Received 15 November 2020; revised 8 March 2021; accepted 24 March 2021. Date of current version 13 May 2021.

Digital Object Identifier 10.1109/OJCAS.2021.3073891

Multi-Feature 360 Video Quality Estimation

ROBERTO G. DE A. AZEVEDO [1], NEIL BIRKBECK[2], IVAN JANATRA [2], BALU ADSUMILLI[2], AND PASCAL FROSSARD [1] (Fellow, IEEE)

[1]Signal Processing Laboratory (LTS4), École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland
[2]YouTube Media Algorithms Team, Google Inc., Mountain View, CA 94043 USA

This article was recommended by Associate Editor M. Trocan

CORRESPONDING AUTHOR: R. G. de A. AZEVEDO (e-mail: roberto.azevedo@acm.org)

This work was supported in part by the Youtube, Google Inc.

ABSTRACT We propose a new method for the visual quality assessment of 360-degree (omnidirectional) videos. The proposed method is based on computing multiple spatio-temporal objective quality features on viewports extracted from 360-degree videos. A new model is learnt to properly combine these features into a metric that closely matches subjective quality scores. The main motivations for the proposed approach are that: 1) quality metrics computed on viewports better captures the user experience than metrics computed on the projection domain; 2) the use of viewports easily supports different projection methods being used in current 360-degree video systems; and 3) no individual objective image quality metric always performs the best for all types of visual distortions, while a learned combination of them is able to adapt to different conditions. Experimental results, based on both the largest available 360-degree videos quality dataset and a cross-dataset validation, demonstrate that the proposed metric outperforms state-of-the-art 360-degree and 2D video quality metrics.

INDEX TERMS Visual quality assessment, omnidirectional video, 360-degree video, multi-metric fusion.

I. INTRODUCTION

DRIVEN by the growing interest in virtual and augmented reality, omnidirectional (or 360-degree) videos are becoming prevalent in many immersive applications, e.g., medicine, education, and entertainment. Omnidirectional videos are spherical signals captured by cameras with a full 360-degree field-of-view (FoV). When consumed via head-mounted displays (HMDs) omnidirectional videos allow the user to be immersed in the content. During runtime, based on the user's head motion, the portion of the sphere in the user's field of view, named *viewport*, is seamlessly updated following the user's head motion, thus providing an improved sense of presence. Together, the new immersive features and interactive dimension change the end user perceived quality of experience (QoE) in many ways when compared to traditional videos [2]. Similarly to traditional audiovisual multimedia content, methods for assessing the QoE of omnidirectional content plays a central role in shaping processing algorithms and systems, as well as their implementation, optimization, and testing [24]. In particular, the visual quality assessment of omnidirectional videos is one of the most important aspects of users' QoE when consuming such immersive content.

Quality assessment of 360-degree visual content consumed through HMDs brings its own specificities. For instance, to reuse existing image and video processing technologies, the 360-degree visual content is commonly mapped to a 2D plane (the projection domain) and stored as a rectangular image [1], [2]. Examples of commonly used projections include: equirectangular (ERP), truncated square pyramid (TSP), cube map (CMP), and equiangular cube map (EAC) [10]. The coupled interaction between projection and compression of the resulting rectangular images, however, brings new types of visual distortions [2]. Also, the magnification of the content, the supported increased field-of-view, the fact that the user is completely immersed, and the new interactive dimension, all contribute to the overall perceived visual quality and QoE [2]. Such new features call for the development of new methods and good practices related to both subjective and objective quality assessment of 360-degree visual content [1], [2], [13].

Subjective video quality assessment (VQA) methods collect quality judgments from human viewers through psychophysical experiments. Subjective VQA has the advantage of being more reliable, since humans are the ultimate receivers of the multimedia content. Subjective VQA, however, is expensive, time-consuming, and not suitable for real-time processing quality control. Thus, objective VQA algorithms are required to estimate video quality automatically. Based on the amount of the reference content they use, objective VQA metrics can be divided into: full-reference (FR) methods, which require complete access to the reference video; reduced-reference (RR) methods, which do not require the complete reference, but some features that characterize the reference video; and no-reference (NR) methods, which do not require any information about the reference video. With regards to the prediction accuracy, FR methods are in general more accurate and more widely applied. This paper proposes a new FR method for 360-degree VQA; thus the metrics discussed hereafter are FR schemes.

PSNR-based objective image quality assessment (IQA) metrics that take into account the properties of 360-degree images have been recently proposed in the literature, e.g., S-PSNR [46] CPP-PSNR [47], and WS-PSNR [38]. Those methods are easy to implement and can be efficiently integrated into video coders, but their correlation with subjective judgements are far from satisfactory. Moreover, when used for video quality assessment they lack a proper modelling of the temporal characteristics of the human-visual system (HVS). VQA methods must also consider the temporal factors apart from the spatial ones and the contribution and their interaction to the overall video quality. Therefore, more perceptually-oriented metrics are still required for 360-degree VQA.

In contrast to previous work, we propose a viewport-based multi-metric fusion (MMF) approach for 360-VQA. The proposed approach extends [4][1] and is based on: i) extracting spatio-temporal quality features (i.e., computing objective IQA metrics) from viewports; ii) temporally pooling them taking the characteristics of the human-visual system (HVS) into consideration, and; iii) then training a regression model to predict the 360-degree video quality.

On one hand, working with viewports allows us to better account for the final viewed content and naturally supports different projections [3], [6]. On the other hand, the use of multiple objective metrics computed on these viewports allows our method to have a good performance for the complex and diverse nature of visual distortions appearing in 360-degree videos. Indeed, previous work in both traditional 2D [25], [32], [34] and 360-degree [3] VQA have recognized that even with the multitude of available objective IQA metrics, there is no single one that always performs best for all types of distortions. The combination of multiple metrics is thus a promising approach that can take advantages of the power of individual metrics to properly estimate 360 video quality [26], [34].

Experimental results, based on the largest publicly available 360-degree video quality dataset, VQA-ODV [21], and on the VR-VQA48 [43] dataset show the viability of our proposal, which outperforms state-of-the-art methods for 360-degree VQA.

The rest of the paper is organized as follows. Section II presents the related work. Section III describes our proposal and highlights the main contributions of our proposal. Section IV presents the experimental setup used to validate our approach and the experimental results. Section V-A provides further experiments through ablations studies. Finally, Section VI brings our conclusions and future work.

II. RELATED WORK

A. TRADITIONAL IQA/VQA METHODS

Traditional 2D image and video quality assessment have attracted a lot of attention in recent years [24], [29]. Some objective IQA metrics are well-established today, e.g., Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM) [42], Multiscale Structural Similarity (MS-SSIM) [41], Most Apparent Distortion (MAD) [8], Feature Similarity (FSIM) [48], Gradient Magnitude Similarity Deviation (GMSD) [44], and Haar wavelet-based Perceptual Similarity Index (HaarPSI) [33]. Different from IQA metrics, VQA metrics must also take into consideration the temporal dimension and model the temporal aspects of the HVS. A straightforward and convenient approach to VQA is to use an IQA method on a frame-by-frame basis and then compute the global quality score by simple average of Minkowski summation. However, such approaches do not correctly model the temporal dimension.

Examples of metrics specifically developed for VQA include VQM [30], MOVIE [36], Vis3 [40], and SSTS-GMSD [45]. Although such metrics have produced interesting results, currently the best results are achieved by MMF approaches, from which VMAF is one of the most prominent examples. VMAF [32][2] is an MMF approach for traditional visual content. The objective metrics used as features to train an SVM (Support vector machine) regressor are: Visual Information Fidelity (VIF) [37], Detail Loss Metric (DLM) [23], Anti-Noise Signal-to-Noise Ratio (AN-SNR), and Mean Co-located Pixel Difference (MCPD). MCPD is used as a simple metric for the temporal dimension. The SVM-based regressor is trained to provide an output in the range of 0–100 per video frame. By default, VMAF final output is the average of the individual frame VMAF scores, which is clearly not the best approach for modeling the temporal aspects of the HVS [35].

In this paper, similar to VMAF, we also propose an MMF approach, but we focus on omnidirectional visual content

[1] Compared to [4], we extend the individual features used by our model, propose an adapted temporal pooling method, and provide an extensive new set of experiments, including different regression methods and a cross-dataset validation.

[2] https://github.com/Netflix/vmaf

and take into consideration the perceptual particularities of this new media type. In particular, we compute the individual features in the viewports domain, which allows our method to better account for the final viewed content. Our method uses a different set of individual spatial and temporal features and pooling method (detailed in Section III) which support a better correlation to subjective tests. Finally, in our proposal we use a Random Forest Regression model (RFR) [7] whereas VMAF uses SVR. Such choices are supported by the experimental results in Section IV.

B. OBJECTIVE METRICS FOR 360-DEGREE VQA

Currently, the main approaches for objectively assessing the quality of 360-degree content can be broken into 4 categories: 1) well-known objective metrics for 2D content computed on the projection domain; 2) well-known objective metrics for 2D content computed on the viewports; 3) objective metrics specifically developed for 360-degree visual content; and 4) deep learning techniques.

The use of standard 2D image and video metrics (e.g., the ones discussed in Section II-A) directly in the projection domain is straightforward, but they do not properly model the perceived quality of the 360-degree content. The main issues with such an approach are two-fold: first, it gives the same importance to the different parts of the spherical signal, which not only is sampled very differently from classical images, but also have different viewing probabilities (thus different importance); and second, even for traditional images, these metrics are known to have limitations for different visual distortion types—none is universally satisfactory [25]. In this paper, we address such issues by computing objective metrics in the viewport domain and by employing a MMF fusion approach.

To cope with the sampling issue of the projection domain, recent proposals for omnidirectional image quality assessment have been developed to tackle the specific geometry of 360-degree images: Spherical-PSNR (S-PSNR) [46], Craster Parabolic Projection PSNR (CPP-PSNR) [47], Weighted-to-Spherical-PSNR (WS-PSNR) [38], and S-SSIM [9]. In S-PSNR, sampling points uniformly distributed on a spherical surface are re-projected to the original and distorted images respectively to find the corresponding pixels, followed by the PSNR calculation. In CPP-PSNR, the PNSR is computed between samples in the CPP domain [47], where pixel distribution is closer to that in the spherical domain. The pixels of the original and distorted content are first projected to the spherical domain and then mapped to the CPP domain, where PSNR is computed. In WS-PSNR, the PSNR computation at each sample position is performed directly on the planar domain, but its value is weighted by the area on the sphere covered by the given sample. Different weight patterns may be used for different projections. S-SSIM is a similar approach to S-PSNR, but using SSIM instead of PSNR [9].

The use of objective metrics computed on the viewports is an interesting approach, in which N viewports of different viewing directions are generated for both the original and the distorted content, and the 2D metric is computed individually for each of these viewports. Then, the overall 360-degree quality metric can be computed by aggregating the quality of individual viewports. If the objective metric properly models the human perceptibility, in theory, it could be a good approximation of the overall 360-degree quality. The use of viewports [3], [6] and Voronoi patches [11], [12] for computing individual IQA metrics have also been discussed. Here, we acknowledge that the use of viewports (or patches) to compute the 360-VQA is indeed a more perceptually-correct ways of assessing 360 visual quality. Previous methods, however, simply compute the quality of 360 images as the average of the viewports (or patches). In contrast, we use an MMF approach that allows to better account for different visual distortion types, the temporal dimension, and viewing probability of 360-degree videos.

Recent works have also proposed deep learning architectures to estimate 360-degree video quality [20], [21], [22]. One of the main issues with such approaches is that the current 360-VQA datasets are not big enough to satisfactorily train deep learning methods. Thus, they need to perform data augmentation, such as splitting the original image into patches or rotating the original 360-degree images. In both cases, however, it is not clear if the new generated patches or rotated images share the same quality scores as the original content.

Finally, all the metrics proposed for 360-VQA mentioned above do not explicitly model the temporal dimension of 360-degree videos. They usually compute the overall quality simply as the average of the quality of each individual frame. Different from IQA, however, VQA metrics shall ideally take the temporal dimension into consideration and properly integrate the temporal properties of the HVS. As previously mentioned, some traditional objective VQA metrics do consider the temporal dimension but do not take into consideration the characteristics of 360-degree videos.

We address the above mentioned issues by computing per-frame spatio-temporal objective metrics in the viewport domain, temporally pooling them by taking into account the HVS, and employing a multi-metric fusion approach that closely matches subjective scores. As previously mentioned, being an MMF approach, our proposal shares some of the principles of similar methods for 2D videos. However, it: i) takes into account the specific features of 360-degree videos; ii) uses a different set of individual spatial and temporal features; iii) is based on an improved temporal pooling method; and iii) uses a random forest regression model. Considered together, those features allows our method to support a better correlation to subjective scores and more robust results than state-of-the-art method.

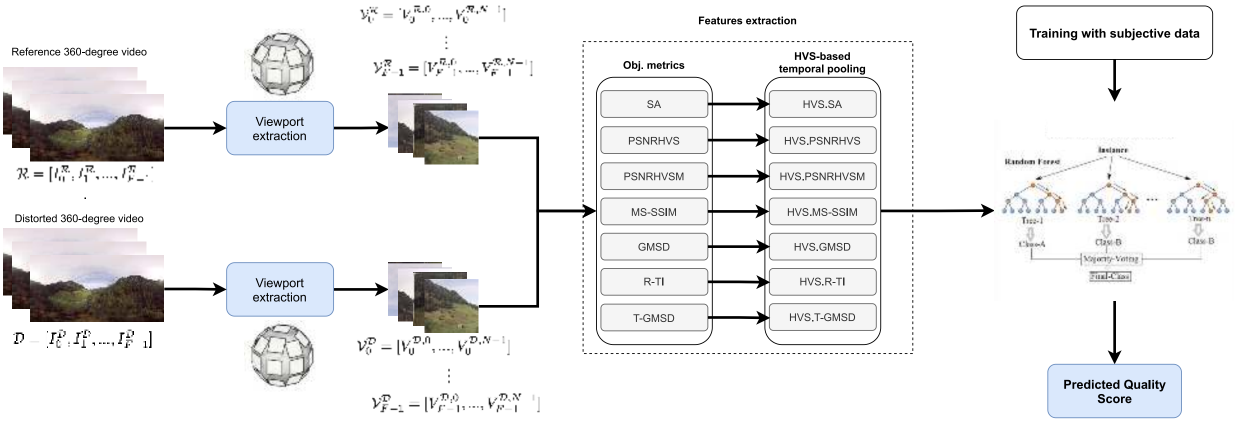

Reference 360-degree video
$\mathcal{R} = [I_0^{\mathcal{R}}, I_1^{\mathcal{R}}, \ldots, I_{F-1}^{\mathcal{R}}]$
Distorted 360-degree video
$\mathcal{D} = [I_0^{\mathcal{D}}, I_1^{\mathcal{D}}, \ldots, I_{F-1}^{\mathcal{D}}]$
Viewport extraction
Viewport extraction
$\mathcal{V}_0^{\mathcal{R}} = [V_0^{\mathcal{R},0}, \ldots, V_0^{\mathcal{R},N-1}]$
$\mathcal{V}_{F-1}^{\mathcal{R}} = [V_{F-1}^{\mathcal{R},0}, \ldots, V_{F-1}^{\mathcal{R},N-1}]$
$\mathcal{V}_0^{\mathcal{D}} = [V_0^{\mathcal{D},0}, \ldots, V_0^{\mathcal{D},N-1}]$
$\mathcal{V}_{F-1}^{\mathcal{D}} = [V_{F-1}^{\mathcal{D},0}, \ldots, V_{F-1}^{\mathcal{D},N-1}]$
Features extraction
Obj. metrics
HVS-based temporal pooling
SA
PSNRHVS
PSNRHVSM
MS-SSIM
GMSD
R-TI
T-GMSD
HVS.SA
HVS.PSNRHVS
HVS.PSNRHVSM
HVS.MS-SSIM
HVS.GMSD
HVS.R-TI
HVS.T-GMSD
Training with subjective data
Random Forest
Instance
Tree-1
Tree-2
Tree-n
Class-A
Class-B
Class-B
Majority-Voting
Final-Class
Predicted Quality Score

FIGURE 1. Proposed 360-VQA multi-method approach overview.

III. VIEWPORTS-BASED MMF FOR 360 VQA

Fig. 1 shows our proposed 360-VQA approach. The possible space of visible viewports is represented by using N viewports rendered from different viewing directions. The same N viewports are rendered from both original and the distorted video content, and 2D objective metrics are computed individually within the viewports and then temporally pooled using an HVS-based method. Finally, based on the per-viewport pooled scores, we train a regression model that is able to learn a combination of the individual objective metrics into a new objective metric that closely relates to subjective scores.

The rest of this section details each of the steps above. In what follows, let $\mathcal{R} = \{R_f, f = 0, 1, \ldots, F-1\}$ and $\mathcal{D} = \{D_f, f = 0, 1, \ldots, F-1\}$ respectively be the reference and distorted sequences of the same 360-degree video content in the projection domain. R_f and D_f denote the f'th frame of $\mathcal{R}$ and $\mathcal{D}$, respectively, and F the total number of frames in both $\mathcal{R}$ and $\mathcal{D}$.

A. VIEWPORT SAMPLING AND FIELD OF VIEW

First, for each frame f, we compute a set of viewports $\mathcal{V}_f^{\mathcal{R}} = \{V_f^{\mathcal{R},0}, \ldots, V_f^{\mathcal{R},N-1}\}$ and $\mathcal{V}_f^{\mathcal{D}} = \{\mathcal{V}_f^{\mathcal{D},0}, \ldots, V_f^{\mathcal{D},N-1}\}$, for the respective reference and distorted frames. A viewport (Fig. 2) is the gnomonic projection [28] of the omnidirectional signal to a plane tangent to the sphere, which is defined by:

- the viewing direction (el_o, az_o), which defines the center O' where the viewport is tangent to the sphere;
- its resolution $[vp_w, vp_h]$; and
- its horizontal and vertical field-of-view, FoV_h and FoV_v, respectively.

When considering a viewport-based metric for 360-degree videos, we need to define a viewport sampling process that, given an omnidirectional image, I, and the viewports parameters vp_w, vp_h, FoV_w and FoV_h, generates N viewports from different viewing directions (i.e., different O's). On one hand, larger FoVs result in both overlapped regions between the viewports and larger geometry distortions. Having duplicated content can be an issue because it increases the importance of such duplicated areas when computing objective metrics on the viewports. On the other hand, smaller FoVs might require more viewports to completely cover the sphere area, which might also result in higher computational costs. Thus, a good balance between the sampling and number of viewports and its distribution to provide good coverage of the sphere is necessary. Ideally, the viewport resolution $[vp_w, vp_h]$ should also match the HMD resolution used to visualize the content.

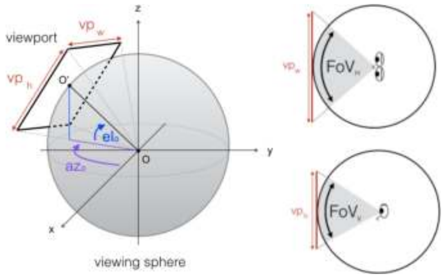


FIGURE 2. Viewport parameters [14].

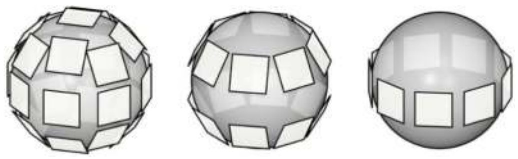
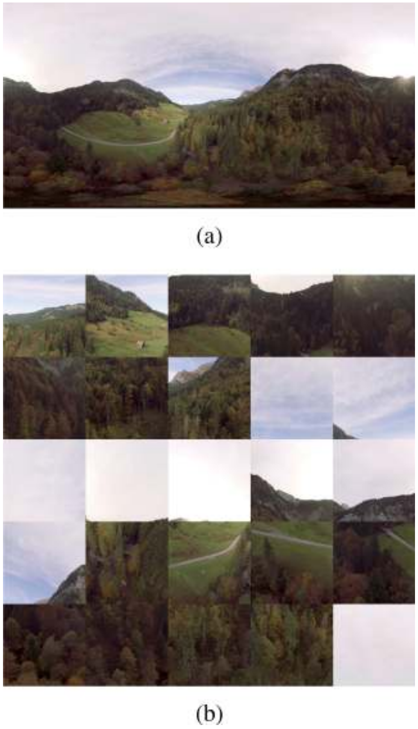


FIGURE 3. Uniform (left), tropical (center), and equatorial (right)viewports sampling for computing viewport-based objective metrics.

Fig. 3 shows three viewport sampling configurations: *uniform*, *tropical*, and *equatorial* [6], that sample, respectively, 25, 16, and 9 viewports. Fig. 4 shows an example of an ERP frame and the viewports generated using the *uniform* sampling method. In the figure, the viewports are aggregated into a single frame that we will refer to as the *viewports collage* (VP-Collage) frame in the rest of the paper.

B. SPATIAL AND TEMPORAL FEATURES

Based on the previously generated viewports, we compute for each pair of reference and distorted viewports, $V_f^{\mathcal{R},n}$ and



(a)

(b)

FIGURE 4. **Examples of a frame on the ERP projection domain (a) and an aggregate frame with viewports (b) computed from it using the uniform sampling process.**

$V_f^{\mathcal{D},n}$, $0 \leq n < N$, $0 \leq f < F$ a set of M objective metrics, denoted as:

$$\mathrm{Q}_f^n = \left\{Q_0\left(V_f^{\mathcal{R},n}, V_f^{\mathcal{D},n}\right), \ldots, Q_{m-1}\left(V_f^{\mathcal{R},n}, V_f^{\mathcal{D},n}\right)\right\} \quad (1)$$

In particular, the following spatial quality metrics are computed for each viewport pair:

- Spatial Activity (SA),
- PSNR-HVS and PSNR-HVS-M [31],
- Multi-Scale Structural Similarity (MS-SSIM) [41], and
- Gradient Magnitude Similarity Deviation (GMSD) [44]

and the following temporal quality metrics:

- Relative change in the temporal information (R-TI), and
- Temporal Gradient Magnitude Similarity Deviation (T-GMSD)

The above metrics were selected mainly because it has been shown that they correlated well with subjective scores for traditional image quality assessment and they complement each other with regards to different 2D image distortions [44]. Nevertheless, the proposed method can also be easily adapted to other metrics, which, for instance, better reflect the distortions in a specific dataset. The rest of this subsection details each of the above metrics.

1) SPATIAL ACTIVITY

The spatial activity (SA) of a pair of frames is defined as the root mean square (RMS) difference between the Sobel maps of each of the frames [16]. The Sobel operator, S, is defined as:

$$S(z) = \sqrt{(G_1 * z)^2 + \left(G_1^T * z\right)^2} \quad (2)$$

where z is the frame picture and $*$ denotes the 2-dimensional convolution operation, G_1 is the vertical Sobel filter, given by:

$$G_1 = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} \quad (3)$$

and G_1^T is the transpose of G_1 (horizontal Sobel filter).

Let u and v be $V_f^{\mathcal{R},n}$ and $V_f^{\mathcal{D},n}$, i.e., the same viewport sampled from the same frame f from both reference and distorted content, respectively. We define the difference between the Sobel maps of both frames as:

$$s = S(u) - S(v). \quad (4)$$

Then, we compute SA as:

$$SA(v, u) = \sqrt{\frac{1}{MN}\sum_{i,j}|s_{ij}|^2} \quad (5)$$

where i, j are respectively the horizontal and vertical indices of s, and M and N are the height and width of the viewports, respectively.

2) PSNR-HVS AND PSNR-HVS-M

PSNR-HVS [15] and PSNR-HVS-M [31] are two models that have been designed to improve the performance of PSNR taking into consideration the HVS properties. PSNR-HVS divides the image into 8x8 pixels non-overlapping blocks. Then the $\delta(i, j)$ difference between the original and the distorted blocks is weighted for every 8x8 block by the coefficients of the Contrast Sensitivity Function (CSF). PSNR-HVS-M [31] is defined similarly, but the difference between the DCT coefficients is further multiplied by a contrast masking metric (CM) for every 8x8 block.

3) MS-SSIM

The structural similarity (SSIM) metric divides the job of computing the similarity between two images into three comparisons: luminance, contrast, and structure, respectively defined as:

$$l(\mathbf{x}, \mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (6)$$

$$c(\mathbf{x}, \mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (7)$$

$$s(x, y) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (8)$$

where μ_x and μ_y denote the mean luminance intensities of the compared image signals $\mathbf{x}$ and $\mathbf{y}$; σ_x and σ_y are the standard deviations of the luminance samples of the two images; σ_{xy} is the covariance of the luminance samples; and C_1 and C_2 are stabilizing constants. For an image with a dynamic range L, $C_1 = (K_1 L)^2$ where K_1 is a small constant such that C_1 only takes effect when $(\mu_x^2 + \mu_y^2)$ is small. The SSIM index is then defined as:

$$SSIM(\mathbf{x}, \mathbf{y}) = [l(\mathbf{x}, \mathbf{y})]^{\alpha} \cdot [c(\mathbf{x}, \mathbf{y})]^{\beta} \cdot s(\mathbf{x}, \mathbf{y})]^{\gamma} \tag{9}$$

where α, β, and γ are positive parameters that adjust the relative importance of the three comparison functions. Setting $\alpha = \beta = \gamma = 1$ and $C_3 = C_2/2$ gives the specific form:

$$SSIM(x, y) = \frac{\left(s\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C2\right)}{\left(\mu^2 + \mu_y^2 + C1\right)\left(\sigma_x^2 + \sigma_y^2 + C2\right)} \tag{10}$$

MultiScale-SSIM (MS-SSIM) is an extension of SSIM for multiple scales. At every scale, MS-SSIM applies a low pass filter to the reference and distorted images and downsample the filtered images by a factor of two. At the mth scale, contrast and structure terms are taken into account:

$$MSSSIM(\mathbf{x}, \mathbf{y}) = [l_M(\mathbf{x}, \mathbf{y})]^{\alpha} \cdot \prod_{m=1}^{M} [c_m(\mathbf{x}, \mathbf{y})]^{\beta} \cdot [s_m(\mathbf{x}, \mathbf{y})]^{\gamma}. \tag{11}$$

4) GMSD

Gradient Magnitude Similarity Deviation (GMSD) [44] is based on the standard deviation of the gradient magnitude similarity map, GMS, which is computed as:

$$GMS(u, v) = \frac{2 \cdot m(u) \cdot m(v) + c}{m(u)^2 + m(v)^2 + c} \tag{12}$$

where u and v are respectively the current and previous frame; c is a positive constant that guarantees stability; and $m(z)$ is:

$$m(z) = \sqrt{(z * G_2)^2 + \left(z * G_2^T\right)^2} \tag{13}$$

where $*$ denotes the convolution operator, G_2 represents the vertical Prewitt filter:

$$G_2 = \begin{bmatrix} \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \\ \frac{1}{3} & 0 & -\frac{1}{3} \end{bmatrix}. \tag{14}$$

G_2^T is the transpose of G_2, i.e., the horizontal Prewitt filter. The GMSD index is then computed as:

$$GMSD(u, v) = \sqrt{\frac{1}{NM} \sum_{i,j} \left(GMS(u, v) - \overline{GMS(u, v)}\right)^2}, \tag{15}$$

where $\overline{GMS(u, v)}$ is the gradient magnitude similarity mean, computed as:

$$\overline{GMS(u, v)} = \frac{1}{NM} \sum_{i,j} GMS(u, v). \tag{16}$$

5) RELATIVE CHANGE IN TEMPORAL INFORMATION

Temporal information (TI) [19] is an indicator that characterizes the amount of motion in a video and is defined as the standard deviation of the difference between two frames:

$$\Delta F_n = F_n - F_{n-1} \tag{17}$$

$$TI[F_n] = std(\Delta F_n) \tag{18}$$

Here, we define the relative change in the temporal information as:

$$TI_{rel}[F_n] = \frac{|TI_{ref}[F_n] - TI_{dist}[F_n]|}{TI_{ref}[F_n]}, \tag{19}$$

where $TI_{ref}[F_n]$ and $TI_{dist}[F_n]$ are respectively the TI for the frame F_n in the reference and distorted videos.

6) TEMPORAL GMSD

is defined as the GMSD score between the difference of two consecutive references frames and two consecutive distorted frames, i.e.:

$$\Delta F_r(f) = F_r(f) - F_r(f-1) \tag{20}$$

$$\Delta F_d(f) = F_d(f) - F_d(f-1) \tag{21}$$

$$\text{T-GMSD}(f) = GMSD(\Delta F_r(f), \Delta F_d(f)). \tag{22}$$

C. TEMPORAL FEATURE POOLING

The per-viewports metrics Q_f^n for each frame, $f \in \{1, \ldots, F\}$ and viewport $n \in \{1, \ldots, N\}$, are integrated to yield the overall quality of each viewport: Q_{pool}^n. This integration is performed by the temporal pooling module. Our proposal is modular and can be adapted to different temporal pooling methods. Based on the experiments of Section V-C, and inspired by [27], we propose a temporal pooling method considering the characteristics of the HVS, in particular:

- *the smooth effect*, i.e., the subjective ratings of the whole video sequence typically demonstrate far less variations than the frame-level quality scores;
- *the asymmetric effect*, i.e., HVS is more sensitive to frame-level quality degradation than to improvement; and
- *recency effect*, i.e., subjects tend to put a higher weight on what they have seen most recently.

More precisely, for each viewport n, we first process the original scores considering both smooth and asymmetric effects as:

$$Q_{LP}^n(f) = \begin{cases} Q_{LP}^n(f-1) + \alpha \cdot \Delta Q(f), & \text{if } \Delta Q^n \leq 0 \\ Q_{LP}^n(f-1) + \beta \cdot \Delta Q(f), & \text{if } \Delta Q^n > 0 \end{cases} \tag{23}$$

where $\Delta Q^n = Q_{frame}^n(f) - Q_{LP}^n(f-1)$ and $Q_{LP}^n = Q_{frame}^n(1)$, and α and β control the asymmetric weights. Then, we perform a weighted-average sum of the above processed scores considering the recency effect:

$$Q_{pool}^n = \frac{1}{F} \sum_{f=1}^{F} Q_{LP}^n(f) w(f). \tag{24}$$

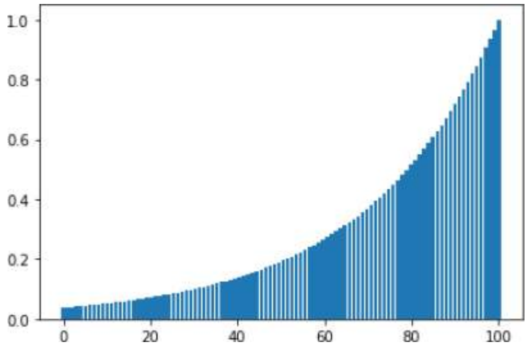


FIGURE 5. Exponential weights used to average per-frame quality metrics.

Here, we define the weights $w(f)$ as an exponential function (see Fig. 5):

$$w(f) = e^{(((f+1)-F)/\tau)}. \tag{25}$$

Similar to [27], in our experiments we use $\alpha = 0.03$ and $\beta = 0.2$.

D. REGRESSION

After the temporal pooling, we end up with M features for each viewport, which are then concatenated as a feature vector, Q $= [Q_0^0, \ldots, Q_{m-1}^0, Q_0^1, \ldots, Q_{m-1}^1, Q_0^{n-1}, \ldots, Q_{m-1}^{n-1}]$. Such a vector is used for learning a non-linear mapping between the computed per-viewport features and the subjective DMOS scores of 360-degree videos. In our framework, we have tested three different regression methods:

- Support Vector Regression (SVR) [5];
- Gradient Boosting regression (GBR) [17]; and
- Random Forest Regression (RFR) [7].

Based on the experiments on Section IV, we have chosen RFR as our final regression method because it significantly outperformed the other methods. Next we detail our experiments setup, hyper-parameter tuning, training, and test processes.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

Unless specified otherwise, we validate our proposal based on the VQA-ODV [21] dataset. VQA-ODV is the largest publicly available dataset today and is composed of three types of projections, ERP, RCMP, and TSP, and 3-levels of H.265 distortions, quantization parameters (QP)=27, 37, and 42. In total, there are 60 different reference sequences (12 in raw format and others downloaded from YouTube VR channel) and 540 distorted sequences that were rated by 221 participants. In all our following experiments, we extract only the ERP sequences from VQA-ODV, resulting in 180 distorted sequences. Both MOS (Mean Opinion Scores) and DMOS (Differential Mean Opinion Scores) are available for the dataset.

We compare our method to PSNR, S-PSNR, WS-PSNR MS-SSIM, and VMAF, using common criteria for the evaluation of objective quality metrics: Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), and Root Mean Squared Error (RMSE). SROCC measures the prediction monotonicity while PLCC and RMSE measure the prediction accuracy. Higher SROCC, PLCC and lower RMSE indicate good correlation with subjective scores.

Moreover, we compare the performance of our method and the other objective metrics when the features are computed in: i) the projection domain (“Proj.”); ii) all viewports merged in a collage frame (“VP-Collage”) (see Fig. 4); and iii) the viewports considered indiviually ("VP"), i.e., the metrics are computed independently for each viewports (as discussed in Section III). Computing the objective metrics in the viewport collage frame is similar to averaging the quality of all the viewports.

Based on the above 3 different modes, we performed the following experiments:

- using the same fixed train/test subset of the VQA-ODV dataset used in [21], [22] (Section IV-A);
- a cross-validation experiment in the whole ERP sequences of VQA-ODV dataset (Section IV-B); and
- a cross dataset validation Section IV-C.

In all the above cases, we perform grouped split of the data. Such approach divides the database into two content-independent subsets (training and testing) ensuring that videos generated from one reference (i.e., the same content) in the testing subset are not present in the training subset, and vice-versa. Also, in the above following, we use a uniform sampling with a 40-deg field of view for the viewports, which resolution matches the HMD resolution used in the dataset (an HTC Vive).

A. VQA-ODV: FIXED TRAIN/TEST SUBSETS

Table 1 shows the results of our method using the same (fixed) train/test selection of [21], [22], which ensures that the same sequence content is not part of both train and test sets. After separating the dataset into train and test sets, we first run a group shuffle cross-validation on the training data to find the best random forest hyper-parameters to predict the 360-VQA. Based on the resulting hyper-parameters we then train and validate the model, respectively using the previously separated test data. Then, we compute the performance of the model to predict the scores of the sequences in the test set. By following the above procedure, we make sure that the hyper-parameter tunning never sees the test data.

We compare the performance of our method against: PSNR, MS-SSIM, and VMAF (state-of-the-art metrics for 2D quality assessment) when the features are computed in both the projection and the viewports domain; and S-PSNR and WS-PSNR, metrics specifically developed for 360 VQA. For the non-learned metrics (PSNR, MS-SSIM, S-PSNR, and WS-PSNR), we emulate the training phase by fitting the following 4-parameter logistic function with the train set and

TABLE 1. Fixed train/test set test results (VQA-ODV dataset).

Metric	PLCC ↑	SROCC ↑	RMSE ↓
PSNR	0.725	0.738	8.176
PSNR (VP-Collage)	0.762	0.763	7.582
S-PSNR	0.751	0.770	7.756
WS-PSNR	0.743	0.561	7.950
MS-SSIM (Proj.)	0.760	0.789	7.874
MS-SSIM (VP-Collage)	0.817	0.841	7.002
VMAF (Proj.)	0.797	0.794	7.248
VMAF (VP-Collage)	0.845	0.856	6.271
Ours (SVR, Proj.)	0.804	0.836	7.251
Ours (SVR, VP-Collage)	0.858	0.905	8.121
Ours (SVR, VP) (Overfitting)	0.513	0.724	10.412
Ours (GB, Proj.)	0.847	0.850	6.474
Ours (GB, VP-Collage)	0.857	0.854	6.252
Ours (GB, VP)	0.915	0.889	5.041
Ours (RFR, Proj.)	**0.856**	**0.869**	**6.359**
Ours (RFR, VP-Collage)	**0.906**	**0.893**	**5.550**
Ours (RFR, VP)	**0.929**	**0.917**	**4.561**

then computing its performance with the test set:

$$s' = \frac{\beta_1 - \beta_2}{1 + e^{-\frac{S-\beta_3}{||\beta_4||}}} + \beta_2. \tag{26}$$

Finally, for our method, we compare different regression techniques: Support Vector Regression (SVR), Gradient Boosting Regression (GBR), and Random Forest Regression (RFR).

Discussion: From Table 1, the best performance for the fixed train/test set on VQA-ODV is achieved by our method using features computed separatly for each viewport ("VP" mode). Such results can be explained by both the viewports being closer to what the users see and the model being able to learn the most important viewports, which is not the case when using the "VP-Collage" mode. Our results in Table 1 also show that computing objective metrics on the viewport domain ("VP-Collage") improve the performance of all the tested metrics (LCC gain around 0.05) validating our hypothesis that viewports better represents the perceived quality when users watch a 360 video on an HMD when compared to the projection domain. Finally, it is also interesting to note that our method (on both "Proj." and "VP-Collage") outperforms VMAF, which can be explained by the choice of objective metrics, the temporal pooling, and regression methods in our method. Finally, with regards to generalization of our method, Table 2 also brings the training performance of our different regression methods. When comparing Table 1 and 2 it is clear that RFR is the most robust regression method among the tested ones, providing PLCC values while avoiding overfitting.

B. VQA-ODV: CROSS-VALIDATION

To avoid bias on the specific train/test set used above, we also performed a full cross-validation on the VQA-ODV dataset. In the cross-validation experiments, we performed a 1000x randomly group selection of 80%/20% train/test splitting of the dataset, and then computed the average PLCC, SROCC, and RMSE of the models. The group selection ensures that there is no overlap between content in the training and test sets. The hyper-parameters used for the RF regression are the same ones found on the fixed train/test experiments above. Table 3 shows the average PLCC, SROCC, and RMSE results for the cross-validation experiments, and Fig. 6 depicts the distribution of the correlation scores through a violin plot. Larger sections of the violin plots depict a higher probability of achieving these correlation scores, while narrower sections depict a lower probability.

TABLE 2. Training performance of the different regression methods on the fixed train set of VQA-ODV.

Metric	PLCC ↑	SROCC ↑	RMSE ↓
Ours (SVR, Proj.).	0.751	0.775	7.788
Ours (SVR, VP-Collage)	0.781	0.810	7.455
Ours (SVR, VP)	1.000	1.000	0.080
Ours (GB, Proj.).	0.990	0.989	1.737
Ours (GB, VP-Collage)	0.989	0.987	1.798
Ours (GB, VP)	0.999	0.999	0.410
Ours (RFR, Proj.).	0.935	0.946	4.259
Ours (RFR, VP-Collage)	0.966	0.972	3.224
Ours (RFR, VP)	0.983	0.984	2.420

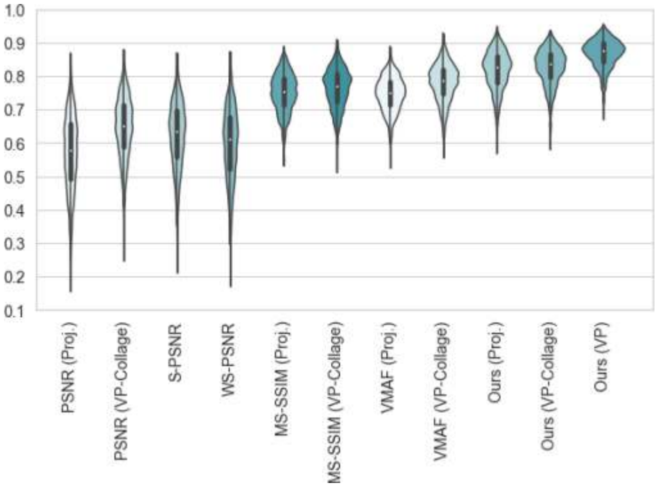


FIGURE 6. Violin plots for the GroupShuffle (80%/20%) cross-validation PLCC performance on VQA-ODV dataset.

TABLE 3. Average of GroupShuffle cross validation (80%/20%) performance on VQA-ODV.

Metric	PLCC ↑	SROCC ↑	RMSE ↓
PSNR (Proj.)	0.572	0.619	9.825
PSNR (VP-Collage)	0.647	0.686	9.122
S-PSNR	0.625	0.667	9.346
WS-PSNR	0.598	0.645	9.598
MS-SSIM (Proj.)	0.750	0.775	7.935
MS-SSIM (VP-Collage)	0.764	0.791	7.758
VMAF	0.747	0.767	7.963
VMAF (VP-Collage)	0.781	0.798	7.515
Ours (RFR, Proj.)	**0.817**	**0.829**	**6.872**
Ours (RFR, VP-Collage)	**0.827**	**0.826**	**6.738**
Ours (RFR, VP)	**0.868**	**0.868**	**5.937**

TABLE 4. Objective metrics performance on VR-VQA48. Our method is trained on VQA-ODV and tested on VR-VQ48.

Metric	PLCC ↑	SROCC ↑	RMSE ↓
PSNR (Proj.)	0.499	0.508	10.732
PSNR (VP-Collage)	0.613	0.612	9.843
S-PSNR	0.569	0.595	10.183
CPP-PSNR	0.567	0.595	10.198
WS-PSNR	0.548	0.562	8.804
MS-SSIM (Proj.)	0.759	0.751	8.490
MS-SSIM (VP-Collage)	0.843	0.829	7.634
VMAF (Proj.)	0.783	0.771	7.712
OV-PSNR[PSNR] [18]	0.837	0.890	6.749
OV-PSNR[S-PSNR] [18]	0.818	0.775	7.123
OV-PSNR[CPP-PSNR] [18]	0.837	0.787	5.181
OV-PSNR[WS-PSNR] [18]	0.838	0.790	5.157
Ours (SVR, Proj.).	0.795	0.775	7.898
Ours (SVR, VP-Collage)	0.876	0.855	6.981
Ours (SVR, VP) (overfitting)	0.699	0.775	9.559
Ours (GB, Proj.).	0.787	0.780	8.238
Ours (GB, VP-Collage)	0.891	0.881	6.031
Ours (GB, VP)	0.949	0.940	5.724
Ours (RFR, Proj.)	**0.837**	**0.820**	**7.689**
Ours (RFR, VP-Collage)	**0.925**	**0.900**	**5.429**
Ours (RFR, VP)	**0.956**	**0.949**	**5.161**

Discussion: Table 3 further validates our previously conclusions from the fixed train/test dataset settings, namely: i) considering viewports ("VP-Collage" and "VP" modes) instead of the projection domain improves the performance of objective quality metrics; ii) considering features on individual viewpors ("VP" mode) allows the model to further weight the different views based on the importance of those regions to the final quality of the 360 video. Moreover, Figure 6 shows that besides having a better average, our method also provide a higher density on the high PLCC values.

C. VQA-ODV X VR-VQA48: CROSS DATASET VALIDATION

To demonstrate the perfomance of our proposal in more than one dataset, Table 4 shows the results of the performance of our model trained on the ERP sequences of VQA-ODV and tested on public available VR-VQA48 dataset [43]. VR-VQA48 is composed of 12 original omnidirectional video sequences (YUV 4:2:0 format at the resolution of 4096 × 2048) and 36 corresponding impaired sequences by encoding each original sequence with 3 different bitrate settings. 48 subjects rated raw subjective quality scores for all the 48 sequences. MOS and DMOS values are available for the sequences per subject. We consider the final quality score of each sequence as the average DMOS. As in the previous setups, we also use the uniform viewport sampling with a 40-degree of field-of-view. In total, the model is trained on the 180 ERP sequences of and tested on the 36 distorted sequences of VR-VQA48. Also, from the previous tables, Table 4 also add the results of the method proposed by [18] as reported in the original work.

TABLE 5. Training performance of the different regression methods trained on VQA-ODV.

Metric	PLCC ↑	SROCC ↑	RMSE ↓
Ours (SVR, Proj.).	0.771	0.796	7.523
Ours (SVR, VP-Collage)	0.792	0.815	7.241
Ours (SVR, VP) (Overfitting)	1.000	1.000	0.080
Ours (GB, Proj.).	0.986	0.820	7.689
Ours (GB, VP-Collage)	0.983	0.979	2.240
Ours (GB, VP)	0.999	0.999	0.470
Ours (RFR, Proj.).	0.945	0.820	7.689
Ours (RFR, VP-Collage)	0.970	0.973	3.002
Ours (RFR, VP)	0.987	0.987	2.109

TABLE 6. Performance of our proposal using SFFS on fixed train/test VQA-ODV dataset.

Selected features (RFR, Proj.)	PLCC ↑	SROCC ↑	RMSE ↓
GMSD	0.808	0.770	6.890
GMSD, R-TI	0.874	0.878	5.831
GMSD, R-TI, PSNR-HVS	0.885	0.879	5.726
GMSD, R-TI, PSNR-HVS, PSNR-HVS-M	0.891	0.909	5.507
Selected features (RFR, VP-Collage)	PLCC ↑	SROCC ↑	RMSE ↓
PSNR-HVS-M	0.841	0.826	6.470
PSNR-HVS-M, R-TI	0.888	0.869	5.586
PSNR-HVS-M, R-TI, GMSD	0.911	0.901	5.372
PSNR-HVS-M, R-TI, GMSD, T-GMSD	0.916	0.901	5.392
Selected features (RFR, VP)	PLCC ↑	SROCC ↑	RMSE ↓
PSNR-HVS-M	0.877	0.874	5.695
PSNR-HVS-M, R-TI	0.921	0.934	5.221
PSNR-HVS-M, R-TI, GMSD	0.946	0.943	4.924
PSNR-HVS-M, R-TI, GMSD, T-GMSD	0.946	0.940	4.971

Discussion: The results in Table 4 confirm our previous conclusions, showing that our method is also robust across different datasets. Of course, if the distortions included in a new dataset are not well modeled by the individual features of our model, we would not expect to achieve good results on that dataset. That does not seem the case for the VR-VQA48 dataset. Table 5 brings the training performance, further validating that RFR is the best regression method among the tested ones.

V. ABLATION STUDIES

In this section we provide additional experiments that allow us to better understand the influence of viewport sampling (Section V-B); the importance of different individual features (Section V-A); and the influence of different temporal pooling methods (Section V-C).

A. FEATURES SELECTION

To study the importance of each feature for our model we performed additional experiments based on the Sequential Forward Feature Selection (SFFS) method. In such a method, we start with an empty set of features ($M = m_i | m = 1, \ldots, f$), and for each step we select the next previously still not selected feature (m^*) that maximize a specific metric. We focus on minimizing PLCC in our experiments. Table 6 shows the results for 1 to 4 selected features on the fixed train/test setup. As can be seen in Table 6, we can improve even further our results with just a subset of the initially

TABLE 7. Individual objective metrics performance on VQA-ODV (mean temporal pooling). Tropical and Uniform metrics are computed on VP-Collage frames.

	Proj. (ERP)		Tropical (30deg)		Uniform (30deg)		Tropical (40deg)		Uniform (40deg)		Uniform (50deg)	
Metric	LCC ↑	SROCC ↑	LCC ↑	SROCC ↑	LCC ↑	SROCC ↑	LCC ↑	SROCC ↑	LCC ↑	SROCC ↑	LCC ↑	SROCC ↑
PSNR	0.571	0.602	0.606	0.639	**0.652**	**0.687**	0.621	0.650	0.648	0.678	0.646	0.676
PSNR-HVS	0.642	0.678	0.648	0.696	0.659	0.695	0.637	0.667	**0.710**	**0.740**	0.710	0.743
PSNR-HVS-M	0.703	0.728	0.673	0.694	0.692	0.725	0.677	0.703	**0.751**	**0.777**	0.748	0.775
SSIM	0.572	0.605	0.584	0.622	0.631	0.668	0.594	0.630	**0.695**	**0.727**	0.691	0.723
MS-SSIM	0.747	0.769	0.718	0.749	0.741	0.769	0.727	0.756	**0.757**	**0.785**	0.748	0.775
GMSD	0.734	0.752	0.694	0.719	0.722	0.748	0.703	0.726	**0.783**	**0.806**	0.779	0.801
VIFP	**0.652**	**0.715**	0.633	0.674	0.654	0.717	0.629	0.639	0.644	0.673	0.644	0.676
SA	0.499	0.532	0.508	0.540	0.564	0.595	0.522	0.552	**0.665**	**0.701**	0.662	0.694

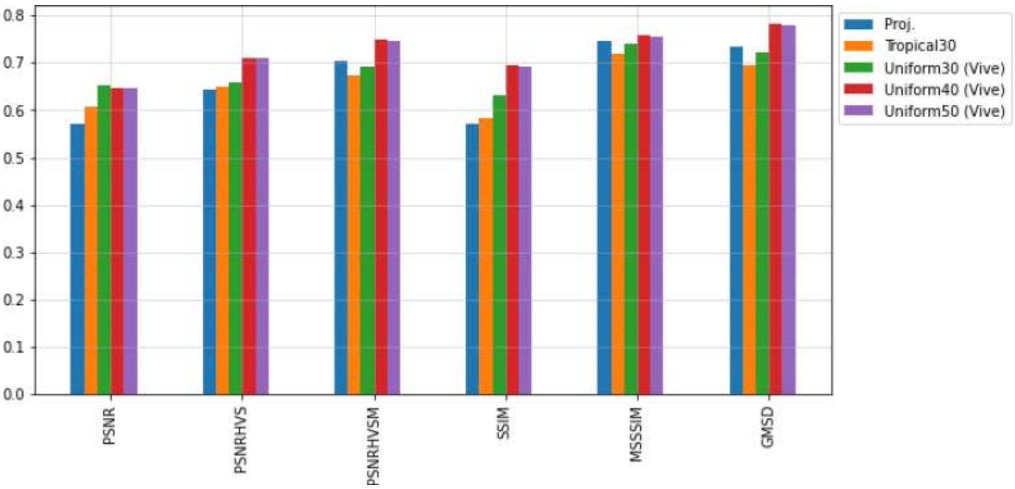


FIGURE 7. Performance (PLCC) of individual objective viewport metrics.

proposed features. By having a subset of features, however, it is possible that the model does not generalize as well to other distortion types. More experiments, with datasets including other distortions type, should be performed in the future. One of the main issues that prevented us for performing such experiments is the lack of availability of such datasets, which we see as an important future work for the research community.

B. INFLUENCE OF VIEWPORT SAMPLING

When considering a viewport-based metric for 360-degree videos there are (in theory) infinite ways on how to sample the viewports on the sphere. To better understand the visual quality of objective metrics computed in the projection domain against the same metrics computed in the viewports, we report here a study on objective quality metrics on the VQA-ODV dataset. For that, we computed the following objective metrics in both the projection and viewport domains: PSNR, PSNR-HVS, PSNR-HVSM, SSIM, MSSSIM, GMSD, Spatial activity, and Temporal activity. In the viewport domain, we consider different field-of-views and viewports sampling patterns. In both cases (projection and viewport domains) the features are computed individually for each frame and then pooled with an average method.

For the following experiments, we have chosen the *uniform* and *tropical* sampling methods shown in Fig. 3 [6] with FoVs of 30, 40, and 50 degrees. In total, there are 16 and 25 viewports for the tropical and uniform sampling methos, respectively. Finally, we compute the Pearson and Spearman Correlation Coefficient between the DMOS and the fitted 4-parameter logistic regression, given by Equation (26).

Table 7 shows the LCC and SROCC performance of the different viewport sampling and field-of-views compared with the same metrics computed in the frame domain. Fig. 7 plots the LCC values performance. From those results, we can conclude that the *uniform* sampling with 40-deg sampling is the one with the best overall performance.

C. INFLUENCE OF TEMPORAL POOLING

Different temporal poolings might also result in different performance [39]. To better understand the performance of our method when using different temporal pooling methods, we also performed an experiment on the same architecture of Figure 1 when only changing the temporal pooling method. Despite our proposed temporal pooling method, we also tested:

a) *Arithmetic mean:* The sample mean of frame-level scores:

$$Q = \frac{1}{N}\sum_{n=1}^{N} q_n. \tag{27}$$

TABLE 8. **Temporal pooling performance on VQA-ODV fixed train/test sets (Ours (RFR, VP)).**

Temporal pooling	PLCC ↑	SROCC ↑	RMSE ↓
Minkowski4	0.900	0.881	5.798
Minkowski2	0.910	0.903	5.361
Mean	0.918	0.908	5.364
Percentile	0.919	0.905	5.415
HVS (Ours)	**0.929**	**0.917**	**4.561**

b) *Minkowski mean:* The L_p Minkowski summation of time-varying quality is defined as:

$$Q = \left(\frac{1}{N} \sum_{n=1}^{p} q_n^p \right)^{1/p}. \tag{28}$$

c) *Percentile:* Percentile pooling is based on observed phenomenon that perceptual quality is heavily affected by the "worst" parts of the content. Many prior works have studied and justified (or challenged) percentile pooling [15]–[18], [20]. The k-th percentile pooling is expressed:

$$Q = \frac{1}{\left| P_{\downarrow k\%} \right|} \sum_{n \in P_{\downarrow k\%}} q_n \tag{29}$$

Table 8 shows the performance of our method (VP) using the different pooling methods on the VQA-ODV dataset. For the percentile method, we use $k = 10$ (i.e., we used only the 10% worst scores of the frames. For the Minkowski mean we report the results on both $p = 2$ and $p = 4$. From the results, we can conclude that our proposal performs better with our HVS-based pooling proposed in Section III.

VI. CONCLUSION

We propose the use of viewport-based multi-metrics fusion for 360-degree VQA and discuss the lessons learned by implementing and evaluating such an approach on two publicly available 360 video datasets. The computation of features in viewports implies that our metric can be applied on a variety of projections, and our experiments demonstrate that the MMF approach is capable of achieving state-of-the-art results while requiring much less training data than deep learning techniques.

As future work, we plan: (i) to consider color and visual attention; (ii) to consider fixation in our current temporal pooling; and (iii) extend our proposed method to no-reference video quality assessment. Finally, it is important to highlight that although our method achieves impressive performance on the available 360 videos quality datasets, there is still need in the immersive media community to produce more challenging datasets, considering different distortions, projections, and HMD devices. Overall, our proposed method can also be easily adaptable to new projection types and other individual objective metrics that better maps the distortions on specific 360 video processing contexts as well, which is also another interesting future work direction.

REFERENCES

[1] C. Li, M. Xu, S. Zhang, and P. L. Callet, "State-of-the-art in 360° video/image processing: Perception, assessment and compression," 2019. [Online]. Available: http://arxiv.org/abs/1905.00161

[2] R. G. D. A. Azevedo, N. Birkbeck, F. De Simone, I. Janatra, B. Adsumilli, and P. Frossard, "Visual distortions in 360° videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 8, pp. 2524–2537, Aug. 2020, doi: 10.1109/TCSVT.2019.2927344.

[3] R. G. D. A. Azevedo, N. Birkbeck, I. Janatra, B. Adsumilli, and P. Frossard, "Subjective and viewport-based objective quality assessment of 360 videos," in *Proc. Electron. Imag. Image Qual. Syst. Perform. XVII* Burlingame, CA, USA, 2020, p. 6.

[4] R. G. D. A. Azevedo, N. Birkbeck, I. Janatra, B. Adsumilli, and P. Frossard, "A viewport-driven multi-metric fusion approach for 360-degree video quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, London, U.K., 2020, pp. 1–6.

[5] D. Basak, S. Pal, and D. Patranabis, "Support vector regression," *Neural Inf. Process. Lett. Rev.*, vol. 11, no. 10, pp. 203–224, 2007.

[6] N. Birkbeck, C. Brown, and R. Suderman, "Quantitative evaluation of omnidirectional video quality," in *Proc. 9th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, Erfurt, Germany, 2017, pp. 1–3, doi: 10.1109/QoMEX.2017.7965684.

[7] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[8] D. M. Chandler and E. C. Larson, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010, Art. no. 011006, doi: 10.1117/1.3267105.

[9] S. Chen, Y. Zhang, Y. Li, Z. Chen, and Z. Wang, "Spherical structural similarity index for objective omnidirectional video quality assessment," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, San Diego, CA, USA, 2018, pp. 1–6, doi: 10.1109/ICME.2018.8486584.

[10] Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation," *Signal Process.*, vol. 146, pp. 66–78, May 2018. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0165168418300057

[11] S. Croci, C. Ozcinar, E. Zerman, J. Cabrera, and A. Smolic, "Voronoi-based objective quality metrics for omnidirectional video," in *Proc. 11th Int. Conf. Qual. Multimedia Exp. (QoMEX)*, Berlin, Germany, 2019, pp. 1–6.

[12] S. Croci, C. Ozcinar, E. Zerman, J. Cabrera, and A. Smolic, "Visual attention-aware quality estimation framework for omnidirectional video using spherical voronoi diagram," *Qual. User Exp.*, vol. 5, p. 4, Apr. 2020.

[13] F. De Simone, P. Frossard, C. Brown, N. Birkbeck, and B. Adsumilli, "Omnidirectional video communications: New challenges for the quality assessment community," *Video Qual Experts Group eLett.*, vol. 3, no. 1, pp. 18–25, 2017.

[14] F. De Simone, P. Frossard, P. Wilkins, N. Birkbeck, and A. Kokaram, "Geometry-driven quantization for omnidirectional image coding," in *Proc. Picture Coding Symp. (PCS)*, 2016, pp. 1–5, doi: 10.1109/PCS.2016.7906402.

[15] K. Egiazarian, J. Astola, N. Ponomarenko, V. Lukin, F. Battisti, and M. Carli, "New full-reference quality metrics based on HVS," in *Proc. 2nd Int. Workshop Video Process. Qual. Metrics*, vol. 4, 2006, pp. 22–24.

[16] P. G. Freitas, W. Y. L. Akamine, and M. C. Q. Farias, "Using multiple spatio-temporal features to estimate video quality," *Signal Process. Image Commun.*, vol. 64, pp. 1–10, May 2018, doi: 10.1016/j.image.2018.02.010.

[17] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[18] P. Gao, P. Zhang, and A. Smolic, "Quality assessment for omnidirectional video: A spatio-temporal distortion modeling approach," *IEEE Trans. Multimedia*, early access, Dec. 16, 2020, doi: 10.1109/TMM.2020.3044458.

[19] "Subjective video quality assessment methods for multimedia applications," ITU, Geneva, Switzerland, ITU-T Recommendation P.910, 1999. [Online]. Available: http://www.itu.int/rec/T-REC-P.910-200804-I

[20] H. G. Kim, H.-T. Lim, and Y. M. Ro, “Deep virtual reality image quality assessment with human perception guider for omnidirectional image,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 917–928, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8638985/

[21] C. Li, M. Xu, X. Du, and Z. Wang, “Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model,” in *Proc. 26th ACM Int. Conf. Multimedia*, Seoul, South Korea, 2018, pp. 932–940, doi: 10.1145/3240508.3240581.

[22] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, “Viewport proposal CNN for 360deg video quality assessment,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, p. 10.

[23] S. Li, F. Zhang, L. Ma, and K. N. Ngan, “Image quality assessment by separately evaluating detail losses and additive impairments,” *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 935–949, Oct. 2011.

[24] W. Lin and C. C. J. Kuo, “Perceptual visual quality metrics: A survey,” *J. Vis. Commun. Image Represent.*, vol. 22, no. 4, pp. 297–312, 2011, doi: 10.1016/j.jvcir.2011.01.005.

[25] T.-J. Liu, W. Lin, and C.-C. J. Kuo, “Image quality assessment using multi-method fusion,” *IEEE Trans. Image Process.*, vol. 22, pp. 1793–1807, 2013, doi: 10.1109/TIP.2012.2236343.

[26] T.-J. Liu, W. Lin, and C.-C. J. Kuo, “A multi-metric fusion approach to visual quality assessment,” in *Proc. 3rd Int. Workshop Qual. Multimedia Exp. (QoMEX)*, Mechelen, Belgium, 2011, pp. 72–77. [Online]. Available: http://ieeexplore.ieee.org/document/6065715/

[27] Y. Lu, M. Yu, and G. Jiang, “Low-complexity video quality assessment based on spatio-temporal structure,” in *Proc. Int. Conf. Inf. Softw. Technol.*, 2019, pp. 408–415. [Online]. Available: http://link.springer.com/10.1007/978-3-030-30275-7_31

[28] F. Pearson, *Map Projections: Theory and Applications*. Boca Raton, FL, USA: CRC Press, 1990.

[29] B. S. Phadikar, G. K. Maity, and A. Phadikar, “Full reference image quality assessment: A survey,” in *Industry Interactive Innovations in Science, Engineering and Technology*, vol. 11, S. Bhattacharyya, S. Sen, M. Dutta, P. Biswas, and H. Chattopadhyay, Eds. Singapore: Springer, 2018, pp. 197–208.

[30] M. H. Pinson and S. Wolf, “A new standardized method for objectively measuring video quality,” *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[31] N. Ponomarenko, J. Astola, V. Lukin, and F. Silvestri, “On between-coefficient contrast masking of DCT basis functions,” in *Proc. 3rd Int. Workshop Video Process. Qual. Metrics*, 2007, p. 4.

[32] R. Rassool, “VMAF reproducibility: Validating a perceptual practical video quality metric,” in *Proc. IEEE Int. Symp. Broadband Multimedia Syst. Broadcast. (BMSB)*, 2017, pp. 1–2, doi: 10.1109/BMSB.2017.7986143.

[33] R. Reisenhofer, S. Bosse, G. Kutyniok, and T. Wiegand, “A haar wavelet-based perceptual similarity index for image quality assessment,” *Signal Process. Image Commun.*, vol. 61, pp. 33–43, Feb. 2018. [Online]. Available: https://doi.org/10.1016/j.image.2017.11.001

[34] M. Rousselot, X. Ducloux, O. L. Meur, and R. Cozot, “Quality metric aggregation for HDR/WCG images,” in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2019, pp. 3786–3790, doi: 10.1109/ICIP.2019.8803635.

[35] K. Sampath, P. Venkatesan, P. Ramachandran, and K. Goswami, “Block-based temporal metric for video quality assessment,” in *Proc. IEEE Int. Conf. Consum. Electron. (ICCE)*, 2019, pp. 1–4.

[36] K. Seshadrinathan and A. C. Bovik, “Motion tuned spatio-temporal quality assessment of natural videos,” *IEEE Trans. Image Process.*, vol. 19, pp. 335–350, 2010.

[37] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” *IEEE Trans. Image Process.*, vol. 15, pp. 430–444, 2006, doi: 10.1109/TIP.2005.859378.

[38] Y. Sun, A. Lu, and L. Yu, “Weighted-to-spherically-uniform quality evaluation for omnidirectional video,” *IEEE Signal Process Lett.*, vol. 24, no. 9, pp. 1408–1412, Sep. 2017, doi: 10.1109/LSP.2017.2720693.

[39] Z. Tu, C.-J. Chen, L.-H. Chen, N. Birkbeck, B. Adsumilli, and A. C. Bovik, “A comparative evaluation of temporal pooling methods for blind video quality assessment,” 2020. [Online]. Available: arxiv abs/2002.10651

[40] P. V. Vu and D. M. Chandler, “ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices,” *J. Electron. Imag.*, vol. 23, no. 1, 2014, Art. no. 013016. [Online]. Available: http://electronicimaging.spiedigitallibrary.org/article.aspx?doi=10.1117/1.JEI.23.1.013016

[41] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in *Proc. 37th Asilomar Conf. Signals Syst. Comput.*, 2003, pp. 1398–1402, doi: 10.1109/ACSSC.2003.1292216.

[42] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, pp. 600–612, 2004, doi: 10.1109/TIP.2003.819861.

[43] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, “Assessing visual quality of omnidirectional videos,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 12, pp. 3516–3530, Dec. 2019.

[44] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, “Gradient magnitude similarity deviation: A highly efficient perceptual image quality index,” *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2014, doi: 10.1109/TIP.2013.2293423.

[45] P. Yan, X. Mou, and W. Xue, “Video quality assessment via gradient magnitude similarity deviation of spatial and spatiotemporal slices,” in *Proc. Mobile Devices Multimedia Enabling Technol. Algorithms Appl.*, San Francisco, CA, USA, 2015, Art. no. 94110M. [Online]. Available: http://proceedings.spiedigitallibrary.org/proceeding.aspx?doi=10.1117/12.2083283

[46] M. Yu, H. Lakshman, and B. Girod, “A framework to evaluate omnidirectional video coding schemes,” in *Proc. IEEE Int. Symp. Mixed Augmented Reality (ISMAR)*, 2015, pp. 31–36, doi: 10.1109/ISMAR.2015.12.

[47] V. Zakharchenko, K. P. Choi, and J. H. Park, “Quality metric for spherical panoramic video,” in *Proc. Opt. Photon. Inf. Process. X*, 2016, Art. no. 99700C, doi: 10.1117/12.2235885.

[48] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” *IEEE Trans. Image Process.*, vol. 20, pp. 2378–2386, 2011, doi: 10.1109/TIP.2011.2109730.