Towards Machine Perception Aware Image Quality Assessment

Alban Marie, Karol Desnos, Chen Fu, Jinjia Zhou, Luce Morin, Lu Zhang

Towards Machine Perception Aware Image Quality Assessment

Alban Marie*, Karol Desnos*, Chen Fu†, Jinjia Zhou†, Luce Morin* and Lu Zhang*
*Univ Rennes, INSA Rennes, CNRS, IETR – UMR 6164, F-35000 Rennes, France
Email: {alban.marie, karol.desnos, luce.morin, lu.ge}@insa-rennes.fr
†Graduate School of Science and Engineering, Hosei University, Tokyo, Japan
Email: chen.fu.6r@stu.hosei.ac.jp, zhou@hosei.ac.jp

***Abstract*—Over the years, the objective of image and video compression has been to preserve perceived quality according to the Human Visual System (HVS) with minimal rate. Traditional encoders achieve this with the use of Rate-Distortion Optimization (RDO) techniques along with Image Quality Assessment (IQA) metrics that are correlated with human perception. Nowadays, a fast-growing number of applications fall within the realm of Video Coding for Machines (VCM), where the final recipient of compressed data is not a human but a machine performing a vision task. Recently, the lack of correlation between existing distortion measures and machine perception has been revealed, especially for RDO algorithms where distortion measures are computed on a local scale. In this paper, we propose a machine perception-aware metric designed to be incorporated into a standard-compliant Versatile Video Coding (VVC) encoder. Our proposed metric relies on a supervised training procedure as well as additional information available on the encoder side. In terms of correlation with machine perception, our metric significantly outperforms existing distortion measures in the literature.**

***Index Terms*—Video Coding for Machines (VCM), Image Quality Assessment (IQA), Full-Reference (FR), Rate-Distortion Optimization (RDO), Versatile Video Coding (VVC)**

I. Introduction

Throughout the decades, traditional standards such as Versatile Video Coding (VVC) [1] have achieved significant advancements in coding efficiency. Coding efficiency refers to the ability of an encoder to jointly minimize both the rate and distortion, where the distortion is being measured by an Image Quality Assessment (IQA) metric. This minimization is accomplished through the use of Rate-Distortion Optimization (RDO) [17]. By ensuring a strong correlation between the selected IQA metric and the perception of the Human Visual System (HVS), RDO techniques allow a compact representation to be found while preserving quality as perceived by humans. However, recent progress in computer vision tasks [2], [7], [20] has resulted in an increasing amount of visual content compressed for machine analysis rather than human consumption [3]. To this end, the Moving Picture Expert Group (MPEG) has established the Video Coding for Machines (VCM) group [24], with the goal of surpassing the trade-offs in bitrate and machine performance encountered by the VVC Test Model (VTM) encoder.

Recently, a study evaluating the correlation of common IQA metrics with machine perception has been conducted [12]. Experiments indicated that when measured on a local scale, the correlation levels between IQA scores and machine perception were low. Therefore, with the use of such IQA metrics, minimizing the distortion measure within the RDO loop does not guarantee the performance of the final vision task to be preserved.

This paper introduces a novel machine perception aware IQA metric that is designed to be integrated within the RDO loop of a VVC-based encoder. Our proposed metric utilizes a Full-Reference (FR) strategy, where a distorted block is compared against its undistorted counterpart. By leveraging additional information available on the encoder side, our metric enhances its ability to provide relevant scores. Moreover, a supervised learning strategy minimizes the distance between labels representing machine perception and metric predictions.

The remainder of this paper is organized as follows. Section II and Section III provide a review of existing works in the literature and background about the VVC standard, respectively. A detailed presentation of the proposed metric is presented in Section IV. Finally, experiments are conducted in Section V, followed by a conclusion.

II. Related Works

The field of Image Quality Assessment (IQA) focuses on finding quality models for images that match the HVS perception. One type of such quality model is referred to as Full-Reference (FR) metrics, where the degradation within an image is compared to a pristine reference image. As it was observed that the legacy Peak Signal to Noise Ratio (PSNR) metric lacks correlation with human perception, many IQA metrics have been proposed over the years. Notable examples include the Structural SIMilarity (SSIM) [21], as well as more recent deep learning-based metrics such as Learned Perceptual Image Patch Similarity (LPIPS) [23], Deep Image Structure and Texture Similarity (DISTS) [5] and NIMA [18]. Despite reaching a high correlation with the HVS, these metrics are unsuited to correlate with machine perception, especially on a block-level [12]. Consequently, minimizing the score returned by a FR IQA metric within a RDO loop is sub-optimal to preserve the machine task performance in a VCM context.

Being able to predict machine perception with a metric is related to the concept of image *utility* [14]. The utility

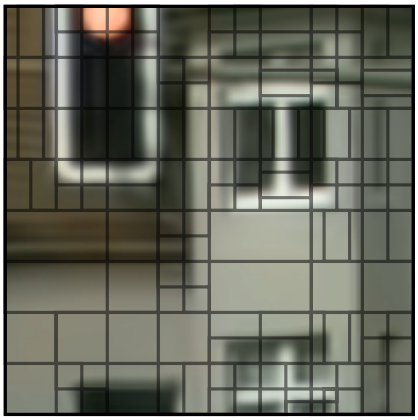

Fig. 1. Illustration of Block Partitioning for a 128×128 CTU. CUs can be subdivided into fine or coarse blocks depending on the content.

of an image refers to the ability of machines to provide accurate predictions on them. Khan et al. [10] showed that a simple FR metric based on a Resnet-18 network [7] can predict if the machine task predictions on both degraded and pristine images would be equivalent. This observation is consistent with the fact that existing FR IQA metrics are able to correlate with machine perception on an image-level [12]. Fischer et al. [6] proposed a Feature-based RDO (FRDO), where the first 5 layers of a pre-trained VGG-16 [16] are used to extract features on which distortion measures such as the Sum of Squared Errors (SSE) or the Sum of Absolute Differences (SAD) are computed. These distortion measures are referred to as Feature-based SSE (FSSE) and Feature-based SAD (FSAD), respectively. However, performing the RDO in the feature domain over the pixel domain might not be preferable, as there is no statistical evidence that measuring distortion on such domain offers a greater correlation with machine perception.

III. VVC Background

One crucial component of a VVC encoder is the Rate-Distortion Optimization (RDO) [17]. In essence, the RDO algorithm tackles an optimization problem to identify the most suitable encoding solution within a search space. The search space can encompass all possible ways to divide a CTU into smaller Coding Unit (CU) using the available splits defined by the VVC standard. Figure 1 illustrates a block partitioning example obtained with a RDO for a given CTU. As it can be seen, larger CU are predominantly selected by the RDO for areas consisting primarily of low frequencies, while CU as small as 4×4 can be utilized to encode complex regions.

The RDO consists in jointly minimizing 2 antagonist terms that are the rate and the distortion. For each tested encoding possibility among the search space, the rate refers to the number of bits required by the entropy coder to encode a block, while the distortion is a measure that quantifies the amount of degradation introduced in the reconstructed block. A bit-rate constraint R_{max} set to a low value may imply a high level of degradation in the encoded block, and vice-versa.

Let ω represent an encoding possibility among the search space Ω, i.e. a set of possible encoding possibilities. Let R and D be functions that return the rate and the distortion of a given encoding possibility ω. The RDO consist in minimizing the distortion D with respect to a bit-rate constraint R_{max}:

$$\omega^* = \underset{\omega \in \Omega}{\operatorname{argmin}}\, D(\omega) \qquad \text{w.r.t. } R(\omega) < R_{max} \tag{1}$$

where ω^* is the optimal solution found by the optimization algorithm. Sullivan et al. [17] solve Equation 1 using a unconstrained Lagrangian optimization that weight the distortion D over the rate R using a Lagrangian multiplier λ:

$$\omega^* = \underset{\omega \in \Omega}{\operatorname{argmin}}\, J(\omega) \qquad \text{s.t. } J(\omega) = D(\omega) + \lambda R(\omega) \tag{2}$$

Minimizing J in Equation (2) for a given Lagrangian multiplier λ leads to an optimal solution of Equation (1) for a particular bit-rate constraint R_{max}. Tuning λ results in different rate-distortion trade-offs. Higher λ values favor the minimization of the rate term R, thus leading to highly compressed visual content. Conversely, lower λ values will encourage the RDO to find a solution with a low distortion D and hence a high-quality output.

The distortion function D measures the quality of the reconstructed block $\hat{B}$ compared to the original block B. Practical use of the RDO most often relies on the SSE or the SAD as the distortion measure D:

$$SSE(B, \hat{B}) = \sum_{i=1}^{W} \sum_{j=1}^{H} |B(x, y) - \hat{B}(i, j)|^2 \tag{3}$$

$$SAD(B, \hat{B}) = \sum_{i=1}^{W} \sum_{j=1}^{H} |B(x, y) - \hat{B}(i, j)| \tag{4}$$

where (x, y), W and H are pixel coordinates, block width and block height, respectively.

IV. Proposed Metric

Figure 2 highlights the architecture of our proposed machine perception aware metric. Similar to SSE or SAD distortion measures, our metric uses a FR strategy where degraded content is compared against a reference.

A. Model Architecture

Within the RDO loop of a VVC based encoder, a distortion measure is used to compare a reconstructed CU over the corresponding CU reference. Because of VVC block partitioning, CU size can vary from 64×64 down to 4×4. It is important

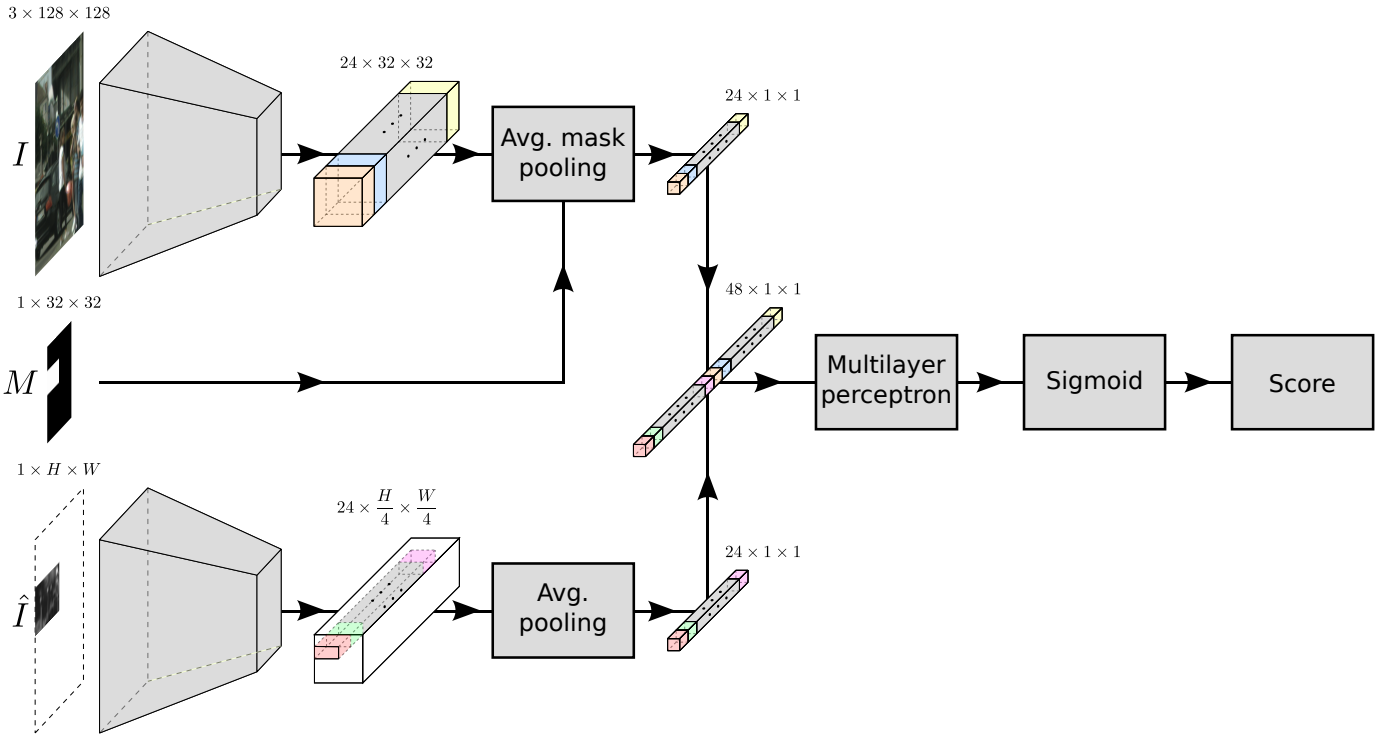

Fig. 2. Proposed machine perception aware metric. The metric uses a Full-Reference (FR) strategy and predict a score based on the pristine Coding Tree Unit (CTU) I_{CTU}, the distorted Coding Unit (CU) $\hat{I}_{CU}$ and a mask M indicating the CU position within the CTU. A score of 0 indicates that the segmentation algorithm predicts identical outcomes for blocks I_{CU} and $\hat{I}_{CU}$, while a score of 1 indicates different classifications for all pixels between blocks I_{CU} and $\hat{I}_{CU}$.

to emphasize that predicting a meaningful score becomes increasingly harder as the CU size decreases. This is due to the reduced amount of information available to the metric at smaller block sizes. Indeed, empirical evidence have shown that correlation between FR IQA metrics scores and machine perception was higher when measured on 128×128 blocks compared to 64×64 or 32×32 blocks [12]. To this end, our proposed metric takes advantage of more information by using as input the whole reference CTU I_{CTU} with both luminance and chrominance components. The added context allows our metric to significantly improve prediction quality, especially on smaller CU, as shown in Section V. Furthermore, the use of the entire reference CTU I_{CTU} instead of the reference CU I_{CU} allows the computational complexity of our metric to be lowered when integrated into the RDO loop of an encoder. This is because the input reference CTU I_{CTU} remains the same for all CUs within the CTU, allowing subsequent computations to be performed only once per reference CTU I_{CTU}. For the distorted block $\hat{I}_{CU}$, note that only the CU and a single component are used, since the reconstructed CTU and other components are not accessible within the RDO encoding process. In this study, the luminance component is considered for the distorted CU $\hat{I}_{CU}$.

The proposed metric start with two feature extraction steps, applied separately on the reference CTU I_{CTU} and the distorted CU $\hat{I}_{CU}$. Both feature extraction models have separate weights but share the same architecture. The architecture is based on the first layers of EfficientNet-b0 [20], which rely on Mobile inverted Bottleneck Convolutions (MBConv) [15], [19] and squeeze-and-excitation modules [9]. More detail on the used architecture is provided in Table I. The weights of both feature extractors are initialized with the *Pytorch* implementation of EfficientNet-b0. Note that the height H and width W of the obtained feature maps are divided by a factor of 4 due to the use of strided convolutions. This design choice is motivated by the requirement that the distortion measure can be computed on blocks as small as 4×4 in the case of VVC.

Thereafter, the feature maps of size $24 \times 32 \times 32$ and $24 \times \frac{H}{4} \times \frac{W}{4}$ undergo downsampling through an average pooling operator, resulting in feature maps of size $24 \times 1 \times 1$. Let $\mathcal{X}$ represent the feature map obtained from the feature extractor which takes the reference CTU I_{CTU} as input. Let M be a mask of size $1 \times 32 \times 32$ that provides information about the CU position within the CTU. A value of 1 for a pixel in the mask M indicates that the corresponding 4×4 area in the CTU belongs to the CU, while a value of 0 indicates that the region lies outside the CU. The average mask pooling first consists in extracting a subregion $\mathcal{Y}$ of the feature map $\mathcal{X}$ using a mask M. $\mathcal{Y}$ can be obtained by keeping values from the feature map $\mathcal{X}$ that correspond to the non-zero area of the mask M. The resulting feature map $\mathcal{Y}$ has dimensions of $24 \times \frac{H}{4} \times \frac{W}{4}$. Afterwards, a conventional average pooling operator is applied to the feature map $\mathcal{Y}$, resulting in a vector $\mathcal{Z}$ with 24 elements. The c^{th} element in the vector $\mathcal{Z}$ is the average of the $\frac{H}{4} \times \frac{W}{4}$

TABLE I
USED ARCHITECTURE FOR BOTH FEATURE EXTRACTORS. OUTPUT RESOLUTION IS SPECIFIED FOR THE REFERENCE CTU I_{CTU} AS INPUT. NOTE THAT THE FIRST CONVOLUTION TAKES 3 AND 1 INPUT CHANNELS FOR THE REFERENCE CTU I_{CTU} AND THE DISTORTED CU $\hat{I}_{CU}$, RESPECTIVELY.

Operator	Output channels	Output resolution
Conv, 3×3	32	64×64
MBConv, 3×3	16	64×64
MBConv, 3×3	24	32×32
MBConv, 3×3	24	32×32
Conv, 1×1	24	32×32

elements from the c^{th} channel of the feature map $\mathcal{Y}$. Note that the mask M is not used for the distorted head that takes as input $\hat{I}$, since the corresponding area already refers to the CU.

Based on both undistorted and distorted heads of our model, two vectors of 24 elements each are obtained. These vectors are then concatenated and fed as input to an Multi-Layer Perceptron (MLP). The employed MLP architecture consists of an input layer with 48 elements, followed by three hidden layers containing 24, 12, and 6 neurons, respectively. The Sigmoid Linear Unit (SiLU) [8] is used in our MLP as the activation function. The output layer comprises a single neuron, which is followed by a sigmoid function to map the possible set of scores of our FR metric between 0 and 1.

B. Training Strategy

In order to train the proposed model, the methodology outlined in [12] is employed to gather training data, evaluation data and associated labels. This process involves leveraging the Cityscapes dataset [4] and utilizing a semantic segmentation algorithm [2]. Following the same methodology, uncompressed images, compressed images, segmentation predictions of uncompressed images, and segmentation predictions of compressed images are acquired. The same coding configurations are considered to obtain compressed images [12], encompassing Joint Photographic Experts Group (JPEG), Advanced Video Coding (AVC), High Efficiency Video Coding (HEVC), and VVC compression applied across a wide range of resolutions and qualities. Subsequently, the uncompressed images and compressed images are utilized to extract reference CTU I_{CTU} and distorted CU $\hat{I}_{CU}$, respectively. The same block sampling strategy as the original paper is employed to ensure that the errors of sampled blocks are uniformly distributed, thus guaranteeing a well-balanced dataset. To generate the training labels representing the machine perception measure, blocks from the segmentation predictions of uncompressed and compressed images are extracted at the corresponding CU coordinates. These blocks are denoted as P_{CU} $\hat{P}_{CU}$, respectively. For more detailed explanations, the reader can refer to the corresponding article [12].

During the training process, the pixel-wise accuracy between the undistorted CU P_{CU} and the distorted CU $\hat{P}_{CU}$ is used as the training label. Notably, the mean Intersection over Union (mIoU) is not employed, despite being the standard metric to assess the relevance of predictions for semantic segmentation. Indeed, the mIoU comes with several limitations when it is computed on a local scale. Due to the high likelihood that a block contains only one or very few classes, the mIoU scores can be significantly impacted even if most pixels within the block are correctly classified. In extreme cases, the misclassification of a single pixel can lead to mIoU scores below 0.5, regardless of the block size [12]. To address these limitations, the pixel-wise accuracy between the undistorted CU P_{CU} and the distorted CU $\hat{P}_{CU}$ is used as a measure of machine perception. The training label used to train our model is determined by subtracting such pixel-wise accuracy from 1. Consequently, a prediction of 0 from our metric indicates that the predictions on the undistorted CU I_{CU} and the distorted CU $\hat{I}_{CU}$ are the same. This behavior is analogous to the SSE and SAD distortion measures, where a score of 0 denotes no difference between the original block B and the reconstructed block $\hat{B}$, as shown in Equation (3) and Equation (4).

TABLE II
CONSIDERED CU SIZES AT TRAINING.

4×4	8×4	16×4	32×4	64×16
4×8	8×8	16×8	32×8	64×64
4×16	8×16	16×16	32×16	
4×32	8×32	16×32	32×32	
		16×64		

To ensure the proper training and evaluation of our model, distinct sets are necessary. As mentioned earlier, the Cityscapes dataset [4] is employed to gather the required training data, evaluation data, and associated labels. Note that images within the Cityscapes training set cannot be used directly for training or evaluating the model. Indeed, these images would not yield representative predictions of the semantic segmentation algorithm performance, as these images have already been seen during the training phase. To this end, the Cityscapes testing set is utilized for training the model, while the validation set is used for evaluation.

As mentioned in Section IV-A, the size of a CU can range from 64×64 to 4×4. To ensure that model predictions accurately reflect machine perception for any given CU size, a CU size is randomly sampled from a set using a uniform distribution function at each batch during training. The considered set of CU sizes, which were determined empirically by examining the CU sizes encountered in the RDO loop of VTM, is presented in Table II.

The optimal learning rate and batch size were determined using the Asynchronous Successive Halving Algorithm (ASHA) [11] hyper-parameter optimization technique, using the Ray Tune library implementation. ASHA objective is to maximize the Pearson Linear Correlation Coefficient (PLCC) between model predictions and labels of the evaluation set by tweaking hyper-parameters. Additionally, L^1 norm, L^2 norm, cosine similarity and Pearson correlation loss functions were compared through the ASHA. Ultimately, a batch size of 64, a learning rate of 2×10^{-2} and the Pearson correlation loss

TABLE III

CORRELATION BETWEEN FR METRIC SCORES AND MACHINE PERCEPTION ACROSS DIFFERENT BLOCK SIZES. THE TABLE PRESENTS THE CORRELATION FOR BOTH SQUARE BLOCKS (TOP) AND RECTANGULAR BLOCKS (BOTTOM).

Metric	4×4			8×8			32×32			64×64		
	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC
SSE	0.009	0.011	0.008	–0.020	–0.030	–0.021	–0.025	–0.021	–0.014	–0.021	–0.047	–0.031
SAD	0.027	0.013	0.010	–0.029	–0.032	–0.022	–0.018	–0.015	–0.010	–0.032	–0.042	–0.028
SSIM [21]	–0.060	–0.024	–0.017	0.003	0.029	0.019	0.036	0.043	0.027	0.029	0.071	0.047
DISTS [5]	0.051	0.047	0.032	–0.001	0.007	0.005	0.062	0.057	0.038	0.025	0.024	0.016
FSSE [6]	0.023	0.012	0.008	–0.021	–0.043	–0.029	–0.050	–0.063	–0.041	–0.068	–0.079	–0.053
FSAD [6]	0.030	0.011	0.008	–0.031	–0.046	–0.031	–0.056	–0.065	–0.042	–0.080	–0.085	–0.057
Ours	**0.288**	**0.257**	**0.182**	**0.370**	**0.335**	**0.235**	**0.422**	**0.406**	**0.280**	**0.358**	**0.345**	**0.239**

Metric	4×32			16×8			16×32			64×16		
	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC	PLCC	SROCC	KROCC
SSE	–0.053	–0.059	–0.039	0.009	–0.024	–0.015	–0.020	–0.012	–0.007	–0.057	–0.124	–0.081
SAD	–0.057	–0.058	–0.038	0.002	–0.022	–0.014	–0.008	–0.005	–0.003	–0.076	–0.117	–0.076
SSIM [21]	0.060	0.063	0.041	0.006	0.023	0.014	0.019	0.017	0.010	0.085	0.143	0.093
DISTS [5]	–0.046	–0.044	–0.029	0.022	0.012	0.008	0.038	0.037	0.025	–0.063	–0.072	–0.047
FSSE [6]	–0.065	–0.078	–0.052	–0.020	–0.040	–0.025	–0.050	–0.052	–0.034	–0.087	–0.160	–0.106
FSAD [6]	–0.074	–0.079	–0.052	–0.035	–0.042	–0.027	–0.057	–0.055	–0.036	–0.121	–0.162	–0.107
Ours	**0.318**	**0.284**	**0.193**	**0.417**	**0.384**	**0.269**	**0.335**	**0.294**	**0.202**	**0.445**	**0.434**	**0.305**

were identified as the optimal configuration. Let N denote the batch size. In the training process, $\mathbf{x}$ and $\mathbf{y}$ represent N-elements vectors corresponding to the model output predictions and training labels, respectively. The Pearson correlation loss function ℓ_{PLCC} is given with the following equation:

$$\ell_{PLCC}(\mathbf{x}, \mathbf{y}) = -\frac{\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})}{\sqrt{\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})^2}\sqrt{\sum_{i=1}^{N}(\mathbf{y}_i - \overline{\mathbf{y}})^2}} \quad (5)$$

where $\overline{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ represents the average of the underlying variable $\mathbf{x}$. By minimizing ℓ_{PLCC}, the Pearson correlation of each batch is maximized. Our model is trained for 200 epochs, each composed of 2^{15} training samples.

V. EXPERIMENTS

To assess the effectiveness of the proposed machine perception aware metric, the correlation between predicted scores and machine perception is measured. The evaluation data and machine perception measures are collected using the methodology described in Section IV. The correlation is measured with PLCC, Spearman Rank-Order Correlation Coefficient (SROCC), and Kendall Rank-Order Correlation Coefficient (KROCC).

Representative FR metrics from the state-of-the-art are compared against our metric. Firstly, the SSE and the SAD are considered, as they are commonly incorporated into RDO loops. The SSIM metric [21] is also included, given the numerousSSIM-based RDO techniques proposed in the literature [13], [22]. Furthermore, DISTS [5] IQA metric is evaluated, as previous research has shown that this metric exhibit the highest correlation with machine perception on a block-level [12]. Despite reaching a correlation comparable

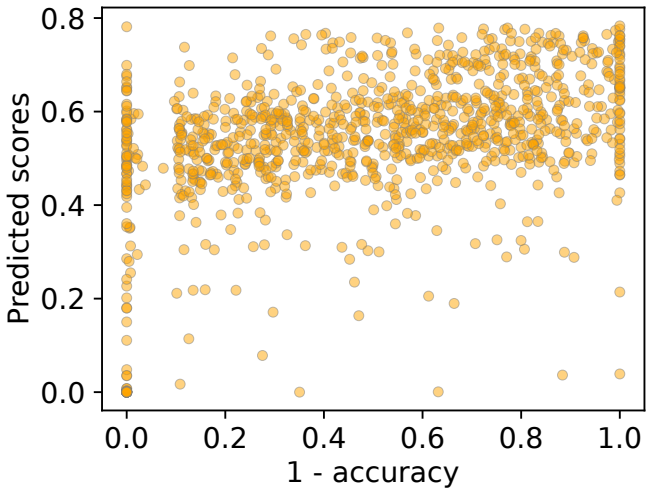

Fig. 3. Scatter plot between machine perception measure and predicted scores by the proposed metric on 64×16 CUs. The machine perception relies on a pixel-wise accuracy measure, as described in Section IV-B.

to that of DISTS, LPIPS [23] is omitted due to its inability to handle blocks smaller than 16×16. However, due to its deep learning-based nature, incorporating DISTS into an RDO loop is hampered because of computational complexity. Finally, both distortion measures FSSE and FSAD proposed by Fischer et al. [6] are also considered.

As indicated in Table II, the RDO loop requires the computation of distortion measures on various CU sizes. To provide a concise analysis, the correlation measurements are presented for eight specific CU sizes: 4×4, 8×8, 32×32, 64×64, 4×32, 16×8, 16×32, and 64×16. Due to the availability of only one image component within the RDO for the distorted CU $\hat{I}_{CU}$, existing metrics from the literature are also constrained to use a single component for both the reference CU I_{CU}

and the distorted CU $\hat{I}_{CU}$. The luminance component for both CU is considered in this experiment. Our metric stands as an exception, as it incorporates both the luminance and chrominance components for the reference CTU I_{CTU} and the luminance component only for the distorted CU $\hat{I}_{CU}$, as described in Section IV.

Table III illustrates the correlation analysis between the considered FR metrics and machine perception. Notably, the metrics from the literature exhibit correlation scores close to 0. As a consequence, minimizing the distortion measure D as shown in Equation (1) and Equation (2) cannot guarantee the performance preservation of the used semantic segmentation algorithm. Hence, it is suggested that FR metrics from the literature are ill-suited in a VCM context.

In contrast, the proposed metric consistently outperforms the existing metrics from the literature by a substantial margin, achieving a PLCC as high as 0.445 on 64×16 CUs. By leveraging the entire CTU I_{CTU} and the associated chrominance components, our metric achieves a higher level of correlation, particularly on smaller blocks [12]. Figure 3 presents a scatter plot illustrating the relationship between machine perception and the predicted scores of our metric. As discussed in Section IV-B, the machine perception measure is obtained by subtracting the accuracy from 1. The figure illustrates that establishing a clear and accurate bijective relationship between the two variables is not straightforward. This underscores the challenges involved in creating an IQA metric that can achieve high levels of correlation with machine perception.

VI. Conclusion

In this paper, a machine perception aware FR IQA metric designed specifically to be integrated within the RDO loop of a VVC-based encoder has been introduced. Our metric takes advantage of additional information available on the encoder side and uses a supervised learning strategy to provide precise scores. A comparative analysis was conducted between our proposed metric and existing metrics from the literature, targeting both human and machine perception. Experimental results revealed that our proposed metric achieved correlation levels as high as 0.445 while existing metrics did not show any significant correlation in the considered experiments. The relatively low correlation levels achieved by our metric underscore the challenges associated with developing a metric that achieves a high correlation with machine perception.

References

[1] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. Overview of the Versatile Video Coding (VVC) Standard and its Applications. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(10):3736–3764, 2021.

[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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Dan Hendrycks and Kevin Gimpel. Gaussian Error Linear Units (GELUs). 2016.

[9] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[10] 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.

[11] Liam Li, Kevin Jamieson, Afshin Rostamizadeh, Ekaterina Gonina, Moritz Hardt, Benjamin Recht, and Ameet Talwalkar. A System for Massively Parallel Hyperparameter Tuning. 2018.

[12] Alban Marie, Karol Desnos, Luce Morin, and Lu Zhang. Evaluation of Image Quality Assessment Metrics for Semantic Segmentation in a Machine-to-Machine Communication Scenario. In *15th International Conference on Quality of Multimedia Experience (QoMEX)*, Ghent, Belgium, June 2023.

[13] Tao-Sheng Ou, Yi-Hsin Huang, and Homer H. Chen. SSIM-Based Perceptual Rate Control for Video Coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(5):682–691, 2011.

[14] 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.

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

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

[18] Hossein Talebi and Peyman Milanfar. NIMA: Neural Image Assessment. *IEEE Transactions on Image Processing*, 27(8):3998–4011, 2018.

[19] M. Tan, B. Chen, R. Pang, V. Vasudevan, M. Sandler, A. Howard, and Q. V. Le. MnasNet: Platform-Aware Neural Architecture Search for Mobile. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2815–2823, 2019.

[20] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International Conference on Machine Learning*, pages 6105–6114. PMLR, 2019.

[21] 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.

[22] Chuohao Yeo, Hui Li Tan, and Yih Han Tan. On Rate Distortion Optimization Using SSIM. *IEEE Transactions on Circuits and Systems for Video Technology*, 23(7):1170–1181, 2013.

[23] 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.

[24] 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.