applied sciences

Article

IV-SSIM—The Structural Similarity Metric for Immersive Video

Adrian Dziembowski *, Weronika Nowak and Jakub Stankowski

Institute of Multimedia Telecommunications, Poznan University of Technology, 60-965 Poznań, Poland; weronika.nowak.code@gmail.com (W.N.); jakub.stankowski@put.poznan.pl (J.S.)
* Correspondence: adrian.dziembowski@put.poznan.pl

Abstract: In this paper, we present a new objective quality metric designed for immersive video applications—IV-SSIM. The proposed IV-SSIM metric is an evolution of our previous work—IV-PSNR (immersive video peak signal-to-noise ratio)—which became a commonly used metric in research and ISO/IEC MPEG standardization activities on immersive video. IV-SSIM combines the advantages of IV-PSNR and metrics based on the structural similarity of images, being able to properly mimic the subjective quality perception of immersive video with its characteristic distortions induced by the reprojection of pixels between multiple views. The effectiveness of IV-SSIM was compared with 16 state-of-the-art quality metrics (including other metrics designed for immersive video). Tested metrics were evaluated in an immersive video coding scenario and against a commonly used image quality database—TID2013—showing their performance in both immersive and typical, non-immersive use cases. As presented, the proposed IV-SSIM metric clearly outperforms other metrics in immersive video applications, while also being highly competitive for 2D image quality assessment. The authors of this paper have provided a publicly accessible, efficient implementation of the proposed IV-SSIM metric, which is used by ISO/IEC MPEG video coding experts in the development of the forthcoming second edition of the MPEG immersive video (MIV) coding standard.

Keywords: image quality; immersive video; video compression; view rendering; structural similarity

Citation: Dziembowski, A.; Nowak, W.; Stankowski, J. IV-SSIM—The Structural Similarity Metric for Immersive Video. *Appl. Sci.* **2024**, *14*, 7090. https://doi.org/10.3390/app14167090

Academic Editor: Atsushi Mase

Received: 23 July 2024
Revised: 8 August 2024
Accepted: 12 August 2024
Published: 13 August 2024

1. Introduction

The core concept behind immersive video is to grant a user the freedom to virtually navigate within a scene [1], which was captured by a multicamera system consisting of several perspective or omnidirectional cameras (blue cameras in Figure 1). The virtual navigation is enabled by a rendering of virtual views [2], i.e., the reprojection of pixels from input cameras to the virtual camera (orange camera in Figure 1) using spatial representation of the scene, usually given as depth maps and camera parameters (DIBR, depth-image-based rendering) [3].

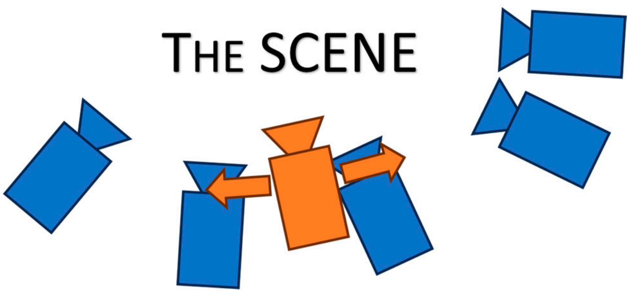

Figure 1. Idea of an immersive video system; blue—real (input) cameras, orange—virtual camera; position of the virtual camera may be arbitrarily changed by the viewer.

In order to provide a high level of user immersion, rendered views should be subjectively perceived as having good quality. Unfortunately, subjective quality evaluation is a tedious and time-consuming process, which is highly inconvenient and expensive in practical systems. Therefore, a natural alternative to performing exhaustive subjective tests is objectively assessing the quality.

Significant challenges in immersive video are the rendering-induced distortions that are unique to this medium. Traditional video quality metrics may not adequately capture these distortions, as they are tailored to conventional two-dimensional video content. This necessitates the development of a specialized objective quality metric that can address the specific issues associated with immersive video, such as the slight shift of reprojected pixels and variations in color characteristics between input views. Development of such a specialized metric is crucial for ensuring that the virtual scenes appear natural and are free from visual discomfort, thereby maintaining the overall immersive experience.

2. Related Work

There are numerous objective quality metrics described in the literature. Many of these methods can properly mimic the behavior of the human visual system in typical image and video processing applications, e.g., several PSNR-based metrics: IV-PSNR (immersive video peak signal-to-noise ratio) [4], CS-PSNR (color-sensitivity-based combined PSNR) [5], PSNR-HA (modified PSNR, which considers the human visual system properties and the change in contrast and mean value), or PSNR-HMA (PSNR-HA, which additionally considers the visual masking of discrete cosine transform coefficients) [6]; metrics based on trained models such as VMAF (video multi-method assessment fusion) [7], VIF (visual information fidelity) [8], LPIPS (learned perceptual image patch similarity) [9], or SFF (sparse feature fidelity) [10]; metrics based on an idea of measuring the structural similarity of images: SSIM (structural similarity index) [11], MS-SSIM (multi-scale SSIM) [12]; the ones based on image feature similarity, e.g., FSIM (feature similarity index) [13], and metrics utilizing temporal information of video sequences [14–16].

All the abovementioned methods were designed for assessing the quality of typical, two-dimensional video; therefore, they are not optimal for video systems that allow the user to look around (360 video, 3DoF—3 degrees of freedom—systems [1]) or navigate within an acquired scene (immersive video systems). Visual quality assessment of 360 videos is widely studied [17], and many efficient quality metrics were proposed. For instance, in [18], the authors have proposed the OV-PSNR metric, integrating both spatial and temporal aspects of distortions and focusing on temporal eye fixation analysis. In [19] and [20], the visual quality assessment is based on neural networks and is focused on the influence of projection of the 360 video into the view watched by a viewer [19] or the influence of viewing directions [20]. Some research focuses on adapting existing methods (e.g., VMAF) to 360 video [21].

However, 360 video systems do not provide a full immersion of the user into the scene, as they do not provide motion parallax, which is inconsistent with human visual perception and leads to visual discomfort [22]. Immersive video introduces the possibility of changing the position of the viewer, which is enabled by view rendering [3]. Most state-of-the-art metrics cannot accurately assess immersive video quality with its typical, rendering-induced distortions [4], such as a slight shift of reprojected pixels and different color characteristics of input views (cf. Section 3).

Most of the existing methods designed for DIBR are computationally complex (e.g., require SIFT—scale-invariant feature transform—for detection of disoccluded areas [23] or perform an exhaustive search for similar regions in compared images [24]), designed for scenarios much simpler than immersive video, e.g., stereoscopic perspective [25–29] or omnidirectional video [30–33]. The approaches to stereoscopic and omnidirectional video quality assessment vary significantly across the different studies. For instance, in [25], the authors introduce a novel blind stereoscopic video quality metric, which incorporates measurement of the disparity entropy for assessing the binocular vision quality. In [26], the authors focus on the efficiency of the proposed metric in terms of processing time, developing a real-time metric designed for stereoscopic video streaming scenarios. Other approaches can require the analysis of 3D saliency maps, both sparse [27] or dense [29], and the metrics designed for stereoscopic perspective video are often based on machine learning techniques [28,29]. The saliency map analysis, as well as neural networks, can also be used for omnidirectional video quality assessment [31]. Other state-of-the-art approaches are focused on the analysis of the behavior used in the omnidirectional scene [32,33].

The state-of-the-art quality metrics designed for stereoscopic perspective or omnidirectional video are often very efficient and competitive in their application. However, as stated above, such applications are much simpler than the immersive video, where a user can freely navigate within the 3D scene but changing his or her viewing angle, as well as viewing position.

Such a characteristic of the immersive video system introduces artifacts and distortions which do not occur in simpler applications, and which are induced by view rendering. On the other hand, the quality metrics designed for immersive video often require the use of depth maps [34], being less versatile (e.g., they cannot be used for the assessment of virtual views presented to the final user of the immersive video system). A more practical approach is based on adapting more straightforward metrics to characteristics of immersive video [4,35,36].

The objective quality metric, which provides the highest correlation with MOS (mean opinion score) in immersive video applications, is IV-PSNR, proposed by the authors of this paper in [4]. The IV-PSNR metric was designed to properly handle and assess distortions typical to immersive video. As written in [4], IV-PSNR provides reliable results despite not utilizing any temporal information and being based on PSNR, which is simple and widely used but certainly not the best objective quality metric.

3. Challenges in Immersive Video Quality Assessment

In this paper, we propose to use structural similarity [11] as a foundation and combine it with the ideas constituting the advantages of IV-PSNR [4]. The usage of a much better base metric (structural similarity instead of MSE—mean squared error [4]) allowed us to increase the correlation with MOS (when compared to IV-PSNR) and resulted in the formulation of IV-SSIM—the structural similarity metric designed for immersive video.

Similarly to IV-PSNR, the proposed IV-SSIM metric does not analyze the temporal consistency of the video, measuring each frame of video independently. However, previous research on the quality of synthesized (rendered) video [37] demonstrated that temporal distortions in immersive video are primarily introduced by changing the viewport rather than by temporal inconsistencies. Therefore, for immersive video applications, it is much more important to properly assess the rendering-induced distortions than the temporal consistency of the content itself.

The most significant source of distortion in immersive video is the process of virtual view rendering (synthesis). The rendering introduces two distinct types of error: corresponding pixel shift error and global color offset error.

3.1. Corresponding Pixel Shift Error

The corresponding pixel shift error (CPSE) is caused by the quantization of depth maps (usually represented with 8 or 10 bits per value), imperfect camera parameters, and inaccuracies in the virtual view rendering process. All mentioned causes are unavoidable during the rendering and induce minor changes (shifts) in the position of rendered objects. This type of error is unnoticeable for viewers; however, it downgrades the "quality" measured by most typical objective quality metrics and makes the results irrelevant.

An example of a rendering-introduced object shift is shown in Figure 2. The figures show an original image (Figure 2A) and a synthesized image (Figure 2B) with the dancer being slightly shifted (by less than 2 pixels). This change in object position is irrelevant to the experience of a viewer; however, it significantly degrades the SSIM metric value (0.9954→0.9893). The proposed IV-SSIM metric is immune to CPSE artifacts (the IV-SSIM metric value is the same for both examples), exposing a better correlation with subjective quality perception. The example shown in Figure 2C presents the object shift exceeding the range expected from reprojection error. In this case, both metrics indicate quality degradation.

(**A**)	SSIM_Y:	0.9847	(**B**)	SSIM_Y:	0.9790	(**C**)	SSIM_Y:	0.9691
	IV-SSIM:	0.9875		IV-SSIM:	0.9875		IV-SSIM:	0.9753

Figure 2. Noticeability of pixel shift in a rendered view: (**A**) correct position of the dancer, (**B**) dancer shifted by two pixels to the right, (**C**) dancer shifted significantly. Sequence Choreo [38].

3.2. Global Color Offset Error

In order to obtain a high-quality rendered view, in practical immersive video systems, it has to be rendered using at least one real view. Real views are often captured by physical cameras, which sometimes differ in capture parameters (exposure setting, color balance), especially when cameras force automatic mode. These sources lead to slight, global differences between the real and the rendered view, called global color offset error (GCOE). Similarly to CPSE, this type of distortion is imperceptible for the viewer while significantly changing metric value.

An example of the abovementioned errors is presented in Figure 3. The image presented in Figure 3B is a result of virtual view synthesis from two real cameras. Those cameras significantly differ in color characteristics (when compared to each other and the reference camera—Figure 3A). It leads to distractive artifacts clearly visible in the background of the exemplary scene. The image in Figure 3C contains a result of virtual view synthesis with color correction applied during the rendering process. The addition of the color correction eliminates background color inconsistency. However, SSIM metric values (0.9216 for B and 0.9107 for C) do not match the subjective perception of image C, which is less distorted than image B. When the IV-SSIM metric is considered, the image quality of example C is evaluated as higher than example B. Such a metric behavior is consistent with subjective perception. Please note that the definition of IV-SSIM (cf. Section 4) implies that its value is usually significantly higher than for SSIM, thus IV-SSIM higher than 0.99 does not mean that the image has near-to-perfect quality (as it would be the case for SSIM). However, for objective quality metrics, the crucial aspect is the correlation with subjective quality perception.

(**A**)	Reference input view	(**B**)	SSIM_Y:	0.9216	(**C**)	SSIM_Y:	0.9107
			IV-SSIM:	0.9900		IV-SSIM:	0.9905

Figure 3. Fragment of a virtual view rendered using views captured by cameras with different color characteristics (**B**) and color-corrected views (**C**), vs. reference input view (**A**); brightness increased by 40% to emphasize differences. Sequence Carpark [39].

4. IV-SSIM

The proposed IV-SSIM metric is based on the structural similarity index (SSIM, [11]), which is one of the most widely used metrics in research on image and video processing. The SSIM metric measures the similarity of two images by using three properties: intensity L, contrast C, and structure S [11]. These properties are defined by analyzing local means (μ_c^I and μ_c^J), standard deviations (σ_c^I and σ_c^J), and covariance ($\sigma_c^{I,J}$) within each color component c of images I and J. The IV-SSIM metric preserves a general idea of the SSIM metric; however, the calculation of several properties and statistics (intensity, local means, standard deviations, and covariance) was significantly changed. These changes adapt the IV-SSIM metric to distortion types occurring in immersive video (described in Section 3).

4.1. Local Image Statistics

In the proposed IV-SSIM metric, the local mean and standard deviation of image I are calculated in the same way as for the classic SSIM

$$\mu_c^I(x,y) = \sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} [\omega(i-x, j-y) \cdot I_c(i,j)], \tag{1}$$

$$\sigma_c^I(x,y) = \sqrt{\sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} \left[\omega(i-x, j-y) \cdot (I_c(i,j))^2\right] - (\mu_c^I(x,y))^2}, \tag{2}$$

where μ_c^I is the local mean and σ_c^I is the standard deviation of the image I, and ω is the weighting mask of size $(2k+1) \times (2k+1)$. Both $\mu_c^I(x,y)$ and $\sigma_c^I(x,y)$ correspond to pixel coordinates (x,y) of the component c of the image I. $I_c(i,j)$ is the single picture element of component c of the image I at coordinates (i,j). The typical weighting mask ω has the size 11×11 ($k = 5$) and is calculated using the circular-symmetric Gaussian weighting function [11].

The IV-SSIM metric includes modifications dedicated to the characteristics of immersive video, implying changes in the calculation of the local statistics of image J and the covariance between images I and J:

$$\mu_c^J(x,y) = \sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} \left[\omega(i-x, j-y) \cdot J_c(i',j')\right], \tag{3}$$

$$\sigma_c^J(x,y) = \sqrt{\sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} \left[\omega(i-x, j-y) \cdot (J_c(i',j'))^2\right] - \left(\mu_c^J(x,y)\right)^2}, \tag{4}$$

$$\sigma_c^{I \to J}(x,y) = \sum_{i=x-k}^{x+k} \sum_{j=y-k}^{y+k} \left[\omega(i-x, j-y) \cdot I_c(i,j) \cdot J_c(i',j')\right] - \mu_c^I(x,y) \cdot \mu_c^J(x,y), \tag{5}$$

where μ_c^J is the local mean and σ_c^J is the standard deviation of the image J, and $\sigma_c^{I \to J}$ is the covariance between images I and J.

The (i',j') is a position of the pixel within image J, defined as:

$$i' = i + s_x(i,j),\ j' = j + s_y(i,j), \tag{6}$$

where $s_x(\cdot), s_y(\cdot)$ are the corresponding pixel shift (in the horizontal and vertical direction, respectively)—cf. Section 3.1. The corresponding pixel shift is calculated independently for each pixel (x,y) of image I.

The shift $s_x(i,j), s_y(i,j)$ is calculated by analyzing the $B \times B$ neighborhood of the pixel (i,j) within image J in search of the pixel, which is most similar to the pixel (i,j) in image I:

$$s_x(x,y), s_y(x,y) \in [-B,B] \cap \mathbb{Z} \ni \left|I_C(x,y) - J_C\left(x + s_x(x,y), y + s_y(x,y)\right)\right| = \min_{\substack{w \in [-B,B] \\ h \in [-B,\ B]}} |I_C(x,y) - J_C(x+w, y+h)|, \tag{7}$$

where B is the maximum considered shift of the corresponding pixel, by default set to 2 (resulting in a 5×5 search range). The 5×5 search range was chosen to preserve consistency with the IV-PSNR metric [4].

It should be emphasized that the modifications of the covariance calculation made it asymmetrical, thus:

$$\sigma_c^{I \to J}(x,y) \neq \sigma_c^{J \to I}(x,y). \tag{8}$$

The problem of IV-SSIM asymmetricity, together with the proposed solution, is presented in Section 4.4.

4.2. Image Properties Similarity

In the second step of IV-SSIM calculation, all the abovementioned local statistics are used for the estimation of the similarity of three properties of images I and J: intensity L, contrast C, and structure S.

$$L_c^{I \to J}(x,y) = \frac{2 \cdot \mu_c^I(x,y) \cdot \left[\mu_c^J(x,y) + s_c^{I \to J}\right] + C_1}{\mu_c^I(x,y)^2 + \left[\mu_c^J(x,y) + s_c^{I \to J}\right]^2 + C_1}, \tag{9}$$

$$C_c^{I \to J}(x,y) = \frac{2 \cdot \sigma_c^I(x,y) \cdot \sigma_c^J(x,y) + C_2}{\sigma_c^I(x,y)^2 + \sigma_c^J(x,y)^2 + C_2}, \tag{10}$$

$$S_c^{I \to J}(x,y) = \frac{\sigma_c^{I \to J}(x,y) + C_3}{\sigma_c^I(x,y) \cdot \sigma_c^J(x,y) + C_3}, \tag{11}$$

where $s_c^{I \to J}$ is the global offset (cf. Section 3.2) of component c between images I and J, defined as:

$$s_c^{I \to J} = \frac{1}{W_c \cdot H_c} \sum_{y=0}^{H_c - 1} \sum_{x=0}^{W_c - 1} (I_c(x,y) - J_c(x,y)), \tag{12}$$

and C_1, C_2, and C_3 are constants providing numerical stability:

$$C_1 = \left(K_1 \cdot (2^b - 1)\right)^2, C_2 = \left(K_2 \cdot (2^b - 1)\right)^2, C_3 = \frac{C_2}{2}, \tag{13}$$

where b is the bit-depth of the image, and:

$$K_1 = 0.01, K_2 = 0.03. \tag{14}$$

W_c and H_c denote the width and height of the image (i.e., width and height of its c-th color plane). The comprehensive description regarding the calculation and usage of C_1, C_2, C_3, K_1, and K_2 coefficients can be found in [40].

In the third step, all property similarities are combined, resulting in local quality scores for component c:

$$Q_c^{I \to J}(x,y) = \left[L_c^{I \to J}(x,y)\right]^{\alpha} \cdot \left[C_c^{I \to J}(x,y)\right]^{\beta} \cdot \left[S_c^{I \to J}(x,y)\right]^{\gamma}, \tag{15}$$

where α, β, and γ are parameters describing the relative importance of $L_c^{I \to J}$, $C_c^{I \to J}$, and $S_c^{I \to J}$, respectively. As proposed in [11], the values $\alpha = \beta = \gamma = 1$ are used.

4.3. Calculation of the IV-SSIM Value

Local quality scores $Q_c^{I\rightarrow J}(x,y)$ could be used for comparing the quality of different parts of the image, being potentially useful for visual localization of distortions. However, these scores have to be combined to assess the quality of the entire image. The IV-SSIM value for each color component c is calculated by averaging these scores over the entire image (entire color plane of component c):

$$IV-SSIM_c^{I\rightarrow J} = \frac{1}{W_c \cdot H_c} \sum_{y=0}^{H_c-1} \sum_{x=0}^{W_c-1} Q_c^{I\rightarrow J}(x,y). \tag{16}$$

In order to produce a single IV-SSIM value, values calculated for single color components are combined using the weighted average:

$$IV-SSIM_{YUV}^{I\rightarrow J} = \frac{IV-SSIM_Y^{I\rightarrow J} \cdot w_Y + IV-SSIM_U^{I\rightarrow J} \cdot w_U + IV-SSIM_V^{I\rightarrow J} \cdot w_v}{w_Y + w_U + w_V}, \tag{17}$$

where w_Y, w_U, and w_V denote weights for luma and both chroma components. By default, $w_Y = 4$, $w_U = 1$, and $w_V = 1$. Such values are in line with default weights for IV-PSNR [4] and ISO/IEC MPEG VC common test conditions for MPEG immersive video (MIV CTC) [41]; they provide a high correlation with subjective quality assessment and refer to the proportion of luma and chroma samples in the most commonly used chroma subsampling format—4:2:0.

4.4. Symmetricity of IV-SSIM

A robust and practical objective quality metric should be symmetrical; thus, it should produce the same output independently of the order of input images. In order to provide the symmetricity, the final value of the IV-SSIM between images I and J is calculated as:

$$IV-SSIM(I,J) = \min\left(IV-SSIM_{YUV}^{I\rightarrow J},\ IV-SSIM_{YUV}^{J\rightarrow I}\right). \tag{18}$$

4.5. IV-SSIM for Video Sequences

The calculation of the IV-SSIM value (18) is performed independently for each frame of a video sequence. In order to assess the quality of the video sequence, IV-SSIM values calculated for all the frames are averaged.

4.6. Publicly Available IV-SSIM Implementation

The reference implementation of the IV-SSIM algorithm is available under the 3-clause BSD license within the Quality Metrics for Immersive Video (QMIV) software v1.0.

The QMIV software was implemented by the authors of this paper and is publicly available in the public git repository of the MPEG Video Coding experts group: https://gitlab.com/mpeg-i-visual/qmiv (accessed on 11 August 2024).

5. Overview of Experiments

The proposed IV-SSIM metric was compared with 16 state-of-the-art quality metrics with publicly available implementations: $PSNR_Y$ (PSNR of luma component), $PSNR_{YUV}$ (weighted average of PSNR for three components with luma weight six times higher than the weights for both chroma components, as described in [42]), IV-PSNR [4], CS-PSNR [5], PSNR-HA and PSNR-HMA [6], VMAF [7], a pixel-based VIF [8], LPIPS [9], SFF [10], $SSIM_Y$ [11] and its multiscale version, MS-SSIM [12], FSIM [13], MW-PSNR (morphological wavelet PSNR) [35], and MP-PSNR (morphological pyramid PSNR) [36]. Moreover, we have added an $SSIM_{YUV}$ metric, calculated for all three color components and weighted in the same way as $PSNR_{YUV}$.

The effectiveness of the proposed IV-SSIM was evaluated in two experiments. In the first one (Section 6), the influence of different techniques of immersive video coding was assessed. In the second experiment (Section 7), IV-SSIM was compared against the commonly used image quality assessment database—TID2013 [43]—in order to present its performance in non-immersive video applications.

In each experiment, the metrics were compared using two commonly used correlation coefficients: PLCC (Pearson's linear correlation coefficient) and SROCC (Spearman's rank-order correlation coefficient) [44].

The calculation of $PSNR_{Y}$, $PSNR_{YUV}$, and IV-PSNR metrics was performed using the IV-PSNR software v4.0 [45] developed by the authors of the IV-PSNR metric [4]. The VMAF metric was calculated using the implementation provided by its authors [46]. Among all tested metrics, only these two software packages include a parallel variant of metric calculation, therefore, in Section 8 we provided results for both single- and multi-threaded computational time evaluation. All metrics were calculated on the same computer equipped with an AMD Ryzen 9 3900XT processor (with 12 identical cores) in connection with 64 GiB of DDR4/3200MT memory. The processor was operating in the default dynamic frequency setting; therefore, the core frequency for single-threaded loads was significantly higher than for multi-threaded variants.

6. Effectiveness in Immersive Video Coding

In the first experiment, IV-SSIM was evaluated using the results of the "MPEG Call for Proposals on 3DoF+ Visual" [47], which contain videos compressed using seven different immersive video coding techniques, five test sequences (both natural and computer-generated), and four rate points (6.5, 10, 15, and 25 Mbit/s). The subjective quality evaluation was performed using the Double Stimulus Impairment Scale test method [48] on a group of 18 naïve viewers [49]. The participants assessed the quality of the "pose trace" videos, simulating the virtual navigation of the viewer within a scene [50]. The objective quality was evaluated by assessing the quality of synthesized input views (virtual views rendered in the position of real cameras that captured the scene). A detailed methodology for subjective and objective evaluation of the immersive video coding dataset can be found in the publicly available ISO/IEC MPEG document [47].

The dataset was used to evaluate a correlation between MOS (mean opinion score) and 17 objective quality metrics. The results are presented in Figure 4. As shown, IV-SSIM achieved the highest SROCC and PLCC among all 17 tested metrics. It outperformed the second-best metric—IV-PSNR—by 0.065 and 0.068 for SROCC and PLCC, respectively. When compared to other metrics, the difference is even higher, showing the superiority of IV-SSIM in the assessment of immersive video.

The difference between correlation coefficients measured for IV-SSIM and $SSIM_{YUV}$ (three-component SSIM, not adapted to immersive video) is 0.205 for SROCC and 0.218 for PLCC (cf. Table 1). A similar difference was measured for IV-PSNR and $PSNR_{YUV}$ (0.201 and 0.211). Such a result shows that the handling of typical immersive video distortions increases the correlation between objective and subjective quality in a consistent way, independently of the objective metric, which was used as a basis (PSNR or SSIM).

Table 1. Comparison between the proposed IV-SSIM metric, typical SSIM [11], IV-PSNR [4], and PSNR. Three-component versions of PSNR and SSIM (i.e., $PSNR_{YUV}$ and $SSIM_{YUV}$) were used.

Scenario	Correlation Coefficient	PSNR	IV-PSNR [4]	SSIM [11]	IV-SSIM
immersive video coding (Section 6)	SROCC	0.527	0.728	0.588	0.793
	PLCC	0.520	0.731	0.581	0.799
general applications—TID2013 (Section 7)	SROCC	0.818	0.825	0.839	0.850
	PLCC	0.816	0.820	0.823	0.825

Despite the fact that the proposed IV-SSIM metric does not assess the temporal consistency of video, it significantly outperforms other metrics, including the ones that measure the temporal similarity (e.g., VMAF). The reason for such a behavior is, as mentioned in Section 3, an atypical characteristic of immersive video, where temporal distortions are introduced not by the timeline, but by changing the viewpoint.

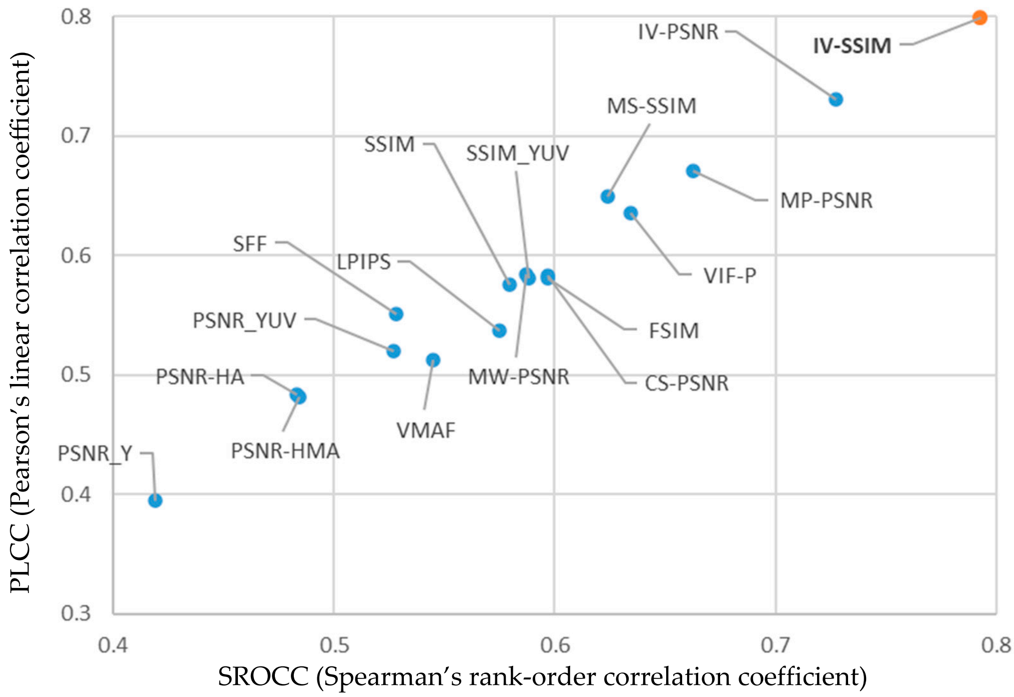

Figure 4. SROCC and PLCC values for all considered metrics; immersive video coding scenario.

7. Effectiveness in General Applications (TID2013)

In the second experiment, all tested metrics were compared against the TID2013 database [43], which includes 24 miscellaneous types of distortions, such as several types of noise, blurring, compression, change in brightness and contrast, as well as compression and transmission errors. It is important to note that the TID2013 database does not contain images with reprojection-related types of distortion (described in Section 3).

The correlation (calculated as SROCC) between subjective and objective quality was estimated independently for each type of distortion. Figure 5 presents SROCC and PLCC values averaged over all 24 types of distortions.

As presented in Figure 5, the proposed IV-SSIM metric outperforms most tested metrics also for non-immersive video applications. For SROCC, IV-SSIM is the second-best metric, and its performance is nearly the same as that of VMAF. VMAF is a meta-metric containing several sub-metrics combined with weights trained for natural images; therefore, its high performance for natural images from the TID2013 database is not surprising. Considering PLCC, it is only worse than VMAF, PSNR-HA, and PSNR-HMA (however, the difference is negligible). Such a discrepancy shows that the correlation between IV-SSIM and MOS is less linear than for these three metrics. However, correlation linearity is less crucial than the ability to assess which video or image is better (what is measured by the rank-order SROCC metric).

The discrepancy between the two used correlation coefficients exists (and it is consistent for all SSIM metrics: IV-SSIM, $SSIM_{YUV}$, and $SSIM_{Y}$), but its impact can be considered negligible, as both coefficients indicate that the proposed IV-SSIM metric performs well even though it was designed for handling a different type of data. Therefore, it could be stated that IV-SSIM properly mimics the subjective perception of quality not only for immersive video but also for any video-related application.

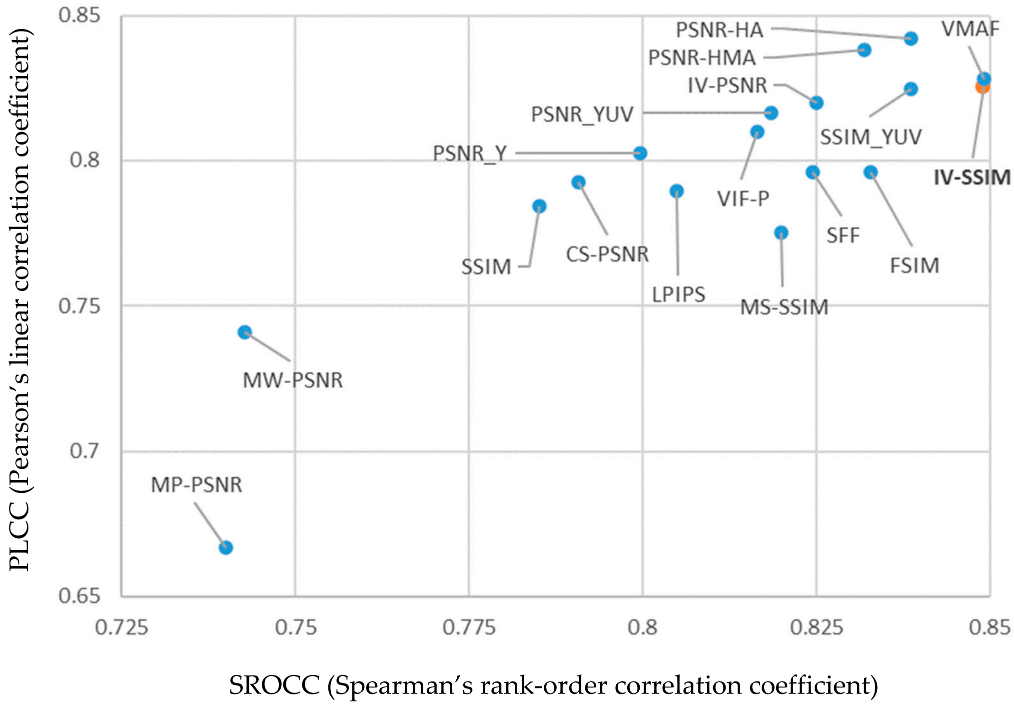

Figure 5. SROCC and PLCC values for all considered metrics; TID2013 database.

8. Computational Time Estimation

A practical objective quality metric should not only correlate with subjective quality perception but should also assess the quality as fast as possible. Unfortunately, the SSIM metric, as defined in [11], introduces high computational and memory bandwidth complexity. Assuming the default size of the Gaussian weighting mask (ω) equal to 11×11, the calculation of SSIM requires 242 memory accesses and ~1000 multiplications for each component of a pixel. To mitigate the mentioned complexity, the proposed IV-SSIM metric was designed to be parallelizable by spreading the metric calculation across multiple CPU cores.

The computational performance was evaluated using our experimental, parallel implementation of $SSIM_Y$, $SSIM_{YUV}$, and IV-SSIM. This experimental implementation is not fully optimized and leaves possibilities for prospective performance improvements. We expect that fully optimized and vectorized implementation would be noticeably faster, but production-quality software implementation is out of the scope of this paper.

Figure 6 compares IV-SSIM with 16 evaluated state-of-the-art metrics in terms of the computational time required for assessing the quality of a single frame of a 4K $\times$ 4K video. These results are presented together with SROCC values estimated in the immersive video experiment (cf. Section 6).

As presented in Figure 6, quality assessment of a single frame of a high-resolution sequence requires only 1.3 s (in a multi-thread implementation), which is faster than most tested implementations of state-of-the-art metrics. Simultaneously, IV-SSIM provides the highest correlation with MOS (mean opinion score).

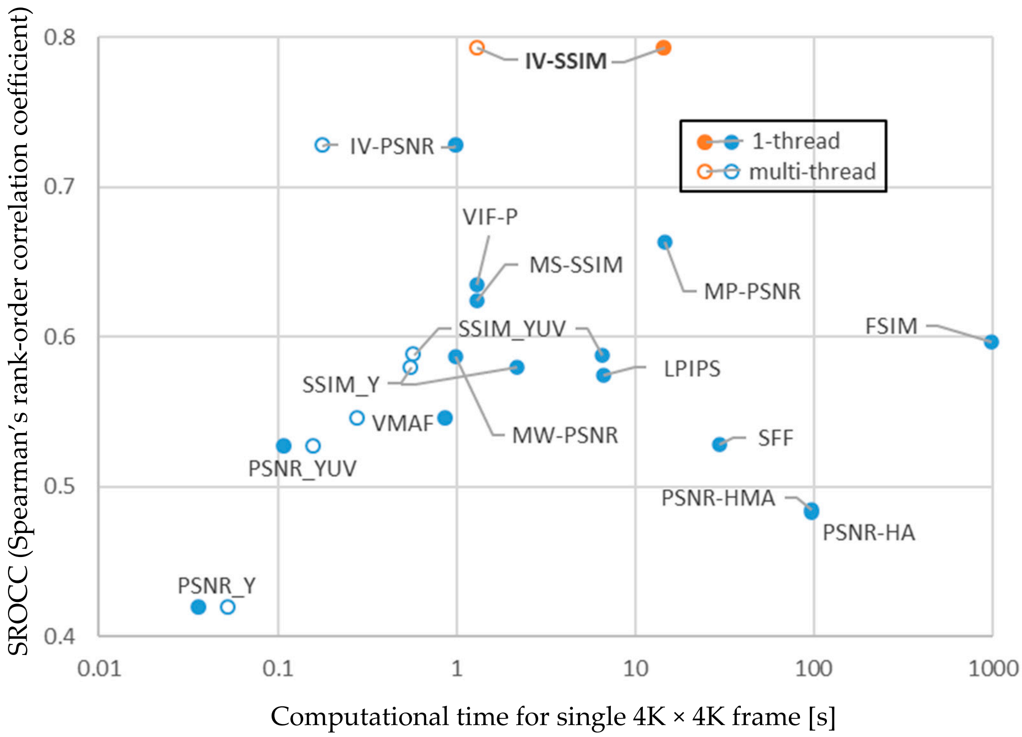

Figure 6. SROCC and computational time for all considered metrics; immersive video coding scenario. Calculations were performed on AMD Ryzen 9 3900XT (12 cores).

9. Conclusions

The proposed IV-SSIM metric is a successor and potential replacement of our previous work—IV-PSNR [4]—which is a PSNR-based metric adapted for the characteristics of immersive video. IV-SSIM is a full-reference objective quality metric based on a calculation of structural similarity (SSIM) between two images.

IV-SSIM combines the advantages of SSIM and IV-PSNR, properly mimicking a human visual system in assessing the quality of immersive video, with its typical distortions caused by reprojection of pixels between multiple views: slight shift of reprojected pixels and color artifacts originated from various color characteristics of different views.

The IV-SSIM metric was compared to 16 state-of-the-art metrics in two experiments. In the first one, its efficiency in an immersive video coding scenario was evaluated. In the second experiment, IV-SSIM was evaluated against the TID2013 database [43], showing its performance in a typical non-immersive video application. The experiments have shown that IV-SSIM clearly outperforms state-of-the-art metrics in assessing the quality of immersive video, also being highly competitive for the assessment of various kinds of distortions introduced to typical 2D images.

Moreover, the calculation of IV-SSIM is straightforward and parallelizable, making the proposed metric practical and convenient to use in real, practical immersive video systems.

Its efficiency was appreciated by the experts of the MPEG Video Coding group, resulting in including IV-SSIM in the common test conditions for MPEG immersive video—MIV CTC [51] and implicit neural visual representation—INVR CTC [52]. The implementation of the IV-SSIM metric was provided by the authors of this paper within the Quality Metrics for Immersive Video (QMIV) software v1.0 [53] and is publicly available on the MPEG Video Coding public git repository: https://gitlab.com/mpeg-i-visual/qmiv (accessed on 11 August 2024).

Currently, IV-SSIM is based on the default definition of the structural similarity index (SSIM) presented in [11]. However, such a definition (e.g., using the Gaussian weighting mask (ω) equal to 11×11) is not efficient in terms of computational complexity. On the other hand, several more efficient implementations of SSIM are known and commonly used, e.g., the rectangular mask used in the ffmpeg framework [54]. Therefore, in the future, different implementations of the structural similarity in the IV-SSIM metric should be analyzed.

Moreover, similarly to SSIM, the proposed IV-SSIM lacks temporal video stability analysis. In the vast majority of scenarios, rendering-induced distortions in immersive video dominate over temporal distortions. However, in some cases (e.g., significant scene geometry changes over a short period of time, resulting in dramatically unstable depth maps), the metric may be insensitive to video flickering, which decreases the subjective quality of video. Therefore, the second direction for future work is to increase the efficiency of IV-SSIM by implementing temporal consistency analysis.

Author Contributions: Conceptualization, A.D. and J.S.; methodology, A.D.; software, J.S.; validation, A.D. and W.N.; writing, A.D. and J.S. All authors have read and agreed to the published version of the manuscript.

Funding: The research was supported by the Ministry of Science and Higher Education of the Republic of Poland.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The experiment described in Section 7 was conducted using the publicly available dataset TID2013 [43], https://www.ponomarenko.info/tid2013.htm, accessed on 11 August 2024. The implementation of the proposed IV-SSIM metric is available within the QMIV v1.0 framework, https://gitlab.com/mpeg-i-visual/qmiv, accessed on 11 August 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wien, M.; Boyce, J.; Stockhammer, T.; Peng, W.H. Standardization status of immersive video coding. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2019**, *9*, 5–17. [CrossRef]
2. Dziembowski, A.; Mieloch, D.; Stankiewicz, O.; Domański, M.; Lee, G.; Seo, J. Virtual view synthesis for 3DoF+ video. In Proceedings of the Picture Coding Symposium (PCS), Ningbo, China, 12–15 November 2019. [CrossRef]
3. Müller, K.; Merkle, P.; Wiegand, T. 3-D video representation using depth maps. *Proc. IEEE* **2011**, *99*, 643–656. [CrossRef]
4. Dziembowski, A.; Mieloch, D.; Stankowski, J.; Grzelka, A. IV-PSNR—The objective quality metric for immersive video applications. *IEEE Trans. Circuits Syst. Video Technol.* **2022**, *32*, 7575–7591. [CrossRef]
5. Shang, X.; Liang, J.; Wang, G.; Zhao, H.; Wu, C.; Lin, C. Color-sensitivity-based combined PSNR for objective video quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 1239–1250. [CrossRef]
6. Ponomarenko, N.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Carli, M. Modified image visual quality metrics for contrast change and mean shift accounting. In Proceedings of the 11th International Conference the Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), Polyana, Ukraine, 23–25 February 2011.
7. Li, Z.; Aaron, A.; Katsavounidis, I.; Moorthy, A.; Manohara, M. Toward a Practical Perceptual Video Quality Metric. In *Netflix Technology Blog*; Technical Report; 2016. Available online: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652 (accessed on 11 August 2024).
8. Sheikh, H.R.; Bovik, A.C. Image information and visual quality. *IEEE Trans. Image Process.* **2006**, *15*, 430–444. [CrossRef]
9. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 586–595. [CrossRef]
10. Chang, H.W.; Yang, H.; Gan, Y.; Wang, M.H. Sparse feature fidelity for perceptual image quality assessment. *IEEE Trans. Image Process.* **2013**, *22*, 4007–4018. [CrossRef] [PubMed]
11. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error measurement to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [CrossRef] [PubMed]
12. Wang, Z.; Simoncelli, E.P.; Bovik, A.C. Multiscale structural similarity for image quality assessment. In Proceedings of the 37th Asilomar Conference on Signals, Systems & Computers, Pacific Grove, CA, USA, 9–12 November 2003. [CrossRef]
13. Zhang, L.; Zhang, L.; Mou, X.; Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [CrossRef]
14. Zeng, K.; Wang, Z. 3D-SSIM for video quality assessment. In Proceedings of the 19th IEEE International Conference on Image Processing, Orlando, FL, USA, 30 September–3 October 2012. [CrossRef]

15. Mantiuk, R.; Denes, G.; Chapiro, A.; Kaplanyan, A.; Rufo, G.; Bachy, R.; Lian, T.; Patney, A. FovVideoVDP: A visible difference predictor for wide field-of-view video. *ACM Trans. Graph.* **2021**, *40*, 49. [CrossRef]
16. Wang, Y.; Jiang, T.; Ma, S.; Gao, W. Spatio-temporal SSIM index for video quality assessment. In Proceedings of the 2012 Visual Communications and Image Processing, San Diego, CA, USA, 27–30 November 2012. [CrossRef]
17. Zakharchenko, V.; Choi, K.P.; Park, J.H. Quality metric for spherical panoramic video. *Proc. SPIE* **2016**, *9970*, 57–65. [CrossRef]
18. Gao, P.; Zhang, P.; Smolic, A. Quality assessment for omnidirectional video: A spatio-temporal distortion modeling approach. *IEEE Trans. Multimed.* **2020**, *24*, 1–16. [CrossRef]
19. Guo, J.; Huang, L.; Chien, W.C. Multi-viewport based 3D convolutional neural network for 360-degree video quality assessment. *Multimed. Tools Appl.* **2022**, *81*, 16813–16831. [CrossRef]
20. Guo, J.; Luo, Y. No-reference omnidirectional video quality assessment based on generative adversarial networks. *Multimed. Tools Appl.* **2021**, *80*, 27531–27552. [CrossRef]
21. Orduna, M.; Díaz, C.; Muñoz, L.; Pérez, P.; Benito, I.; García, N. Video Multimethod Assessment Fusion (VMAF) on 360VR contents. *IEEE Trans. Consum. Electron.* **2020**, *66*, 22–31. [CrossRef]
22. Vadakital, V.K.M.; Dziembowski, A.; Lafruit, G.; Thudor, F.; Lee, G.; Alface, P.R. The MPEG immersive video standard—current status and future outlook. *IEEE Multimed.* **2022**, *29*, 101–111. [CrossRef]
23. Li, L.; Zhou, Y.; Gu, K.; Lin, W.; Wang, S. Quality assessment of DIBR-synthesized images by measuring local geometric distortions and global sharpness. *IEEE Trans. Multimed.* **2018**, *20*, 914–926. [CrossRef]
24. Battisti, F.; Bosc, E.; Carli, M.; Le Callet, P.; Perugia, S. Objective image quality assessment of 3D synthesized views. *Signal Process. Image Commun.* **2015**, *30*, 78–88. [CrossRef]
25. Chen, Z.; Zhou, W.; Li, W. Blind stereoscopic video quality assessment: From depth perception to overall experience. *IEEE Trans. Image Process.* **2018**, *27*, 721–734. [CrossRef] [PubMed]
26. Han, Y.; Yuan, Z.; Muntean, G.M. An Innovative No-Reference Metric for Real-Time 3D Stereoscopic Video Quality Assessment. *IEEE Trans. Broadcast.* **2016**, *62*, 654–663. [CrossRef]
27. Yang, J.; Ji, C.; Jiang, B.; Lu, W.; Meng, Q. No Reference Quality Assessment of Stereo Video Based on Saliency and Sparsity. *IEEE Trans. Broadcast.* **2018**, *64*, 341–353. [CrossRef]
28. Imani, H.; Islam, M.B.; Junayed, M.S.; Aydin, T.; Arica, N. Stereoscopic video quality measurement with fine-tuning 3D ResNets. *Multimed. Tools Appl.* **2022**, *81*, 42849–42869. [CrossRef]
29. Li, C.; Yun, L.; Xu, S. Blind stereoscopic image quality assessment using 3D saliency selected binocular perception and 3D convolutional neural network. *Multimed. Tools Appl.* **2022**, *81*, 18437–18455. [CrossRef]
30. Hu, Z.; Liu, L.; Sang, Q. Omnidirectional Video Quality Assessment with Causal Intervention. *IEEE Trans. Broadcast.* **2024**, *70*, 238–250. [CrossRef]
31. Zhou, M.; Chen, L.; Wei, X.; Liao, X.; Mao, Q.; Wang, H.; Pu, H.; Luo, J.; Xiang, T.; Fang, B. Perception-Oriented U-Shaped Transformer Network for 360-Degree No-Reference Image Quality Assessment. *IEEE Trans. Broadcast.* **2023**, *69*, 396–405. [CrossRef]
32. Jiang, H.; Jiang, G.; Yu, M.; Luo, T.; Xu, H. Multi-angle projection based blind omnidirectional image quality assessment. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 4211–4223. [CrossRef]
33. Sui, X.; Ma, K.; Yao, Y.; Fang, Y. Perceptual quality assessment of omnidirectional images as moving camera videos. *IEEE Trans. Vis. Comput. Graph.* **2022**, *28*, 3022–3034. [CrossRef] [PubMed]
34. Li, L.; Huang, Y.; Wu, J.; Gu, K.; Fang, Y. Predicting the quality of view synthesis with color-depth image fusion. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *31*, 2509–2521. [CrossRef]
35. Sandić-Stanković, D.; Kukolj, D.; Le Callet, P. DIBR-synthesized image quality assessment based on morphological multi-scale approach. *EURASIP J. Image Video Process.* **2016**, *2017*, 4. [CrossRef]
36. Sandić-Stanković, D.; Kukolj, D.; Le Callet, P. DIBR synthesized image quality assessment based on morphological pyramids. In Proceedings of the 2015 3DTV-Conference: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON), Lisbon, Portugal, 26–29 May 2015. [CrossRef]
37. Tian, S.; Zhang, L.; Zou, W.; Su, T.; Morin, L.; Deforges, O. Quality assessment of DIBR-synthesized views: An overview. *Neurocomputing* **2021**, *423*, 158–178. [CrossRef]
38. Klóska, D.; Mieloch, D.; Dziembowski, A.; Szydełko, B.; Stankowski, J.; Lee, G. A new natural content proposal: Choreo. In *Document ISO/IEC JTC1/SC29/WG04 MPEG VC, M66990*; International Organization for Standardization: Rennes, France, 2024.
39. Mieloch, D.; Dziembowski, A.; Domański, M. [MPEG-I Visual] Natural outdoor test sequences. In *Document ISO/IEC JTC1/SC29/WG11 MPEG M51598*; International Organization for Standardization: Brussels, Belgium, 2020.
40. Venkataramanan, A.K.; Wu, C.; Bovik, A.; Katsavounidis, I.; Shahid, Z. A hitchhiker's guide to structural similarity. *IEEE Access* **2021**, *9*, 28872–28896. [CrossRef]
41. ISO/IEC. Common test conditions for MPEG immersive video. In *Document ISO/IEC JTC1/SC29/WG04 MPEG VC N0406*; International Organization for Standardization: Hannover, Germany, 2023.
42. Huang, Y.; Qi, H.; Li, B.; Xu, J. Adaptive weighted distortion optimization for video coding in RGB color space. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014. [CrossRef]
43. Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Process. Image Commun.* **2015**, *30*, 57–77. [CrossRef]

44. Chikkerur, S.; Sundaram, V.; Reisslein, M.; Karam, L. Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison. *IEEE Trans. Broadcast.* **2011**, *57*, 165–182. [CrossRef]
45. Stankowski, J.; Dziembowski, A. IV-PSNR: Software for immersive video objective quality evaluation. *SoftwareX* **2023**, *24*, 101592. [CrossRef]
46. Git Repository. Available online: https://github.com/Netflix/vmaf (accessed on 21 February 2024).
47. ISO/IEC. Call for proposals on 3DoF+ visual. In *Document ISO/IEC JTC1/SC29/WG11 MPEG N18145*; International Organization for Standardization: Marrakech, Morocco, 2019.
48. ITU-T. Subjective video quality assessment methods for multimedia applications. In *Recommendation ITU-T P.910*; International Telecommunication Union: Geneva, Switzerland, 2008.
49. ISO/IEC. Evaluation results of the call for proposals on 3DoF+ visual. In *Document ISO/IEC JTC1/SC29/WG11 MPEG N18353*; International Organization for Standardization: Geneva, Switzerland, 2019.
50. Boyce, J.; Doré, R.; Dziembowski, A.; Fleureau, J.; Jung, J.; Kroon, B.; Salahieh, B.; Vadakital, V.K.M.; Yu, L. MPEG immersive video coding standard. *Proc. IEEE* **2021**, *109*, 1521–1536. [CrossRef]
51. ISO/IEC. Common test conditions for MPEG immersive video. In *Document ISO/IEC JTC1/SC29/WG4 MPEG VC N0539*; International Organization for Standardization: Sapporo, Japan, 2024.
52. ISO/IEC. Common test conditions on radiance field representation and compression. In *Document ISO/IEC JTC1/SC29/WG4 MPEG VC N0561*; International Organization for Standardization: Sapporo, Japan, 2024.
53. ISO/IEC. Software manual of QMIV. In *Document ISO/IEC JTC1/SC29/WG4 MPEG VC N0535*; International Organization for Standardization: Sapporo, Japan, 2024.
54. Ffmpeg Framework. Available online: https://ffmpeg.org (accessed on 7 August 2024).