# Evaluation of Image Quality Assessment Metrics for Semantic Segmentation in a Machine-to-Machine Communication Scenario

Alban Marie, Karol Desnos, Luce Morin, Lu Zhang

# Evaluation of Image Quality Assessment Metrics for Semantic Segmentation in a Machine-to-Machine Communication Scenario

Alban Marie, Karol Desnos, Luce Morin and Lu Zhang
*Univ Rennes, INSA Rennes*
*CNRS, IETR – UMR 6164*
F-35000 Rennes, France
{alban.marie, karol.desnos, luce.morin, lu.ge}@insa-rennes.fr

***Abstract*—Image and video compression aims at finding an optimal trade-off between rate and distortion. This is done through Rate-Distortion Optimization (RDO) in traditional encoders with the use of Image Quality Assessment (IQA) metrics. While it is known that most IQA metrics are designed to be correlated with human perception, there is no evidence that this observation can be generalized in a Video Coding for Machines (VCM) context, where the receiver is not a human anymore but a machine. In this paper, we propose an evaluation protocol to measure the correlation level between conventional Full-Reference (FR) IQA metrics and machine perception through the semantic segmentation vision task. Experiments showed a relatively low correlation between them when measured on the block-level. This observation implies the need of RDO algorithms that are better suited for Machine-to-Machine (M2M) communications. In order to facilitate the emergence of IQA metrics that better reflect machine perception, the code and dataset used to perform this study is made freely available at https://github.com/albmarie/iqa_m2m_segmentation.**

***Index Terms*—Image Quality Assessment (IQA), Video Coding for Machines (VCM), Machine-to-Machine (M2M), block-based, compression, Rate-Distortion Optimization (RDO)**

## I. Introduction

For decades, large gains in image and video coding efficiency have been accomplished with modern compression standards such as Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC) or Versatile Video Coding (VVC). The pursued goal is to achieve an optimal trade-off between the rate and the quality as perceived by the Human Visual System (HVS). However, new kind of transmissions known as Machine-to-Machine (M2M) communications is growing exponentially, with a 4-fold increase in the last 5 years [3]. In 2023, M2M connections represents half of the global connected devices and connections. The objective of visual content compression in the M2M context is now to preserve the quality as perceived by the machines, which refers to the vision task performance of machines. To this end, the Motion Picture Expert Group (MPEG) created the Video Coding for Machines (VCM) group in 2019 to propose a bitstream standardization in the context of M2M transmissions [36].

In conventional compression for HVS, Rate-Distortion Optimization (RDO) is employed, where the distortion level can be quantified by comparing the distorted image to its corresponding pristine reference using Full-Reference (FR) Image Quality Assessment (IQA) metrics. Two of the most common ones are Peak Signal to Noise Ratio (PSNR) and Structural SIMilarity (SSIM) [29]. While existing IQA metrics have shown good correlation with human perception measured by the Mean Opinion Score (MOS) on multiple large-scale IQA datasets [13], [20], there is no evidence that this observation generalizes to machine perception. A low correlation between IQA metrics and vision task algorithms performance would imply at the very least a sub-optimality in the optimization problems using such metrics as a distortion measure in a VCM context.

This paper aims to assess the relevance of existing FR IQA metrics used in RDO methods for the VCM context. More precisely, the correlation between conventional metrics and machines performance when images are subject to various compression artifacts is evaluated. We also propose to measure the correlation on a block-level, since RDO based encoders determine the best encoding modes according to a FR metric for each block separately. To the best of our knowledge, no existing works in the literature have evaluated the correlation between existing metrics and machine perception. Conflicting conclusions could be drawn from the state of the art, as some studies might suggest that the correlation would be high while other studies might indicate the opposite. On top of performing a deep quantitative evaluation, the built dataset used to perform this study is made freely available to facilitate the emergence of novel IQA metrics that better reflect machine perception.

Our work is presented as follows. Related works are presented in Section II. Section III presents the evaluation protocol. Experimental results are presented and discussed in Section IV. Finally, Section V concludes the paper.

## II. Related Work

Image Quality Assessment (IQA) is a research field that aims at finding visual content quality models that match the HVS perception. One metric group, referred to as FR metrics, consists of measuring the degradation within an image compared to a pristine reference image. As it was observed that the

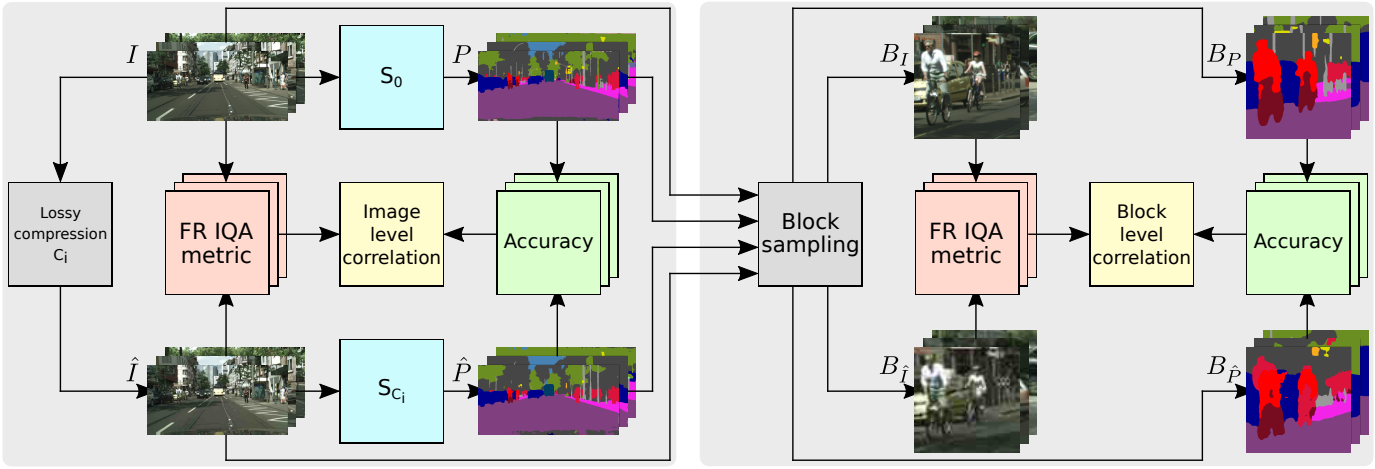



Fig. 1: Used pipeline to compute Image-level (left) and Block-level (right) correlation between conventional metrics and accuracy values pairs. $S_0$ denotes a segmentation algorithm where model weights were trained using pristine images. Models $S_{C_i}$ are obtained by using images compressed by the lossy compression scheme $C_i$ at training, as described in Section III-A. Note the presence of pseudo Ground Truth (GT) $P$ and not the real GT to compute accuracy.

TABLE I: Considered hyper-parameters for the lossy compression scheme.

| Downsampling factors | JPEG qualities | JM, x265 and VVenC QPs |
|---|---|---|
| 0.25 | 5, 7, 10, 15 | 0, 5, 10, 15 |
| 0.5, 0.75 | 20, 25, 30, 35 | 20, 25, 30, 35 |
| 1.0 | 50, 70, 90 | 40, 45, 50 |

legacy PSNR metric lacks correlation with human perception, many IQA metrics have been proposed over the years. These metrics include similarity based metrics such as the SSIM [29] index and its multi-scale variant MS-SSIM [28], Feature SIMilarity (FSIM) [34], Spectral Residual based SIMilarity (SR-SIM) [32]. More recently, Learned Perceptual Image Patch Similarity (LPIPS) [35] Deep Image Structure and Texture Similarity (DISTS) [6], which are deep learning based metrics, have shown good performance on many IQA databases such as TID2013 [20], CSIQ [13] or LIVE [24].

While it is well established that existing IQA metrics correlate well with human perception, few studies related to machine perception have been conducted. Fischer et al. [8] proposed to replace default distortion metric in VVC Test Model (VTM) by a learning based feature extractor and they showed that $5.49\%$ bitrate saving can be achieved at equivalent vision task accuracy compared to VTM anchor. Other work [9], [14] showed that remarkable bitrate reduction of around $40\%$ is achievable over VTM at equivalent vision task performance. This is done by training in a end-to-end fashion an auto encoder followed by a network performing a vision task where the objective is to jointly minimize the prediction error and the rate. Studies related to the concept of *utility* [12], [22] found that a metric optimized to predict quality scores might not be able to predict utility scores accurately. On another side, Leszczuk et al. showed that a set of IQA metrics are able to predict with high confidence the performance of machine in a face recognition task [16] and a license plate recognition task [15], indicating that correlation between existing metrics and machines performance may be high.

## III. Evaluation Protocol

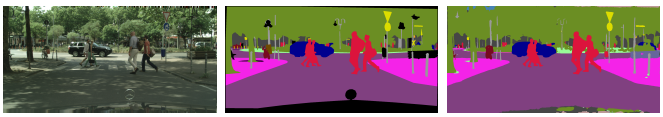
This section introduces the method used to evaluate the correlation of FR IQA metrics with the accuracy of Deep Neural Network (DNN) performing a vision task algorithm.

### A. Built dataset

Figure 1 presents the pipeline that is used to measure the correlation between IQA metrics and DNN performance.

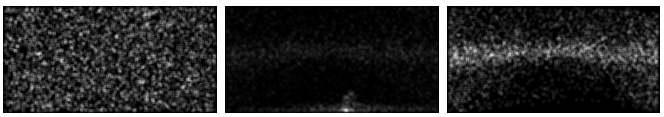
First, a set of pristine images $I$ must be selected to perform the evaluation. The $\#D = 500$ validation images from the Cityscapes [5] dataset are considered in this study, which contains urban landscapes as seen from a driving car perspective.

In the context of VCM, it is desirable to send a minimal amount of information through M2M transmission. Transmitted images are thus compressed images $\hat{I}$, containing compression artifacts compared to raw images $I$. Thus, a lossy compression scheme $C$ is used to obtain compressed images $\hat{I}$ from pristine images $I$. We refer to a lossy compression scheme as a scheme involving the following steps: image downsampling, encoding, decoding and upsampling back to original image resolution. As it has been shown that image downsampling was a crucial part to obtain optimal trade-off between rate and DNN performance [14], [18], we include it in the lossy compression scheme through bicubic interpolation under $4$ different downsampling factors. JPEG, JM, x265 and VVenC lossy compression algorithms, each with $11$ different levels of quantification are applied. For JM, x265 and VVenC that are respectively AVC, HEVC and VVC based video encoders, all-intra configuration is used as the Cityscapes dataset is composed of still images. JM-19.0 [26], VideoLAN organisation implementation for x265 [19] with preset *slow* and VVenC [1] with preset *fast* are used. In total,



(a) (b) (c)

Fig. 2: (a) Example of an original image $I$, (b) associated real GT and (c) pseudo GT $P$.



(a) (b) (c)

Fig. 3: Probability density functions of $2^{12}$ blocks of size $32 \times 32$ using (a) uniform random distribution, proposed block sampling (b) without exclusion of blocks containing unlabeled pixels in real GT and (c) with exclusion. Brighter areas indicates areas that happened to be within a large number of randomly drawn blocks.

$\#C = 4 \times 4 \times 11 = 176$ coding configurations for the lossy compression scheme are involved to encode all $\#D = 500$ images from the dataset. More details are provided in Table I for hyper-parameters such as downsampling factors, JPEG qualities, and JM/x265/VVenC Quantization Parameter (QP) within each codec.

In this study, a DNN performing semantic segmentation vision task is considered as a measure of machine perception in a VCM context. Since pixel-wise predictions are obtained through the segmentation vision task, correlation measurement can be performed on a local scale as depicted is Section III-E. This does not apply for other vision tasks such as classification or object detection where some image areas may not contain any object. DNN model DeepLabV3+ [2] with a ResNet50 [11] backbone is employed for this purpose, using MMSegmentation [4] implementation. We denote the model trained with pristine image $I$ and GT labels as $S_0$. Once trained, pseudo GT predictions $P$ can be obtained by inputting the $\#D$ validation images $I$ to the DNN model $S_0$.

As shown by the literature, a DNN trained on losslessly compressed images $I$ such as original Cityscapes dataset generalizes poorly to compressed images $\hat{I}$, as DNN would encounter artifacts that were not present at training time [7], [17], [18]. To mitigate this bias, progressive training [18] is employed to obtain segmentation models $S_{C_i}$ that are resilient to artifacts generated from coding configuration $C_i$, $i \in \{1, 2, 2 \ldots, \#C\}$. In a nutshell, progressive training allows training one DNN model on multiple coding configurations at once by *progressively* strengthening the degradation level as training progresses. Once trained, predictions on compressed images $\hat{P}$ are obtained by inputting the $\#D$ validation images $I$ to $S_{C_i}$. Note that the design choice of having separate DNN weights for each coding configuration is a limitation in a real world application. However, this design choice does not introduce any bias about the DNN ability to generalize on broad set of distortions.

### B. Considered IQA metrics

Multiple FR IQA metrics are considered in this study, namely PSNR, SSIM [29] and its multiscale variant MultiScale Structural SIMilarity (MS-SSIM) [28], FSIM [34], SR-SIM [32], Gradient Magnitude Similarity Deviation (GMSD) [30] and its multiscale variant MultiScale Gradient Magnitude Similarity Deviation (MS-GMSD) [31], Visual Saliency based Index (VSI) [33], Haar Perceptual Similarity Index (HaarPSI) [21], Mean Deviation Similarity Index (MDSI) [37], LPIPS [35] and DISTS [6]. For LPIPS, the VGG [25] network is used. Metrics are computed using grayscale component for both image and block level. PIQ and PIQA [23] libraries are considered for IQA metrics implementation.

### C. Machine perception measure

To measure DNN performance, one could simply use the mean Intersection over Union (mIoU) score, as it is one of the reference DNN performance measure for segmentation. However, mIoU comes with several limitations, especially for the block-level experiment where blocks may contains very few classes if not only one. mIoU is computed by averaging the IoU of each class within an image. IoU is given by the formula:

$$\text{IoU} = \frac{p \cap t}{p \cup t} \tag{1}$$

where $p$ and $t$ are general notations that represents a prediction image and a true label image, respectively. Consider an extreme case where a true label block contains only one class e.g. *road*, except for one pixel belonging to another class, such as *car*. In this example where the block contain a total of $N^2$ pixels, predicting the class *road* for the whole block would give an IoU of $\frac{N^2-1}{N^2}$ for the class *road*, and an IoU of 0 for the class *car*. The final mIoU would be $\frac{N^2-1}{2N^2} \approx 1/2$. Note that the mIoU is far from a perfect score of 1, even though only one pixel was misclassified. Similar case can occur since predictions $\hat{P}$ tends to be unstable on object edges when $\hat{I}$ is subject to various artifacts. Therefore, instead of mIoU, pixel-wise accuracy $\in [0, 1]$ is used as the DNN performance measure to measure correlation with IQA metrics. In the above example, accuracy would be $1 - 1/N^2$, which is close to 1.

Accuracy can be computed by comparing prediction $\hat{P}$ with a reference denoted as real GT. However, it should be noted that predictions $P$ on losslessly compressed images $I$ may differ from real GT, whereas a IQA metric would return a score indicating no degradation if an image $I$ is compared to itself. In order to mitigate the bias where DNN prediction on pristine images $I$ are not always equal to real GT, pseudo GT is used as proposed by Fischer et al. [10]. Pseudo GT consists of using DNN prediction on pristine data $I$ as a GT instead of using the real GT at evaluation. Figure 2 illustrates the difference between real GT and pseudo GT for one given image. Using pseudo GT ensures that a score indicating no degradation is obtained for the DNN performance when

an undistorted image $I$ is fed to the vision task algorithm, similarly to FR IQA metrics.

Correlation between IQA metrics and DNN performance is measured with Spearman Rank-Order Correlation Coefficient (SROCC), Pearson Linear Correlation Coefficient (PLCC) and Kendall Rank-Order Correlation Coefficient (KROCC). The correlation measurement is also performed both on image-level and block-level, as described in Section III-D and Section III-E.

### D. Image Level

On the image-level, the whole images $I$ and $\hat{I}$ are used to compute FR IQA metrics, and the whole pseudo GT $P$ and prediction $\hat{P}$ are used to compute accuracy. For this experiment, conventional metrics and accuracy on all $\#D$ images across the dataset are computed exhaustively, using all $\#C$ compression schemes as described in Section III-A. Therefore, correlation is measured on a total of $\#C \times \#D$ pairs of values.

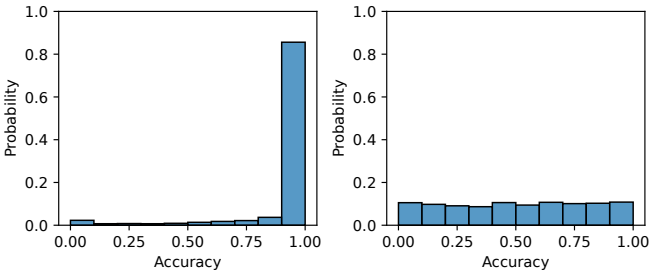
### E. Block Level

When image or video is encoded using RDO, conventional metrics are not computed on the whole image $I$ and $\hat{I}$, but rather on smaller images sections denoted as *blocks* [27]. Ultimately, an optimal trade-off between the rate and the quality as perceived by the used metric is found. For this reason, a block-level experiment is conducted on a subpart of images $I$, $\hat{I}$, $P$ and $\hat{P}$. Using a block sampling algorithm explained below, correlation is then measured on blocks $B_I$, $B_{\hat{I}}$, $B_P$ and $B_{\hat{P}}$. A total of $\#B$ blocks for each considered block size is considered for this experiment.

The proposed block sampling algorithm aims at drawing the same number of blocks with low and high accuracy of a fixed size $N \times N$. Let $k \in \mathbb{N}$ be a parameter used to divide the possible set of accuracy values $\in [0, 1]$ into $k$ disjoint sets such that the $j^{th}$ subset with $0 \leq j < k$ corresponds to the subset $[\frac{j}{k}, \frac{j+1}{k}[$ of accuracy scores, except for $j = k-1$ where $\frac{j+1}{k} = 1$ is included in the subset. First, let $j$ be the $j^{th}$ subset drawn with a uniform probability distribution function out of the $k$ subsets. Then, an image $I$ and a coding configuration $C_i$ is drawn with a uniform probability distribution among the $\#D$ images from the dataset and the $\#C$ coding configurations, respectively. Once corresponding pseudo GT $P$ and prediction $\hat{P}$ are obtained, a mask $M$ is built such that:

$$M(x,y) = \begin{cases} 1 & \text{if } P(x,y) = \hat{P}(x,y) \\ 0 & \text{else} \end{cases} \tag{2}$$

where $x$ and $y$ refers to coordinates in the image domain. In order to count the number of correctly classified pixels in a $N \times N$ window around each pixel, an image $\mathcal{S} = M * K$ is obtained by performing a convolution on mask $M$ with a separable kernel $K = \mathbb{1}_{N \times N}$ containing only ones. Let $I_W$ and $I_H$ be the image width and height in pixel respectively. Based on $\mathcal{S}$, we can determine a set $\mathcal{V}$ of valid block coordinates $x \in [0, I_W[, y \in [0, I_H[$ with an accuracy in the $j^{th}$ subset, such that $(x, y) \in \mathcal{V}$ if:




(a) (b)

Fig. 4: Obtained accuracy scores for $2^{12}$ blocks of size $32 \times 32$ using (a) uniform random and (b) proposed block sampling.

$$\begin{cases} \mathcal{S}(x,y) \in [jN^2/k, N^2] & \text{if } j = k-1 \\ \mathcal{S}(x,y) \in [jN^2/k, (j+1)N^2/k[ & \text{else} \end{cases} \tag{3}$$

Finally, a block with coordinates $(x, y)$ can be drawn with a uniform probability distribution from all candidates in $\mathcal{V}$. If no valid candidates were found and $\#\mathcal{V} = 0$, then the block search is performed again with the same subset $j$, but on a new image $I$ and coding configuration $C_i$ till $\#\mathcal{V} \neq 0$. As this block sampling algorithm allows to find exhaustively all valid block coordinates within a image, any infinite loop problem when $\#\mathcal{V} = 0$ is avoided. As shown in Figure 4.b, accuracy of drawn blocks are now perfectly balanced across the $k$ subsets. Note that sampling blocks randomly within an image using a uniform probability distribution function would have lead to a very strong imbalance of accuracies values. Figure 4.a highlight that more than $80\%$ of blocks would have an accuracy greater than $0.9$ with such uniform block sampling strategy, which is not ideal to compute correlation.

Figure 2 illustrates the difference between real GT and pseudo GT $P$. As it can be seen, some pixels do not belong to any class in the real GT, i.e. pixels of the car being driven. Since DNN was trained only with labeled pixels, pseudo GT tends to be noisy in these areas with only slight artifacts. Consequently, unlabelled areas seem to be privileged by the proposed block sampling algorithm, since these noisy areas tend to have a wide range of accuracy scores from low to high, as opposite to other part of the image where there are mostly blocks with high accuracy. This phenomenon can be observed in Figure 3.b. To mitigate this bias of selecting mostly blocks of low relevance, a block coordinate in $\mathcal{V}$ is considered valid if the corresponding block in the real GT does not contain any pixel with unlabelled data. Figure 3.c illustrates the new probability density function, where most blocks are drawn in meaningful areas, i.e. in areas that are the most critical when driving.

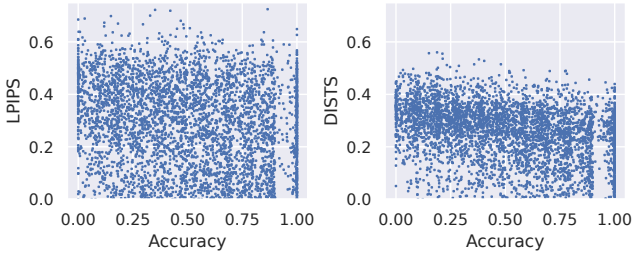
For this experiment, a total of $\#B = 2^{12}$ blocks are sampled for each block size using the proposed block sampling algorithm with $k = 10$. Considered block sizes are $32 \times 32$, $64 \times 64$ and $128 \times 128$.

TABLE II: Image-level and block-level correlations scores.

| Metric | Image-level | | | Block-level, $32 \times 32$ | | | Block-level, $64 \times 64$ | | | Block-level, $128 \times 128$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | PLCC | SROCC | KROCC | PLCC | SROCC | KROCC | PLCC | SROCC | KROCC | PLCC | SROCC | KROCC |
| PSNR | 0.5807 | **0.6992** | **0.5065** | 0.0036 | 0.0008 | 0.0003 | 0.0474 | 0.0500 | 0.0330 | 0.1589 | 0.1567 | 0.1042 |
| SSIM | 0.6060 | 0.6723 | 0.4823 | 0.0032 | 0.0031 | 0.0023 | 0.0245 | 0.0222 | 0.0149 | 0.1058 | 0.1099 | 0.0736 |
| MS-SSIM | 0.6527 | 0.6975 | 0.5038 | NaN | NaN | NaN | 0.0425 | 0.0693 | 0.0456 | 0.1308 | 0.1696 | 0.1129 |
| FSIM | **0.6748** | 0.6805 | 0.4887 | 0.0153 | 0.0312 | 0.0213 | 0.1234 | 0.1355 | 0.0903 | 0.2491 | 0.2575 | 0.1732 |
| SR-SIM | 0.6617 | 0.6922 | 0.4999 | 0.0095 | 0.0062 | 0.0041 | 0.0182 | 0.0175 | 0.0122 | 0.0715 | 0.0903 | 0.0612 |
| GMSD | 0.6669 | 0.6973 | 0.5047 | 0.0515 | 0.0438 | 0.0294 | 0.1075 | 0.1193 | 0.0788 | 0.2273 | 0.2340 | 0.1568 |
| MS-GMSD | 0.6718 | 0.6989 | 0.5063 | 0.0497 | 0.0377 | 0.0255 | 0.1032 | 0.1116 | 0.0739 | 0.2258 | 0.2315 | 0.1550 |
| VSI | 0.6148 | 0.6538 | 0.4664 | 0.0372 | 0.0383 | 0.0258 | 0.1202 | 0.1299 | 0.0870 | 0.2420 | 0.2603 | 0.1760 |
| HaarPSI | 0.6679 | 0.6820 | 0.4900 | **0.0658** | 0.0530 | 0.0356 | 0.1301 | 0.1347 | 0.0894 | 0.2711 | 0.2738 | 0.1840 |
| MDSI | 0.6630 | 0.6888 | 0.4968 | 0.0293 | 0.0224 | 0.0154 | 0.0742 | 0.0728 | 0.0484 | 0.1712 | 0.1652 | 0.1108 |
| LPIPS | 0.5821 | 0.6636 | 0.4742 | 0.0561 | **0.0552** | **0.0371** | **0.1751** | **0.1747** | **0.1165** | 0.3067 | 0.3181 | 0.2139 |
| DISTS | 0.6633 | 0.6692 | 0.4781 | 0.0496 | 0.0441 | 0.0296 | 0.1629 | 0.1634 | 0.1092 | **0.3160** | **0.3276** | **0.2216** |

(a) Pseudo GT $P$ (b) Comp. image $\hat{I}$ (c) Prediction $\hat{P}$

Fig. 5: Example indicating that blocks containing simple areas such as *road*, *building* or *vegetation* can stay correctly classified even under extreme degradation. It is possible to find blocks with perfect accuracy where $\hat{I}$ is composed of a single luminance value.




(a) (b)

Fig. 6: Scatter plots between accuracy and (a) LPIPS on $64 \times 64$ blocks, (b) DISTS on $128 \times 128$ blocks.

## IV. Results and Discussion

Correlation of IQA metrics with DNN accuracy for image-level and block-level is given in Table II. Note that absolute value of IQA metrics are used to compute SROCC, PLCC and KROCC in order to obtain positive correlation scores for all metrics. As it can be seen, there is a quite high correlation between IQA metrics and DNN accuracy on the image-level experiment, with SROCC scores lower but close to $0.7$ for most metrics. In other terms, given a IQA metric score, one can predict with a relatively high confidence the proportion of correctly classified pixels. Intuitively, a low level of degradation in compressed images $\hat{I}$ keep predictions $\hat{P}$ close to pseudo GT $P$, while a higher level of degradation will generally imply lower accuracies.

On the block-level, correlation is much lower. As block size gets lower, IQA metrics and accuracy become less and less correlated. While LPIPS and DISTS manage to achieve a correlation above $0.3$ using SROCC or PLCC on $128 \times 128$ blocks, no metric is able to reach a correlation of $0.1$ on $32 \times 32$ blocks. Indeed, the block-level experiment highlight the lack of ability of existing IQA metrics to predict DNN accuracy on a local standpoint. Figure 5 shows that it is possible to find blocks where IQA metrics indicate a very poor quality while the accuracy is perfect. In contrast, slightly distorted image $\hat{I}$ could alter predictions $\hat{P}$ from pseudo GT $P$ since DNN can lack of resilience to adversarial attacks. Therefore, the higher correlation on the image-level comes from the greater content diversity within each image, since conventional metrics are unable to give accurate predictions on a local scale.

The lack of correlation on the block-level has profound implication on encoder design. Nowadays, most encoders are based on AVC, HEVC or VVC standards, which imply the use of a RDO algorithm to determine appropriate encoding decisions. RDO aims at minimizing blocks degradation according to a FR metric under a given rate constraint. Note that the RDO algorithm may have to tackle even smaller blocks than considered ones in this study. As an example, some VVC based encoders may use blocks down to $4 \times 4$, where correlation would be even harder to obtain. Since all considered metrics fails to achieve high level of correlation with DNN accuracy, minimizing such metric may not translate in better DNN performance.

Figure 6 presents scatter plots with FR metrics against accuracy. Considered metrics are the ones with the best correlation, namely LPIPS and DISTS. Even though these metrics are the ones with the best correlation, DNN accuracy cannot be inferred based on the metric score since a clear relation between the two variable cannot be found. Note that LPIPS and DISTS are based on deep models such as VGG [25], which are unsuited for RDO since the distortion measure may be called millions of times in modern codecs for each frame in a video when a high quality is desired.

Based on the observation that existing metrics are not correlated with machine perception, it is desirable to propose a suitable metric in the VCM context. To this end, the built dataset to perform this study is made freely accessible to facilitate future works on this topic.

## V. Conclusion

In this paper, we evaluated correlation of FR IQA metrics with DNN prediction accuracy for the semantic segmentation vision task on various coding configurations including multiple image resolutions and multiple encoders. A novel evaluation

protocol in the context of VCM is used to perform this experiment, using pseudo GT and DNN trained on distorted data. Experiments indicate a low correlation of conventional IQA metrics with machine perception, especially on a block-level. Therefore, existing FR metrics used in RDO process or end-to-end encoders appear inappropriate in the VCM context. Novel FR metrics that achieve higher correlation with machine perception would enable better encoding choice and therefore greater trade-offs between rate and DNN accuracy. In order to encourage the emergence of such metrics, the dataset used in this work is made publicly available.

## REFERENCES


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

[2] 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.

[3] U Cisco. Cisco annual internet report (2018–2023) white paper. 2020. *Acessado em*, 10(01), 2021.

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

[5] 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.

[6] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[7] 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.

[8] 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.

[9] Kristian Fischer, Fabian Brand, and André Kaup. Boosting Neural Image Compression for Machines Using Latent Space Masking. *IEEE Transactions on Circuits and Systems for Video Technology*, pages 1–1, 2022.

[10] 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.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] Zohaib Amjad Khan, Aladine Chetouani, Giuseppe Valenzise, and Frédéric Dufaux. Towards an Image Utility Assessment Framework for Machine Perception. In *European Signal Processing Conference (EUSIPCO 2022)*, 2022.

[13] Eric Cooper Larson and Damon Michael Chandler. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of Electronic Imaging*, 19(1):1 – 21, 2010. Publisher: SPIE.

[14] 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.

[15] Mikołaj Leszczuk, Lucjan Janowski, Jakub Nawała, and Atanas Boev. Method for Assessing Objective Video Quality for Automatic License Plate Recognition Tasks. In Andrzej Dziech, Wim Mees, and Marcin Niemiec, editors, *Multimedia Communications, Services and Security*, pages 153–166, Cham, 2022. Springer International Publishing.

[16] Mikołaj Leszczuk, Lucjan Janowski, Jakub Nawała, and Atanas Boev. Objective Video Quality Assessment Method for Face Recognition Tasks. *Electronics*, 11(8), 2022.

[17] 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.

[18] Alban Marie, Karol Desnos, Luce Morin, and Lu Zhang. Video Coding for Machines: Large-Scale Evaluation of Deep Neural Networks Robustness to Compression Artifacts for Semantic Segmentation. In *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, pages 01–06, 2022.

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

[20] Nikolay Ponomarenko, Lina Jin, Oleg Ieremeiev, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi, Marco Carli, Federica Battisti, and C.-C. Jay Kuo. Image database TID2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, 30:57–77, 2015.

[21] Rafael Reisenhofer, Sebastian Bosse, Gitta Kutyniok, and Thomas Wiegand. A Haar wavelet-based perceptual similarity index for image quality assessment. *Signal Processing: Image Communication*, 61:33–43, 2018.

[22] David Rouse, Romuald Pépion, Sheila Hemami, and Patrick Le Callet. Image Utility Assessment and a Relationship with Image Quality Assessment. In *Human Vision and Electronic Imaging XIV 2009*, volume 7240, pages pp. 724010–724010–14 (2009), San José, California, United States, January 2009.

[23] Francois Rozet. PyTorch Image Quality Assessment, 2021.

[24] H.R. Sheikh, M.F. Sabir, and A.C. Bovik. A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms. *IEEE Transactions on Image Processing*, 15(11):3440–3451, 2006.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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

[27] G.J. Sullivan and T. Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Processing Magazine*, 15(6):74–90, 1998.

[28] Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pages 1398–1402 Vol.2, 2003.

[29] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[30] Wufeng Xue, Lei Zhang, Xuanqin Mou, and Alan C. Bovik. Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index. *IEEE Transactions on Image Processing*, 23(2):684–695, 2014.

[31] Bo Zhang, Pedro V. Sander, and Amine Bermak. Gradient magnitude similarity deviation on multiple scales for color image quality assessment. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1253–1257, 2017.

[32] Lin Zhang and Hongyu Li. SR-SIM: A fast and high performance IQA index based on spectral residual. In *2012 19th IEEE International Conference on Image Processing*, pages 1473–1476, 2012.

[33] Lin Zhang, Ying Shen, and Hongyu Li. VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281, 2014.

[34] Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. FSIM: A Feature Similarity Index for Image Quality Assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

[35] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CoRR*, abs/1801.03924, 2018. _eprint: 1801.03924.

[36] 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.

[37] Hossein Ziaei Nafchi, Atena Shahkolaei, Rachid Hedjam, and Mohamed Cheriet. Mean Deviation Similarity Index: Efficient and Reliable Full-Reference Image Quality Evaluator. *IEEE Access*, 4:5579–5590, 2016.