq 35$), and it becomes significantly higher at very high levels of compression of the transmitted data (up to an AP increase of 10%-14%). This result shows that specifically when numerous compression artefacts start to arise in the coded/decoded frames (for some examples see Table II) compression tuning can mitigate the performance drop shown in Fig. 2a. Moreover, for these extreme cases, it is beneficial to tune the network with higher levels of compression (QP_T = 41 or 51), as potentially the compression tuned NNs learn to recognise the vehicles with the associated compression artefacts. It is worth noting that some artefacts become significant at higher levels of compression, see Table II, and it is not surprising that the NN performance significantly decreases for $QP_I \geq 41$. By inspecting the decoded frames, it is possible to observe objects blurring into the background, such as vehicles and bikes, loss of edge definition (e.g. the bike sign on the road), blocking artefacts and removal of high frequency details, ghost artefacts due to temporal predictions (e.g. see the prediction of the bike movement using AVC). To further inspect the artefacts due to high compression regimes, we have plotted the loss per pixels between the original dataset frames and the compressed frames in AVC and HEVC respectively, an example is reported in Fig. 3b and 3c for QP = 51 compressed frames. As expected, there is a clear and substantial loss around small details (e.g. the vehicle licence plate), around the edges of objects, visible temporal ghosting artefacts around some objects and blocking artefacts (e.g. in the sky). Interestingly, there is a higher green loss in HEVC with respect to AVC.

In terms of a selection of an optimal compression rate for training, i.e. optimising QP_T for a range of transmitted compressed data (yellow plots in Fig. 2b-2e), retraining with AVC can be optimised by selecting $23 \leq QP_T \leq 29$; using this range of values for compression tuning will guarantee an improvement in performance for both HEVC and AVC compressed transmitted data, up to a QP_I of 35. Similarly, re-training the NN with HEVC compressed data can be optimised by selecting $17 \leq QP_T \leq 23$; using this range of values allows performance to be enhanced for both HEVC and AVC compressed transmitted data, up to a QP_I of 35. Finally, Fig. 4 shows the precision versus recall plots calculated across the frames while the NN is evaluating the transmitted datasets. By comparing lower levels of compression for training data in Fig. 4a-4d, QP_T = 35, with the maximum compression in Fig. 4e and 4f, QP_T = 51, we can clearly observe a degradation in the performance of precision and recall. Moreover, the trends in the plots are more influenced by the standard used for training than the standard used for transmitted data. This aspect demonstrates that re-training has a strong impact on the inference performance of the NN; on the contrary the compression standard used for the transmitted data does not have a strong influence on the performance. When training with AVC (at QP_T =35), the data transmitted with QP_I =41, 51 have clearly worse performance than the other curves, whereas when training with HEVC (at QP_T =35) only QP_I =51 is visibly worse than the other compression rates.

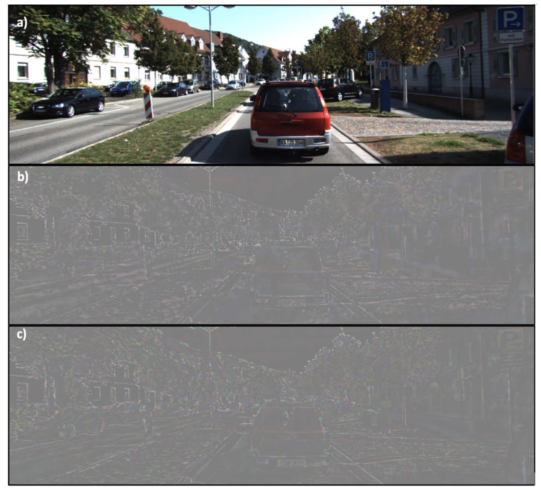


Fig. 3. a) A frame from the original KITTI MoSeg dataset, and the loss per pixels between the original frame and the compressed frames (QP=51) using (b) AVC and (c) HEVC standards.

V. Conclusion

This work presents an impactful strategy to use compressed video camera data to support assisted and automated driving functions, thereby reducing the bandwidth requirements and lowering the overall cost of the fusion and processing solution. The proposed strategy is based on tuning the perception step (namely object detection) using compressed data. In fact, the presented results imply not only that compressed video data can be fed into object detectors based on a deep neural network without affecting the detection performance, but in addition to that the detection precision can be significantly enhanced when using compressed data to re-train the NN. Interestingly, compression tuning the object detector provides results surpassing the performance of a system not employing compression based re-training, in the case of transmitted data which can be compressed or uncompressed, and therefore enhancing the robustness of the system. The NN hyperparameters have been optimised in the case of re-training with not compressed data, therefore there is the possibility to further improve the presented results of the compression tuned NNs by optimizing each one of them. One key aspect that we have demonstrated is that when compression tuning with

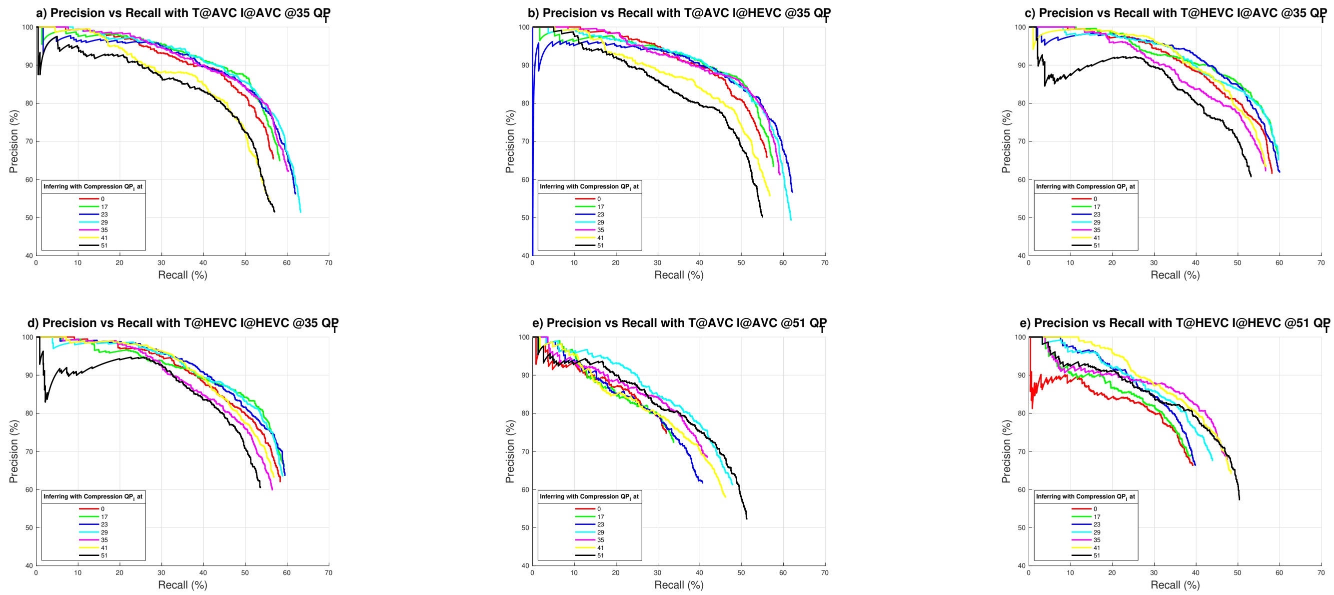


Fig. 4. Precision versus recall plots calculated across the frames while the NN is evaluating the transmitted datasets, a)-d), QP_T=35, e)-f), QP_T=51.

AVC compressed data an optimal value for the training data compression is $23 \leq QP_T \leq 29$, whereas when compressing with HEVC is $17 \leq QP_T \leq 23$. The presented results are focused on real-time automotive applications, however data compression will be key for curated dataset, storage, black boxes, etc., and we expect that this work can impact several different applications.

References

[1] Zenzic, "UK Connected and Automated Mobility Roadmap CAM Creators Update," Zenzic-UK Ltd, Tech. Rep., 2020. [Online]. Available: https://zenzic.io/roadmap/

[2] SAE On-Road Automated Driving (ORAD) committee, "J3016 - Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles," *SAE International*, p. 41, 2021.

[3] C.-P. Hsu, B. Li, B. Solano-Rivas, A. R. Gohil, P. H. Chan, A. D. Moore, and V. Donzella, "A Review and Perspective on Optical Phased Array for Automotive LiDAR," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 27, no. 1, pp. 1–16, 2020.

[4] S. Heinrich, "Flash memory in the emerging age of autonomy," *Flash Memory Summit*, pp. 1–10, 2017.

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

[6] K. Fischer, C. Herglotz, and A. Kaup, "On Intra Video Coding and In-Loop Filtering for Neural Object Detection Networks," in *2020 IEEE International Conference on Image Processing (ICIP)*, Abu Dhabi, 2020, pp. 1147–1151.

[7] D. J. Yeong, G. Velasco-hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, pp. 1–37, 2021.

[8] M. Labussière, J. Laconte, and F. Pomerleau, "Geometry Preserving Sampling Method Based on Spectral Decomposition for Large-Scale Environments," *Frontiers in Robotics and AI*, 2020.

[9] J. Forster, X. Jiang, A. Terzis, and A. Rothermel, "Evaluation of compression algorithms for automotive stereo matching," in *2012 IEEE Intelligent Vehicles Symposium*, 2012, pp. 1017–1022.

[10] Y. Wang, P. H. Chan, and V. Donzella, "A two-stage h. 264 based video compression method for automotive cameras."

[11] P. H. Chan, G. Souvalioti, A. Huggett, G. Kirsch, and V. Donzella, "The data conundrum: compression of automotive imaging data and deep neural network based perception," in *London Imaging Meeting 2021: Imaging for Deep Learning*. London: Society for Imaging Science and Technology, 2021, pp. 78–82.

[12] M. S. Beratoğlu and B. U. Töreyin, "Vehicle License Plate Detector in Compressed Domain," *IEEE Access*, vol. 9, pp. 95 087–95 096, 2021.

[13] T. Tanaka, A. Harell, and I. V. Bajić, "Does Video Compression Impact Tracking Accuracy?" in *IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2022. [Online]. Available: http://arxiv.org/abs/2202.00892

[14] M. S. Aktar and Y. Horita, "Performance analysis of vehicle detection based on spatial saliency and local image features in H.265 (HEVC) 4K video for developing a relationship between iou and subjective evaluation value," *IEEJ Transactions on Electrical and Electronic Engineering*, vol. 15, no. 4, pp. 563–569, 2020.

[15] J. Lohdefink, A. Bar, N. M. Schmidt, F. Huger, P. Schlicht, and T. Fingscheidt, "On low-bitrate image compression for distributed automotive perception: Higher peak snr does not mean better semantic segmentation," in *2019 IEEE Intelligent Vehicles Symposium (IV)*, 2019, pp. 424–431.

[16] G. Lu, W. Ouyang, D. Xu, X. Zhang, C. Cai, and Z. Gao, "Dvc: An end-to-end deep video compression framework," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2019-June, 2019, pp. 10 998–11 007.

[17] A. Sankisa, A. Punjabi, and A. K. Katsaggelos, "Video Error Concealment Using Deep Neural Networks," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 380–384.

[18] P. H. Chan, A. Huggett, G. Souvalioti, P. Jennings, and V. Donzella, "Influence of AVC and HEVC compression on detection of vehicles through Faster R-CNN," *Submitted to IEEE Transactions on Intelligent Transportation Systems*, 2022.

[19] J. Huang, V. Rathod, C. Sun, M. Zhu, A. Korattikara, A. Fathi, I. Fischer, Z. Wojna, Y. Song, S. Guadarrama, and K. Murphy, "Speed/accuracy trade-offs for modern convolutional object detectors," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3296–3305.

[20] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015, pp. 91–99.