Video Coding for Machines: Large-Scale Evaluation of Deep Neural Networks Robustness to Compression Artifacts for Semantic Segmentation

Alban Marie, Karol Desnos, Luce Morin, Lu Zhang

Video Coding for Machines: Large-Scale Evaluation of Deep Neural Networks Robustness to Compression Artifacts for Semantic Segmentation

Alban Marie, Karol Desnos, Luce Morin and Lu Zhang
Univ Rennes, INSA Rennes
CNRS, IETR - UMR6164
Rennes, France
{alban.marie, karol.desnos, luce.morin, lu.ge}@insa-rennes.fr

***Abstract*—In the Video Coding for Machines (VCM) context where visual content is compressed before being transmitted to a vision task algorithm, appropriate trade-off between the compression level and the vision task performance must be chosen. In this paper, a Deep Neural Networks (DNN) based semantic segmentation algorithm robustness to compression artifacts is evaluated with a total of 1486 different coding configurations. Results indicate the importance of using an appropriate image resolution to overcome the block-partitioning limitations in existing compression algorithms, allowing 58.3%, 49.8%, 33.5% and 24.3% bitrate savings at equivalent prediction accuracy for JPEG, JM, x265 and VVenC, respectively. Surprisingly, JPEG can achieve 73.41% bitrate reduction with the inclusion of compressed images at training time over VVC Test Model (VTM) with a DNN trained on pristine data, which implies that DNN generalization ability must not be overlooked.**

***Index Terms*—Video Coding for Machines, Machine-to-Machine communication, Semantic Segmentation**

I. Introduction

Conventional image and video coding aims at achieving an optimal trade-off between bitrate and perceived quality by human observers. Many compression standards have been proposed to fulfill this purpose, such as Joint Photographic Experts Group (JPEG), Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC) or Versatile Video Coding (VVC). However, with the emergence of Machine-to-Machine (M2M) communications, the receiver of visual content is no longer necessarily human. According to Cisco [6], the total amount of M2M connections increased exponentially from 1 to 3.9 billions in the last five years. Furthermore, nearly 80% of the world's total bandwidth is used for image and video transmission. In 2019, a bitstream standardization group called Video Coding for Machines (VCM) was created by the Motion Picture Expert Group (MPEG) in order to address M2M transmissions of multimedia contents [35], where the main objective is to achieve greater trade-offs between bitrate and vision task performance compared to the VVC Test Model (VTM).

In order to address a M2M communication scenario, one could use an encoder expressly designed for a machine receiver, or a conventional encoder originally designed for transmission to a human receiver. While older compression standards such as JPEG or AVC are still wide-spread in embedded systems nowadays, it is reasonable to evaluate their relevance in a M2M communication context. In this paper, we propose to assess the suitability of conventional encoder in a VCM context. Specifically, the robustness to compression artifacts of a Deep Neural Networks (DNN) semantic segmentation algorithm is evaluated on 1486 different coding configurations. Coding configurations include multiple encoders, image quality and resolution, with and without grayscale conversion. The evaluation is performed using a novel progressive training strategy to enhance DNN robustness to various artifacts.

Our work is presented as follows. Section II reviews existing works in the literature. The considered evaluation protocol is detailed in Section III. Experimental results are introduced and discussed in Section IV, followed by a conclusion.

II. Related Work

In order to reduce the amount of information to transmit over a M2M connection, redundant information in visual content must be discarded. Therefore, evaluating DNN robustness to artifacts is crucial in the context of VCM. It has been shown that DNN solving vision tasks such as classification, object detection or segmentation are highly affected by distortion in visual data [13], [19], [23].

Many papers in the literature propose to evaluate DNN robustness to compression artifacts. In order to evaluate which information is relevant in a VCM scenario, related works can be organised in accordance to the following criteria: (I) DNN are re-trained on compressed data to achieve optimal trade-offs between bitrate and vision task performance. (II) Multiple encoders are evaluated. (III) Images are encoded at multiple resolutions. (IV) Experiments are conducted on grayscale images. (V) A high number of coding configurations are considered, which implies to take into account a wide range of image quality.

Table I compares related works in the literature according to the aforementioned criteria. Note that most related works do not attempt to perform a large-scale evaluation of DNN

978-1-6654-7189-3/22/$31.00

TABLE I

RELATED WORK. (I) COMPRESSED DATA AT TRAINING TIME. (II) MULTIPLE CODECS. (III) MULTIPLE IMAGE RESOLUTIONS. (IV) CHROMINANCE DEGRADATION. (V) TOTAL NUMBER OF CODING CONFIGURATION.

	[12]	[10]	[25]	[30]	[22]	[20]	[21]	[27]	[1]	[3]	[14]	[34]	[16]	[31]	[18]	[26]	[24]	[15]	[17]	[19]	[28]	Ours
(I)					✓	✓		✓				✓					✓	✓			✓	✓
(II)	✓	✓	✓			✓		✓			✓				✓	✓				✓	✓	✓
(III)		✓			✓		✓			✓						✓						✓
(IV)		✓												✓								✓
(V)	21	432	19	8	8	16	10	69	12	25	20	8	4	24	46	28	14	6	4	12	9	1486

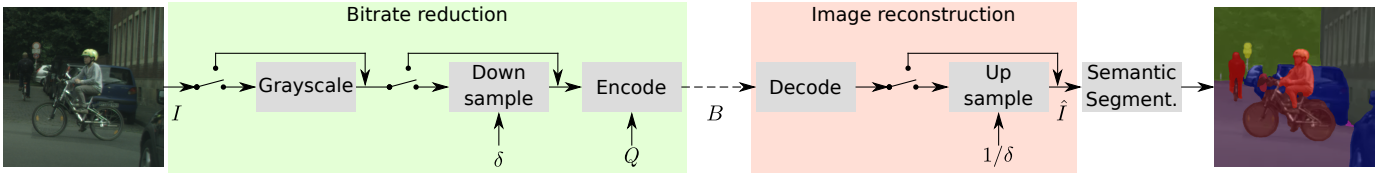

Fig. 1. Pipeline used to evaluate DNN resilience to compression artifacts. An image I, with or without chrominance channels is first downsampled by a factor δ and encoded with a codec using quality Q. Compressed image $\hat{I}$ is obtained by decoding bitstream B and upsampling reconstructed image by a factor $1/\delta$. It is then given to the semantic segmentation algorithm to obtain the prediction.

resilience to compression artifacts, but were included because of their proximity to this study. To the best of our knowledge, there is no related work that meet simultaneously all the mentioned requirements.

Regarding (V) the number of considered distortions, existing studies consider at most 50 coding configurations, at the exception of Dejean-servieres et al. [10]. This is because some studies only compare themselves to HEVC test Model (HM) or VVC Test Model (VTM) with few Quantization Parameter (QP) after proposing a new method to reach better trade-offs between vision task performance and bitrate [14]–[17], [19], [25], [26]. Some papers also evaluate DNN resilience to JPEG/JPEG2000 compression [10], [12], [15], [18], [21], [24], [27], [28], [30], AVC [1], [31] or auto-encoders [18], [26], [27], but no paper consider all mentioned image and video codec generations in a unified framework (II). Note that older codecs such as JPEG or AVC achieve lower trade-offs between bitrate and vision task performance, but their low-complexity compared to modern codecs makes them more suitable to some applications using low-power devices [31], especially when hardware implementation of AVC encoders is still widespread nowadays. Very few papers consider other hyper-parameters than QP such as (III) lowering image resolution [3], [10], [22], [26] or (IV) removing color information [10], [31].

The use of large-scale datasets is also criteria of major importance. A great amount of studies include experiments on large-scale datasets such as ImageNet [11] or Cityscapes [8], allowing general conclusion to be drawn with higher confidence. Note that Dejean-servieres et al. [10] did not address this point, since a subset of 55 images of the original ImageNet dataset were used, while the original database contains over 50000 validations images.

Few papers attempted to (I) use compressed data at training time in order to improve DNN resilience to compression artifacts [15], [20], [22], [24], [27], [28], [34]. All studies converge to show that adding compressed data at training time allows to reach much higher trade-offs between bitrate and vision task performance. While training and evaluation with the same encoding configuration allows to reach the best trade-offs, training and evaluating with different codec is still beneficial for DNN resilience [15], [27], [28].

Note that considering (I) compressed data at training time while using (V) a large range of coding configurations implies tremendous increase of computational resources. In Section III-B, we define explicitly the used training procedure to reduce training complexity, which allows us to consider more coding configurations, while achieving better trade-offs between bitrate and vision task performance.

III. BENCHMARK METHODOLOGY

A. Considered Coding Configurations

In this section, the method used to evaluate DNN robustness to compression artifacts is defined explicitly. Figure 1 illustrates the pipeline used to perform the benchmark that fulfill criteria presented in Section II. An image I is first captured by a camera. Depending on the coding configuration, criteria (IV) is considered by discarding or not image I chrominance channels according to the ITU-R BT.601 standard. Resulted image is then downsampled by a factor δ and compressed with one of the considered codecs, using quality Q. Bitstream B can then be sent to the vision task side using minimal bitrate since most of the irrelevant information was discarded. Reconstructed image $\hat{I}$ is obtained after decoding bitstream B and applying an upsampling by a factor $1/\delta$. Finally, reconstructed image $\hat{I}$ can be fed to a vision task algorithm. Note that artifacts may have been introduced in $\hat{I}$ because of grayscale conversion, downsampling and encoding steps. Therefore, in order to fulfill criteria (I), the used DNN was trained with images containing such artifacts, as explained in Section III-B.

Criteria (II) is met by considering a total of 5 compression algorithms, namely JPEG, JPEG2000, JM, x265 and

TABLE II
NUMBER OF CONSIDERED CODING CONFIGURATIONS PER CODEC.

JPG	JPG2K	JM	x265	VVenC	**Total**
672	352	154	154	154	**1486**

VVenC [4]. JPEG and JPEG2000 are two of the most widely used lossy image codecs. JM [32], x265 and VVenC are AVC, HEVC and VVC based video codecs, respectively. Since mentioned datasets are composed of still images, ALL-Intra configuration is used for these three video codecs. Note that HM [33] and VTM [2] are not used in this study, since they are not meant to be used in a real world scenario due to their extreme complexity. Therefore, lower complexity codecs such as x265 and VVenC were used instead. JM-19.0 and VideoLAN organisation implementation for x265 [29] are used. Slow and fast preset are selected for x265 and VVenC, respectively. For JPEG2000, a single tile is considered for the whole image. Because JPEG2000 do not employ any block-partitioning within a tile, image downsampling is not considered with JPEG2000 as it would not bring any gains in terms of rate-distortion trade-off.

Criteria (III) is also satisfied by taking into account a wide range of downsampling factor δ, where δ refers to the factor by which the width and height of the image are multiplied. A total of $\#\delta = 7$ image resolutions using bicubic interpolation are evaluated, with $\delta \in \{0.25, 0.375, 0.5, 0.625, 0.75, 0.875, 1.0\}$. A report of considered coding configurations is summarized in Table II. With a total number of 1486 distortions, the last criteria (V) is fulfilled.

B. Progressive Training Procedure

DNN are not inherently resilient to degradation in visual content, including compression artifacts. In order to achieve optimal trade-off between bitrate and vision task performance, the inclusion of compressed data at training time must be considered (criterion (I)).

For this purpose, we propose to use a progressive training, behind which the key idea is to train one DNN on a large amount of distortion at once, ranging from undistorted to highly distorted. This is done by increasing at training time the distortion strength *progressively*, where the distortion strength parameter start at p_0 and end at p_∞. The distortion strength can be characterised by the downsampling factor δ, or by the parameter allowing to control the amount of quantification in each codec, such as QP for AVC, HEVC or VVC based codecs. At each epoch e, the distortion strength parameter p is equal to $f(e)$, which is determined by the following equation:

$$f(e) = p_\infty + \Delta p \lfloor \frac{1}{\Delta p}(p_0 - p_\infty)\exp(-se) \rceil \qquad (1)$$

where $s \in \mathbb{R}^{+*}$ controls the speed at which the distortion level p converges towards p_∞, and where $\Delta p \in \mathbb{R}^{+*}$ refers to the step size between two consecutive distortion level p. The intuition behind our progressive training is that achieving high accuracy on images with distortion strength $f(e)$ is easier if the DNN is already robust to images of slightly higher qualities $f(e-1)$, and so on. Given that converging to a minima of the DNN loss function is increasingly harder as the distortion level p increases, an exponential decay function is used to decrease the pace at which image quality is reduced as the training progresses. In order to obtain model weights for adaptive training initialization, note that the DNN is trained on uncompressed pristine images before.

TABLE III
PROGRESSIVE TRAINING HYPER-PARAMETERS USED FOR EXPERIMENTS.

	p_0	p_∞	Δp	s	$\#e$
JPG	100	1	1	0.05	65
JPG2K	0	500	1	0.015	200
JM	0	50	5	0.04	70
x265	0	50	5	0.04	70
VVenC	0	60	5	0.025	72

Listed coding configurations in Table II vary across distortions introduced by different types of processing, such as grayscale conversion, image downsampling or compression artifacts. Therefore, multiple progressive training are interlaced together to train the DNN on all considered coding configurations. First, a progressive training with varying downsampling factor δ is done. This training allows us to obtain model weights θ_{δ_i} for each considered δ_i by taking DNN model weights at last epoch e where $f(e) = \delta_i$. Afterwards, model weights θ_{δ_i} can be used at initialization for other progressive training with varying compression strength, where each images are downsampled by a factor θ_{δ_i}. Such training with varying compression strength is done for every image resolution on JPEG, JM, x265 and VVenC codecs, resulting in a total of $4 \times \#\delta = 28$ progressive trainings.

The described training strategy can be done both on color and grayscale images. For color images, DNN weights trained on pristine data are used as initialization for the progressive training with varying image resolutions. For grayscale, the same model is first fine-tuned on grayscale images before being used for initialization. At the end, a total of $2 \times (1 + 1 + 4 \times \#\delta) = 60$ progressive trainings are done to cover the 1486 coding configurations described in Table II. Hyper-parameters p_0, p_∞, Δp, s and the number of epochs $\#e$ used by each progressive training are given in Table III.

C. Used Dataset and Architecture

In the context of this study, the Cityscapes [8] dataset is considered. Cityscapes is composed of urban landscapes represented with losslessly compressed images of resolution 2048×1024. The lowest image resolution considered in this study is therefore 512×256 with the downsampling factor $\delta = 0.25$.

DeepLabV3+ [5] with a ResNet50 backbone, a state-of-the-art semantic segmentation algorithm is used to evaluate DNN robustness to compression artifacts. MMSegmentation library [7] was used for DeepLabV3+ implementation, training, evaluation and model weights. For all training depicted in Section III-B, SGD optimizer with mini-batch size of 4, a learning rate of 10^4 and a polynomial decay of 0.9 are used. Note that the learning rate is resetted to 10^4 at each epoch

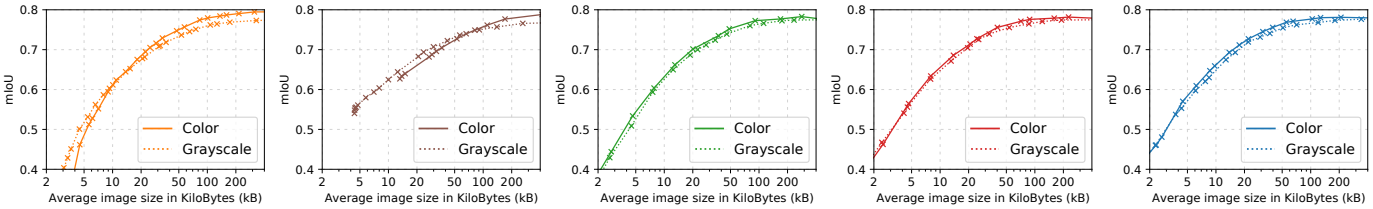

Fig. 2. Trade-offs between bitrate and mIoU with and without chrominance channels. From left to right, each plot correspond to JPEG, JPEG2000, JM, x265 and VVenC. Criteria (I) and (III) defined in Section II are considered for each curve.

TABLE IV
TRAINING COMPLEXITY AND BDR RELATIVE TO SEPARATE TRAINING (ST). LOWER BDR VALUES MEANS LOWER BITRATE AT EQUIVALENT SEMANTIC SEGMENTATION ACCURACY IN TERMS OF MIOU.

	Complexity↓	BDR↓
Baseline	—	486.57%
ST	100.00%	0.00%
DA	46.99%	21.92%
ours, $s = 0.085$	**26.51%**	16.44%
ours, $s = 0.045$	48.19%	9.51%
ours, $s = 0.025$	86.75%	**–2.62%**

e_i as long as $f(e_i - 1) \neq f(e_i)$, based on Eq. 1. The vision task performance is measured with the mean Intersection over Union (mIoU) metric.

IV. EXPERIMENTS

A. Progressive training evaluation

Firstly, we briefly evaluate the effectiveness of the progressive training procedure by comparing training complexity and reached accuracy on a small subset of coding configurations. 11 VVenC coding configurations are used for this experiment, with QP $Q \in \{5n | n \in \mathbb{Z}, 0 \leq n \leq 10\}$, with chrominance channels and no image downsampling. Multiple convergence speed s from Eq. (1) are considered, namely 0.085, 0.045 and 0.025. We compare our method to a Separate Training (ST) approach that perform one training per coding configuration. Final complexity of ST strategy is obtained by adding up the number of epochs required until converge for each coding configuration. In addition, Data Augmentation (DA) [15] training procedure is considered, which consists of training a single model on all coding configurations at once by selecting a coding configuration randomly for each image during the training. Progressive training, ST and DA uses model weights trained on raw images I at the initialization. Baseline refers to the model trained on undistorted data, which thus does not validate criteria (I) from Section II.

Results are presented in Table IV, where Bjøntegaard-Delta Rate (BDR) metric represents the average bitrate savings at equivalent DNN accuracy in term of mIoU. ST achieve good rate-mIoU trade-offs by considering criteria (I). However, it is at the cost of training complexity since one training has to be done for each coding configuration. DA training procedure mitigate this issue by training one model on multiple coding configurations at once, but it is at the cost of BDR, especially when a very diverse set of coding configurations is considered. Note that DA and ST could achieve comparable rate-mIoU trade-offs on a set of similar coding configurations. The proposed progressive training procedure allows optimal rate-accuracy trade-offs to be reached with a lower training complexity. Tuning the convergence speed s parameter from Eq. (1) allows to find a suitable balance between BDR values and training complexity. With a convergence speed of $s = 0.025$, we are able to reduce the training complexity over the ST procedure approach with better BDR values.

Based on selected parameters given in Table III, note than a total of 4278 epochs are required to cover the 1478 considered coding configurations using the proposed progressive training strategy. Therefore, to be comparable in terms of training complexity, the use of ST strategy instead of the proposed progressive training would have required less than 3 epochs per coding configuration on average, which is not enough to converge.

B. Coding Configuration Comparison

In this subsection, all considered coding configurations are compared together as depicted in Section III.

Table V presents BDR values with mIoU as the quality metric between every codec, either with or without the consideration of (III) all image resolutions with the convex hull. Note that image downsampling is not considered along with JPEG2000, as explained in Section III-A.

It can be noted that using appropriate image resolution is of major importance, saving 58.3%, 49.8%, 33.5% and 24.3% bitrate at equivalent prediction accuracy for JPEG, JM, x265 and VVenC, respectively. Block-partitioning is the main limitation of older codecs such as JPEG or JM to obtain better rate-accuracy trade-offs. As an example, the fixed 8×8 bloc size of JPEG may be unsuited to areas with low spatial information such as *sky* or *road*, especially on high resolutions images. If a downsampling is applied on the same image, a 8×8 bloc will correspond to a larger area, which could be more suited depending on the image content. Even from codecs such as x265 or VVenC that allows larger blocs to be selected at encoding, BDR savings can be obtained with the use of convex hull.

We suggest that using (III) appropriate image resolution is of greater importance compared to using a more complex codec in terms of rate-accuracy trade-off. When appropriate image resolution is used, it can be noted that JPEG, JM and x265 achieve 43.1%, 14.2% and 19.9% bitrate savings over JM, x265 and VVenC, respectively. Note that this observation cannot be extrapolated to datasets containing smaller images

TABLE V

BDR WITH MIOU METRIC COMPARISON. NEGATIVE VALUES REPRESENT BITRATE REDUCTION AT EQUIVALENT PREDICTION ACCURACY IN TERMS OF MIOU. SYMBOL * DENOTES THE USE OF (III) CONVEX HULL CURVES ACROSS CONSIDERED IMAGE RESOLUTIONS.

	JPG	JPG*	JPG2K	JPG2K*	JM	JM*	x265	x265*	VVenC	VVenC*
JPG	0.0	139.6	52.9	—	53.3	125.7	123.3	178.6	193.8	224.1
JPG*	−58.3	0.0	−35.1	—	−43.1	27.6	−2.7	68.6	26.8	73.0
JPG2K	−34.6	54.1	0.0	—	−13.1	32.8	29.6	62.2	72.0	88.4
JPG2K*	—	—	—	—	—	—	—	—	—	—
JM	−34.8	75.9	15.0	—	0.0	99.1	71.8	157.5	115.3	179.6
JM*	−55.7	−21.6	−24.7	—	−49.8	0.0	−14.2	34.6	4.3	37.1
x265	−55.2	2.7	−22.9	—	−41.8	16.5	0.0	50.3	24.7	63.0
x265*	−64.1	−40.7	−38.4	—	−61.2	−25.7	−33.5	0.0	−19.9	2.0
VVenC	−66.0	−21.1	−41.9	—	−53.5	−4.2	−19.8	24.9	0.0	32.1
VVenC*	−69.1	−42.2	−46.9	—	−64.2	−27.1	−38.7	−2.0	−24.3	0.0

TABLE VI

BEST ACHIEVABLE COMPRESSION RATIO WITH RESPECT TO A MINIMAL MIOU CONSTRAINT. CONSIDERED DNN ACHIEVE 0.8027 MIOU ON RAW IMAGES.

t_0	mIoU	JPG	JPG2K	JM	x265	VVenC
97.5%	0.783	15.9	5.6	N.A.	N.A.	N.A.
95.0%	0.763	33.2	18.8	30.2	39.9	43.4
90.0%	0.723	52.6	56.4	66.4	96.6	107.3
80.0%	0.642	172.0	188.9	182.3	231.9	273.9
70.0%	0.562	358.3	N.A.	313.0	507.7	529.3

TABLE VII

BDR VALUES USING MIOU COMPARED TO VTM ANCHOR USING JPEG ENCODER UNDER VARIOUS CODING CONFIGURATIONS. (I) COMPRESSED DATA AT TRAINING TIME. (III) MULTIPLE IMAGE RESOLUTIONS. (IV) CHROMINANCE DEGRADATION.

(I)	(III)	(IV)	BDR
			644.68%
✓			–4.06%
✓	✓		–73.41%
✓		✓	–25.65%
✓	✓	✓	–76.13%

or with higher spatial information, since larger bloc size may not be desirable in such contexts.

Figure 2 compares rate-mIoU trade-offs with and without chrominance information. Using grayscale images allows JPEG and JPEG2000 to achieve greater compromise at lower bitrates, while the same does not apply for JM, x265 and VVenC video codecs. Note that an accuracy drop is observable at high bitrate for video codecs. As shown by Fischer et al. [14], artifacts generated by in-loop filtering tend to worsen DNN performance, even with near lossless image quality.

Achievable compression ratio with respect to a minimal mIoU constraint is provided in Table VI. Multiple threshold t_0 are used to define the minimal acceptable mIoU score. Among all coding configurations that have greater mIoU than the constraint, the one with the lowest bitrate is selected to compute the highest achievable compression ratio with respect to the mIoU constraint. The average bitrate using PNG compression of 2362.90 kB per image in the original Cityscapes dataset is used as the anchor to compute the compression ratio.

C. Comparison with CTC

In order to assess the importance of using (I) compressed data at training time, (III) multiple image resolution and (IV) chrominance degradation, a comparison with VTM is provided as recommended by Common Test Conditions (CTC) from the VCM MPEG standardization group [9]. Convex hull of VTM-12.0 with ALL-Intra configuration is used as an anchor, where QP $\in 22, 27, 32, 37, 42, 47$ and downsampling factor $\delta \in \{0.25, 0.5, 0.75, 1.0\}$ are considered.

Table VII highlights JPEG BDR gains using mIoU metric compared to VTM anchor. As expected, JPEG performs poorly compared to VTM when criteria (I), (III) and (V) are not considered, which results in a bitrate increase of 644.68% at equivalent DNN performance in terms of mIoU. Surprisingly enough, the use (I) compressed images $\hat{I}$ at training time allows JPEG to outperform the VTM anchor with a bitrate reduction of 4.06%. Bitrate savings can be further increased to 73.41% if criteria (I) and (III) are satisfied jointly, which emphasizes the importance of criteria (III) for JPEG codec because of the fixed 8×8 bloc size. JPEG achieving BDR gains compared to the VTM anchor highlight the lack of resilience to compression artifacts that DNN can have if criteria (I) is overlooked. Note that this experiment was performed with a DNN model trained on Cityscapes dataset, which does not contain any compression artifacts. Less extreme results would be obtained with a dataset that includes compression artifacts, since artifacts created by different compression algorithms share similarities on which DNN are able to generalize [15], [28]. As shown by Figure 2, removing (IV) chrominance information is beneficial for JPEG at lower bitrates. Therefore, higher bitrate gains can be achieved by being able to remove colors. Nevertheless, criteria (IV) appears less detrimental compared to criteria (I) or criteria (III).

V. CONCLUSION

In this paper, we evaluate in the VCM context the impact of compression artifacts on deep based semantic segmentation algorithms at an unprecedented scale. A wide range of image degradations are considered in order to measure which distortions allows to maximize DNN performance at equivalent bitrate. Experiments showed that using appropriate image resolution is the most crucial parameter to achieve optimal rate-accuracy trade-off, achieving 58.3%, 49.8%, 33.5% and 24.3% bitrate savings at equivalent prediction accuracy for

JPEG, JM, x265 and VVenC, respectively. Significant bitrate reductions can also be obtained with newer codecs, but at the cost of encoding complexity. Surprisingly, VVenC achieve a very low bitrate saving over x265 of 2.00% with optimal image resolution, suggesting that the main limitation of older codecs is their limited block partitioning. Removing chrominance channels appears as an unsuitable strategy, as it can worsen DNN performance even at very low bitrates. At high bitrates, the poor generalization ability of DNN models to video codecs artifacts, such as in-loop filtering, makes them inferior to simpler codecs like JPEG in terms of rate-accuracy trade-off.

REFERENCES

[1] Miloud Aqqa, Pranav Mantini, and Shishir K Shah. Understanding How Video Quality Affects Object Detection Algorithms. In *VISIGRAPP (5: VISAPP)*, pages 96–104, 2019.

[2] Frank Bossen. H.266/VVC Software Coordination and VTM Reference Software, 2020.

[3] Imene Bouderbal, Abdenour Amamra, and Mohamed Akrem Benatia. How Would Image Down-Sampling and Compression Impact Object Detection in the Context of Self-driving Vehicles? In *CSA*, pages 25–37, 2020.

[4] Jens Brandenburg, Adam Wieckowski, Tobias Hinz, and Benjamin Bross. VVenC Fraunhofer versatile video encoder. *cit. on*, page 3, 2020.

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[6] V Cisco. Cisco visual networking index: Forecast and trends, 2017–2022. *White Paper*, 1(1), 2018.

[7] MMSegmentation Contributors. MMSegmentation: OpenMMLab Semantic Segmentation Toolbox and Benchmark, 2020.

[8] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The Cityscapes Dataset for Semantic Urban Scene Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[9] Igor Curcio. Common Test Conditions and Evaluation Methodology for Video Coding for Machines. *ISO/IEC JTC 1/SC 29/WG 2, MPEG Technical requirements, Finland*, 2022.

[10] Mathieu Dejean-Servières, Karol Desnos, Kamel Abdelouahab, Wassim Hamidouche, Luce Morin, and Maxime Pelcat. *Study of the impact of standard image compression techniques on performance of image classification with a convolutional neural network*. PhD Thesis, INSA Rennes; Univ Rennes; IETR; Institut Pascal, 2017.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[12] S. Dodge and L. Karam. Understanding how image quality affects deep neural networks. In *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6, 2016.

[13] Samuel Dodge and Lina Karam. Human and DNN classification performance on images with quality distortions: A comparative study. *ACM Transactions on Applied Perception (TAP)*, 16(2):1–17, 2019. Publisher: ACM New York, NY, USA.

[14] 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)*, pages 1147–1151, 2020.

[15] Kristian Fischer, Christian Blum, Christian Herglotz, and André Kaup. Robust Deep Neural Object Detection and Segmentation for Automotive Driving Scenario with Compressed Image Data. In *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 1–5, 2021.

[16] Kristian Fischer, Fabian Brand, Christian Herglotz, and André Kaup. Video Coding for Machines with Feature-Based Rate-Distortion Optimization. In *2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2020.

[17] Kristian Fischer, Felix Fleckenstein, Christian Herglotz, and André Kaup. Saliency-Driven Versatile Video Coding for Neural Object Detection. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1505–1509, 2021.

[18] Kristian Fischer, Christian Forsch, Christian Herglotz, and André Kaup. Analysis Of Neural Image Compression Networks For Machine-To-Machine Communication. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 2079–2083, 2021.

[19] Kristian Fischer, Markus Hofbauer, Christopher Kuhn, Eckehard Steinbach, and André Kaup. Evaluation of Video Coding for Machines without Ground Truth. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1616–1620, 2022.

[20] S. Ghosh, R. Shet, P. Amon, A. Hutter, and A. Kaup. Robustness of Deep Convolutional Neural Networks for Image Degradations. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2916–2920, 2018.

[21] Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *CoRR*, abs/1903.12261, 2019. _eprint: 1903.12261.

[22] Suresh Prasad Kannojia and Gaurav Jaiswal. Effects of varying resolution on performance of CNN based image classification: An experimental study. *Int. J. Comput. Sci. Eng*, 6(9):451–456, 2018.

[23] Samil Karahan, Merve Kilinc Yildirum, Kadir Kirtac, Ferhat Sukru Rende, Gultekin Butun, and Hazim Kemal Ekenel. How Image Degradations Affect Deep CNN-Based Face Recognition? In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*, pages 1–5, 2016.

[24] Karim El Khoury, Martin Fockedey, Eliott Brion, and Benoît Macq. Improved 3D U-Net robustness against JPEG 2000 compression for male pelvic organ segmentation in radiotherapy. *Journal of Medical Imaging*, 8(4):1–20, April 2021.

[25] L. Kong and R. Dai. Object-Detection-Based Video Compression for Wireless Surveillance Systems. *IEEE MultiMedia*, 24(2):76–85, 2017.

[26] Nam Le, Honglei Zhang, Francesco Cricri, Ramin Ghaznavi-Youvalari, and Esa Rahtu. Image Coding For Machines: an End-To-End Learned Approach. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1590–1594, 2021.

[27] J. Löhdefink, A. Bär, N. M. Schmidt, F. Hüger, 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)*, pages 424–431, 2019.

[28] Alban Marie, Karol Desnos, Luce Morin, and Lu Zhang. Expert Training: Enhancing AI Resilience to Image Coding Artifacts. In *Electronic Imaging, Image Processing: Algorithms and Systems XX*, San Francisco, United States, January 2022.

[29] VideoLAN organisation. x265 software library, 2013.

[30] Prasun Roy, Subhankar Ghosh, Saumik Bhattacharya, and Umapada Pal. Effects of Degradations on Deep Neural Network Architectures. *CoRR*, abs/1807.10108, 2018. _eprint: 1807.10108.

[31] Benno Stabernack and Fritjof Steinert. Architecture of a Low Latency H.264/AVC Video Codec for Robust ML Based Image Classification. In *Workshop on Design and Architectures for Signal and Image Processing (14th Edition)*, DASIP '21, pages 1–9, New York, NY, USA, 2021. Association for Computing Machinery. event-place: Budapest, Hungary.

[32] Karsten Suehring. H.264/AVC Software Coordination JM Reference Software, 2003.

[33] Karsten Suehring. H.265/HEVC Software Coordination and HM Reference Software, 2013.

[34] Suvash Sharma, Christopher Hudson, Daniel Carruth, Matt Doude, John E. Ball, Bo Tang, Chris Goodin, and Lalitha Dabbiru. Performance analysis of semantic segmentation algorithms trained with JPEG compressed datasets. volume 11401, April 2020.

[35] Y Zhang and P Dong. MPEG-M49944: Report of the AhG on VCM. *Moving Picture Experts Group (MPEG) of ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, Tech. Rep.*, 2019.