applied sciences

MDPI

Article

Temporally Aware Objective Quality Metric for Immersive Video

Jakub Stankowski, Bartosz Sojka, Tomasz Grajek * and Adrian Dziembowski

Institute of Multimedia Telecommunications, Poznan University of Technology, 60-965 Poznań, Poland; jakub.stankowski@put.poznan.pl (J.S.); bartosz.sojka@put.poznan.pl (B.S.); adrian.dziembowski@put.poznan.pl (A.D.)
* Correspondence: tomasz.grajek@put.poznan.pl

Abstract

State-of-the-art objective quality metrics designed for immersive content typically prioritize spatial distortions; therefore, they can omit temporal artifacts introduced by view synthesis and dynamic scene rendering. Consequently, metrics such as the commonly used peak signal-to-noise ratio for immersive video (IV-PSNR) are "temporally blind", creating a conceptual gap where temporally stable distortions cannot be distinguished from disruptive temporal flickering. To address this limitation, we propose a temporal extension of the IV-PSNR metric that incorporates motion information into the quality assessment process. The method augments the traditional Y, U, and V color components with a fourth channel representing motion vectors (M), enabling the proposed four-component IV-PSNR$_{\mathrm{YUVM}}$ metric to account for dynamic distortions introduced by view rendering. To evaluate the effectiveness of the proposed approach, multiple configurations of motion integration were tested, including metrics based solely on motion consistency, metrics combining motion with texture, and several dense optical flow algorithms with different parameter settings. Extensive experiments performed on immersive video sequences demonstrate that the proposed four-component IV-PSNR$_{\mathrm{YUVM}}$ achieves the highest correlation with subjectively perceived video quality. These results confirm that combining texture with motion information provides a benefit, making the proposal a valuable addition for real-world immersive video systems.

Keywords: video quality; immersive video; video compression; view rendering

1. Introduction

While immersive video [1,2] is by far more advanced than traditional two-dimensional video, the ultimate goal remains unchanged: delivering the highest possible quality of experience (QoE) by maximizing subjective satisfaction for users. Unfortunately, conducting comprehensive subjective tests for quality assessment is a laborious and time-consuming endeavor [3], rendering it highly impractical. Consequently, a natural alternative is to employ objective quality assessment.

The field of objective quality evaluation is vast, especially within image and video processing. In many scenarios, such as traditional two-dimensional video, cutting-edge metrics effectively emulate the subjective perception of video quality, e.g., structural similarity image assessment (SSIM) [4], video multi-method assessment fusion (VMAF) [5], or learned perceptual image patch similarity (LPIPS) [6].

It is important to note that within the domain of traditional, two-dimensional video, the significance of temporal consistency is well-established. Metrics such as video quality model (VQM) [7] and spatiotemporal reduced reference entropic differencing (STRRED) [8]

Academic Editor: Thomas Lindner

Received: 14 November 2025
Revised: 18 December 2025
Accepted: 23 December 2025
Published: 26 December 2025

efficiently assess temporal changes to detect artifacts (e.g., flickering) which are invisible in static frame analysis. Furthermore, advanced methods explicitly use the optical flow to model the human visual system's (HVS) sensitivity to motion. For instance, motion-based video integrity evaluation (MOVIE) index [9] utilizes optical flow to efficiently evaluate both motion and spatial distortions, while other approaches employ motion-compensated strategies to weight structural similarity (motion-compensated SSIM, MC-SSIM) [10]. Despite these advancements in two-dimensional video, the distinctive nature of immersive video, involving the reprojection of data captured by multiple cameras [11], sets it apart as an exception. In this context, neither traditional spatial nor temporal objective quality metrics do not perform very well. To account for the specific artifacts inherent to immersive video, the PSNR for immersive video (IV-PSNR) metric has been introduced [3].

The IV-PSNR metric [3] is an extension of the classical peak signal-to-noise ratio (PSNR), specifically designed for video quality assessment of immersive content. Unlike basic PSNR, which simply computes pixel-wise differences, IV-PSNR accounts for the characteristics of immersive content by incorporating and compensating corresponding pixel shift (CPS) search (Figure 1) within local blocks, compensating small geometric misalignments introduced by view synthesis. Additionally, IV-PSNR analyzes the global component differences (GCD) between compared frames (Figure 1), providing a more reliable assessment for content with inter-view illumination variations. These techniques allow the IV-PSNR metric to achieve a strong correlation with mean opinion scores (MOS), outperforming even more sophisticated state-of-the-art metrics in immersive video applications.

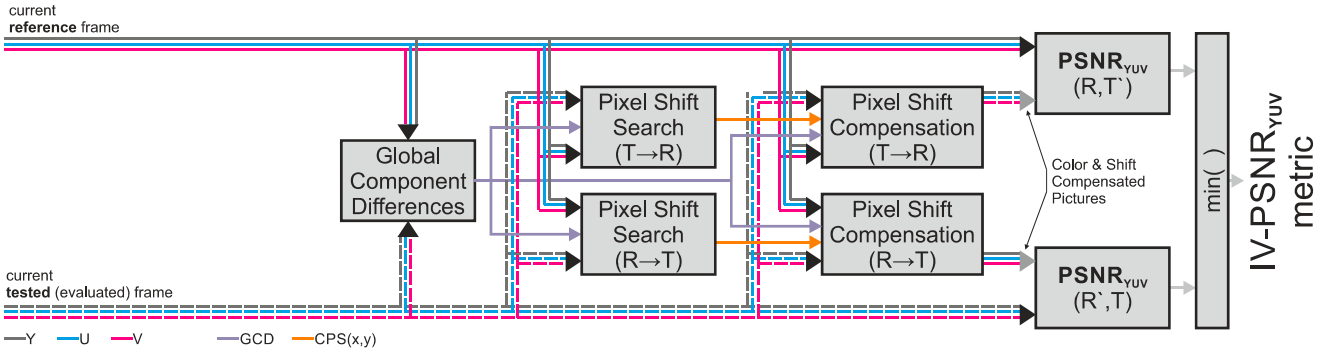

Figure 1. Schematic diagram of the IV-PSNR metric. Arrow color notation: grey—luma component (Y); blue—first chroma component (U); pink—second chroma component (V); violet—global component differences (GCD) vector (one value per Y, U, and V channels); orange—estimated corresponding pixel shift (CPS) between two compared pictures (one value per pixel); light gray—metric values; solid lines represent reference sequence; dashed lines represent tested sequence.

As presented in Figure 2, the basic version of the IV-PSNR metric exhibits a significantly higher correlation with subjective quality evaluation (SROCC > 0.63) than other state-of-the-art metrics in immersive video applications, including perceptual metrics like VMAF (SROCC < 0.41) or LPIPS (SROCC < 0.55). This performance gap is caused by the fact that traditional 2D quality metrics are not designed to assess specific geometric distortions inherent to view synthesis. Consequently, since IV-PSNR is currently the most reliable objective metric for this domain, we select it as the main baseline for our proposed temporal extension, excluding less correlated metrics from further validation.

However, the primary goal of this work is to bridge the conceptual gap in the IV-PSNR metric's definition. The basic version of IV-PSNR [3] is "temporally blind" and practically limits the analysis to a single frame (simply averaging frame-level scores when assessing video sequence). Consequently, it cannot distinguish between a video with stable distortions and one with significant, disruptive temporal flickering if their average spatial quality is the same. This limitation is particularly valid in the considered context,

as in immersive video, viewers freely navigate through a 3D scene, and even subtle inconsistencies over time—such as flickering edges in input views and, what is crucial for subjective quality [12], in depth maps—can strongly degrade perceived quality even if individual frames have high quality.

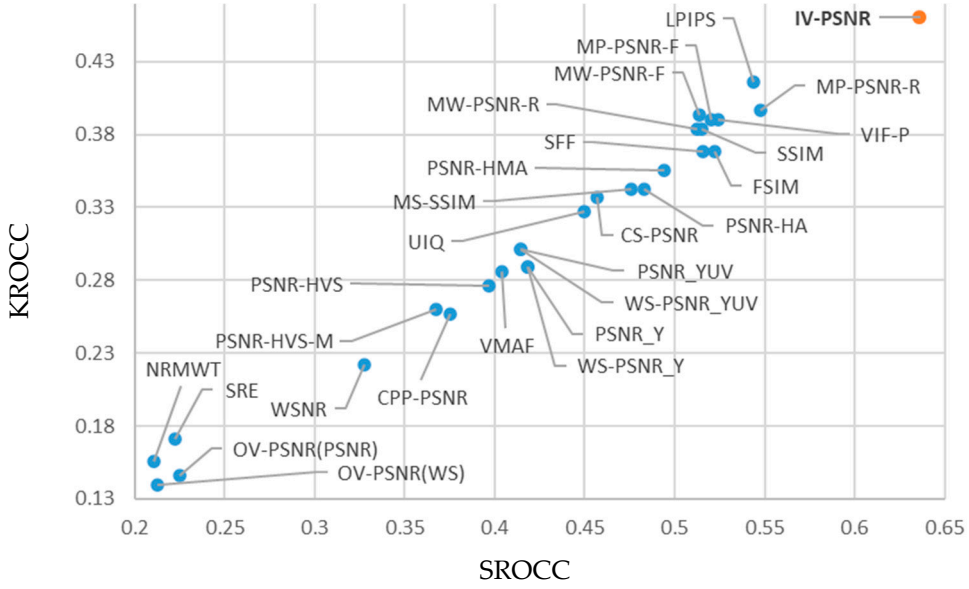

Figure 2. Correlation between subjective quality and state-of-the-art objective quality metrics in immersive video applications. Figure from [3].

Therefore, providing a methodologically sound way to incorporate motion analysis is a necessary evolution of the tool, providing the mathematical framework to capture temporal consistency, which should be a prerequisite for any reliable immersive video quality assessment tool.

2. Temporal Extension of the IV-PSNR Metric

The goal of this work is to extend the IV-PSNR metric with information about changes in the temporal characteristic of the measured sequence. The most widespread quality metrics used in immersive video applications (PSNR, SSIM, IV-PSNR, weighted-to-spherically uniform PSNR (WS-PSNR) [13], LPIPS [6], and SSIM for immersive video (IV-SSIM) [14]) operate on a single image, and sequence quality is calculated by a simple average over frames. Unfortunately, such an approach makes the abovementioned metrics insensitive to the deterioration of motion consistency. Some temporal-related artifacts, like edge flickering of objects, are highly disturbing to viewers and should be included in measured subjective quality.

The proposed approach includes the motion field analysis [15] to evaluate sequence motion consistency. This strategy aligns with state-of-the-art methodologies in traditional video quality assessment (e.g., [9,10]), where incorporating optical flow has been shown to be crucial for accurately modeling the human perception of temporal artifacts. The motion field is determined using a dense optical flow (DOF) algorithm, and the optical flow calculation is performed for both sequences (reference and tested). A visual difference between motion fields calculated for the original and the distorted images is presented in Figure 3.

Building on this, we propose to include the analysis of the similarity between the temporal characteristics of two compared video files in the IV-PSNR metric calculation by augmenting the typical Y, U, and V components with a fourth channel representing motion vectors (M) derived using a dense optical flow algorithm.

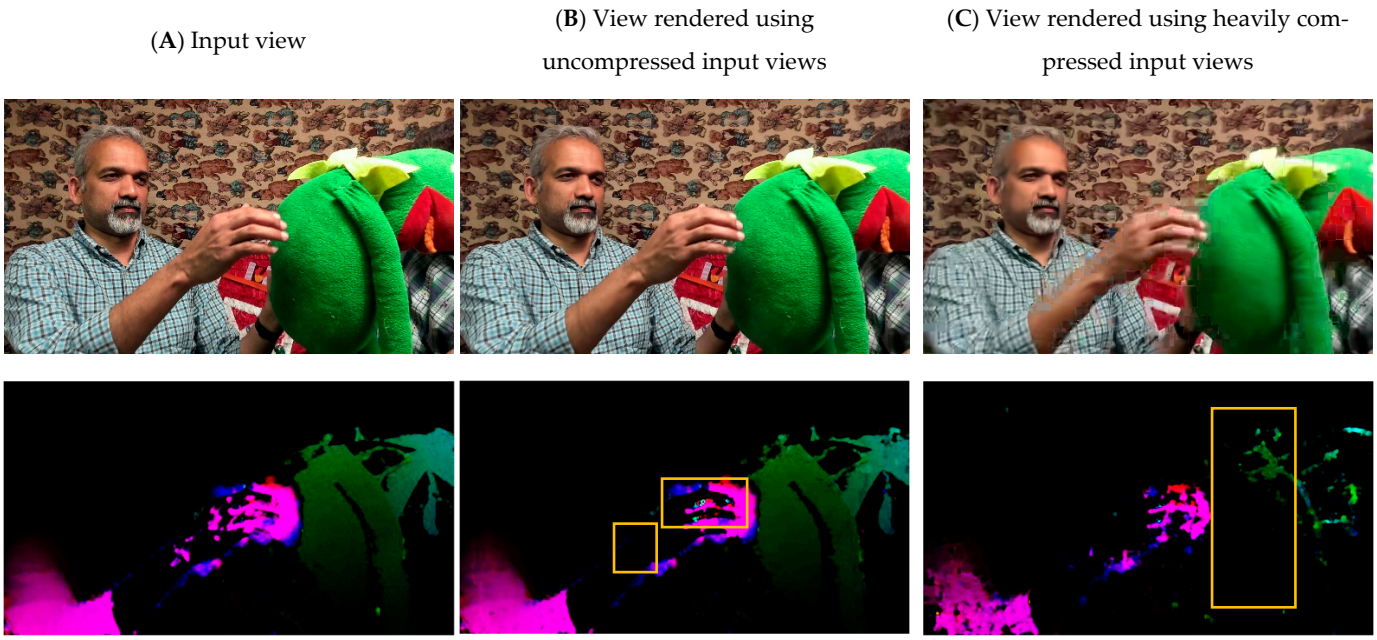

Figure 3. Example of motion fields calculated (using Farneback's algorithm) for original (**A**) and distorted (**B**,**C**) sequences; yellow boxes highlight regions with significantly different characteristics of motion fields calculated for differently processed sequences.

2.1. Considered Dense Optical Flow Estimation Algorithms

The proposed temporal extension of the IV-PSNR metric relies on the accurate and efficient estimation of dense optical flow (DOF) to assess motion consistency within two compared video sequences. While there are numerous DOF algorithms, ranging from classical optimization methods [16,17] to deep learning-based approaches [18,19], in this paper, we focus on practical applicability, accessibility, and the ability to fine-tune performance. For these reasons, we have considered two algorithms available through the commonly used OpenCV library: Farneback's algorithm [20] and the robust local optical flow (RLOF) algorithm [21].

Farneback's algorithm is a well-known global motion estimation technique that approximates the displacement field as a polynomial expansion. Its primary advantages include its robustness to noise and its ability to capture large displacements, making it suitable for a variety of video content, including immersive video. Furthermore, its implementation in OpenCV ensures public availability, facilitating reproducibility of the research. The algorithm's parameters, such as polynomial expansion degree, window size, number of iterations, and number of pyramid levels, offer considerable flexibility for optimization.

The RLOF algorithm is another powerful dense optical flow method. Unlike global approaches, RLOF focuses on local motion estimation, which can be particularly beneficial in scenes with complex or non-uniform motion patterns (e.g., in rendered views in immersive video systems). Its robust characteristics imply insensitivity to outliers and illumination changes, often found in immersive video sequences captured by a set of cameras. Similarly to Farneback's algorithm, the OpenCV implementation of RLOF provides an accessible and well-documented tool. The algorithm's configurable parameters, including regularization weights and iteration counts, enable detailed tuning to meet specific application requirements, thereby ensuring a balance between accuracy and computational cost.

In the context of this study, both algorithms can be integrated into the IV-PSNR v6.0 software and used to compute dense optical flow fields, allowing us to analyze how different DOF estimation algorithms influence the performance of the proposed temporal IV-PSNR extension.

2.2. Dense Optical Flow in IV-PSNR

As discussed earlier, conventional single-image quality metrics, by simply averaging over the entire sequence, are insensitive to crucial temporal artifacts such as object edge flickering, which significantly degrade the subjective quality of immersive video. To address this limitation, the proposed approach extends the IV-PSNR metric by incorporating information derived from motion consistency analysis, which is performed using a dense optical flow algorithm.

Importantly, the temporal IV-PSNR extension does not require the use of a specific DOF algorithm; both Farneback's and RLOF produce the two-dimensional motion vector fields, which can be used as an additional motion channel M, integrated into the metric alongside the Y, U, and V components.

A crucial aspect of integrating DOF into any PSNR-based metric (including IV-PSNR) is to adapt the calculation to work on two-dimensional motion vectors, since they typically operate on scalar image components (e.g., luma and chroma values), calculating the sum of squared differences (SSD) between two images. For a scalar component, the calculation is straightforward:

$$SSD_c = \sum_y \sum_x (R_c(x,y) - T_c(x,y))^2 ,$$

where $R_c(x,y)$ and $T_c(x,y)$ are values of a color component c of pixel (x,y) in the reference (R) and tested (T) sequences, respectively.

To combine spatial distortions with motion-based distortions within a unified IV-PSNR framework, we define the motion distortion term SSD_M by analogy to SSD_c. The SSD_M value is calculated based on two corresponding motion vectors from the reference motion vector field ($\overrightarrow{R_M}$) and the tested motion vector field ($\overrightarrow{T_M}$) calculated for reference (R) and the tested (T) sequences, respectively. Finally, the SSD_M is defined as a sum of the squared Euclidean distances between corresponding motion vectors $\overrightarrow{R_M(x,y)}$ and $\overrightarrow{T_M(x,y)}$ and is described by the following equation:

$$SSD_M = \sum_y \sum_x \left| \overrightarrow{R_M(x,y)} - \overrightarrow{T_M(x,y)} \right|^2 .$$

This formulation ensures that the motion component is fully compatible with the existing SSD-based structure of the IV-PSNR metric and can be integrated as a fourth distortion channel.

The integration of DOF into the IV-PSNR framework can be realized through several methods, resulting in various approaches that leverage temporal characteristics to improve objective quality assessment. Among these, the four-component IV-PSNR (Section 2.2.1) constitutes the main contribution of this paper, jointly evaluating spatial and temporal video aspects. For completeness, additional variants are also examined (Section 2.2.2). These variants are anchor baselines, designed to isolate individual aspects of motion analysis and demonstrate that relying on motion-only or motion-assisted algorithms does not fully capture the whole spatiotemporal characteristics of immersive video.

2.2.1. Proposed Four-Component IV-PSNR: IV-PSNR$_{\mathrm{YUVM}}$

In this approach, the DOF field is computed for both the reference and the tested video sequence. These DOF fields are treated as an additional, fourth component of the image, supplementing three color components (i.e., Y, U, and V)—Figure 4. This allows for the calculation of IV-PSNR for both the spatial and temporal aspects of the video, resulting in the comprehensive IV-PSNR$_{\mathrm{YUVM}}$ (with "YUV" in the name representing color components and "M" representing the motion component) metric.

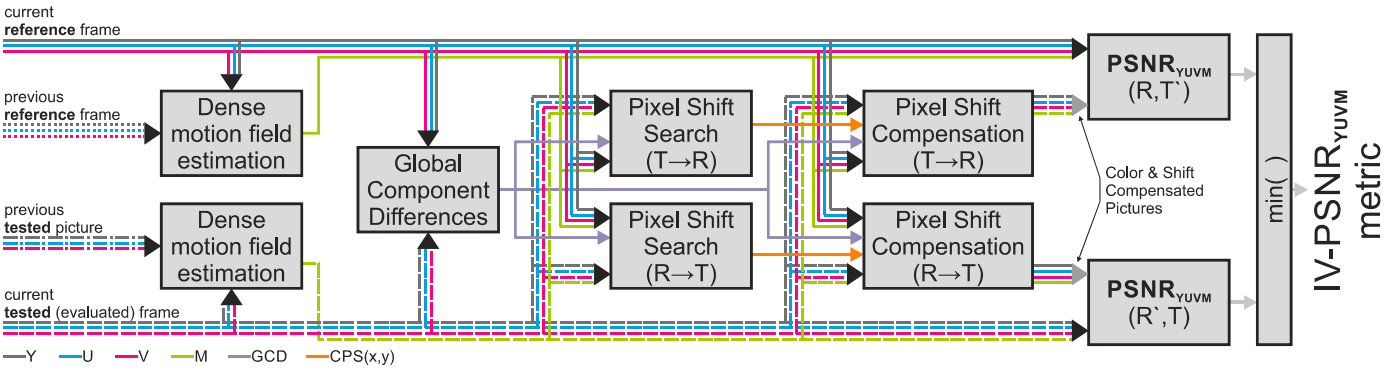

Figure 4. Schematic diagram of the proposed IV-PSNR$_{YUVM}$ metric. Arrow color notation: grey—luma component (Y); blue—first chroma component (U); pink—second chroma component (V); violet—global component differences (GCD) vector (one value per Y, U, and V channels); green—motion field component (M); orange—estimated corresponding pixel shift (CPS) between two compared pictures (one value per pixel); light gray—metric values; solid lines represent current frame reference sequence; dashed lines represent current frame tested sequence; dotted lines represent previous frames of reference and tested sequence.

Since the motion field (M) and the three color components (Y, U, V) operate on different numerical scales, their distortions cannot be combined directly. Luma and both the chromas operate on quantized ranges determined by the bit depth of the input video (typically 8 or 10 bits). At the same time, dense optical flow values are expressed in pixel units and are independent of bit depth. Without normalization, SSD_M could either dominate the combined metric or become negligible, depending on the video bit depth. To ensure that all four components contribute in a comparable and bit-depth-independent manner, the motion-related distortion SSD_M is rescaled to match the dynamic range of the quantized video components. Using 10-bit videos as the reference scale, the corrected distortion is defined as:

$$SSD'_M = SSD_M \cdot 2^{(b-10)} .$$

This normalization does not assume that optical flow vectors scale with bit depth. Instead, it equalizes the numerical dynamic range of all four components, enabling a meaningful and stable weighted combination within the IV-PSNR$_{YUVM}$ metric.

A combination of the IV-PSNR values for each component is performed by using the weighted average, following the concept from the original IV-PSNR metric. In IV-PSNR [3], the weights are set by default to 4:1:1 for luma and two chromas, respectively. In the proposed IV-PSNR$_{YUVM}$, the weights are set to 4:1:1:M, where M is the weight for motion vectors. While a broader range of M values (0.125, 0.25, 0.5, 1, 2, 4, 8, and 16) was initially evaluated, the extreme weights consistently led to lower correlation with subjective quality. Therefore, for clarity and conciseness, only the representative values $M \in \{1, 2, 4\}$ are reported in this paper.

2.2.2. Other DOF Integration Approaches

Two other variants are not alternative proposed metrics. Instead, they serve as anchors designed to isolate the effect of motion analysis in IV-PSNR, helping to demonstrate that the proposed four-component IV-PSNR$_{YUVM}$ metric is the most efficient and conceptually justified approach.

A. DOF-Only IV-PSNR

The DOF-only approach focuses exclusively on the motion consistency. In this method, the IV-PSNR metric is calculated solely on the DOF fields (Figure 5). The comparison is

thus performed directly between the motion sequences of the reference and tested videos, rather than their original texture content. This provides a dedicated measure of motion distortion, making the metric highly sensitive to temporal inconsistencies and artifacts that the texture-based quality assessment might overlook.

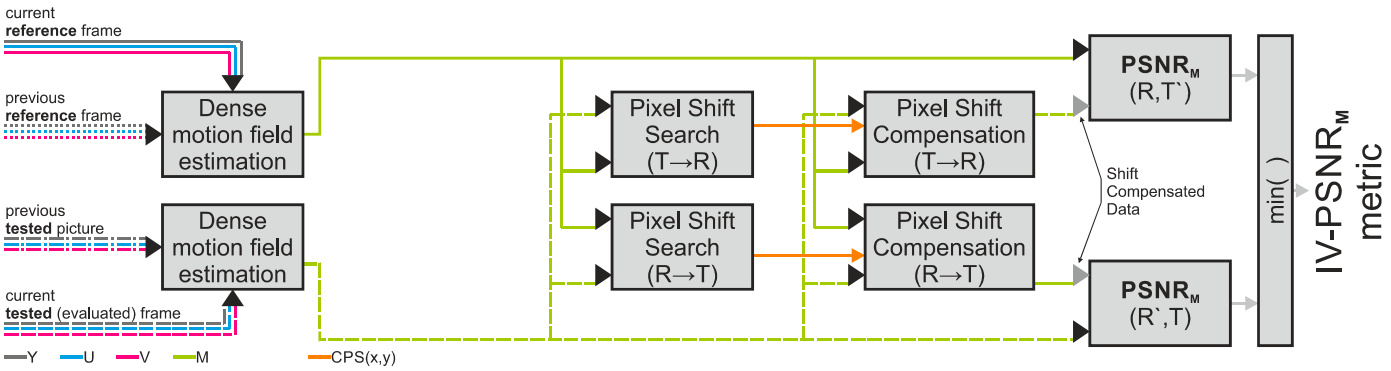

Figure 5. Schematic diagram of the proposed IV-PSNR$_M$ metric. Arrow color notation: grey—luma component (Y); blue—first chroma component (U); pink—second chroma component (V); green—motion field component (M); orange—estimated corresponding pixel shift (CPS) between two compared pictures (one value per pixel); light gray—metric values; solid lines represent current frame reference sequence; dashed lines represent current frame tested sequence; dotted lines represent previous frames of reference and tested sequence.

In total, two versions of the DOF-only IV-PSNR were tested. In the first one, the DOF is calculated for the luma component only. This metric uses a single M channel representing motion information; thus, it can be named IV-PSNR$_M$.

In the second approach, separate DOF is calculated for each color component of the input video sequences. Then, motion field values for luma and two chroma components are combined by using the weighted average (with typical IV-PSNR weights: 4:1:1)—Figure 6. This metric uses three motion-related channels: M_Y, M_U, and M_V, and is named IV-PSNR$_{3M}$. This variant has been implemented in order to test whether basing the calculated DOF on chroma components increases the correlation of the modified IV-PSNR metric with subjective quality.

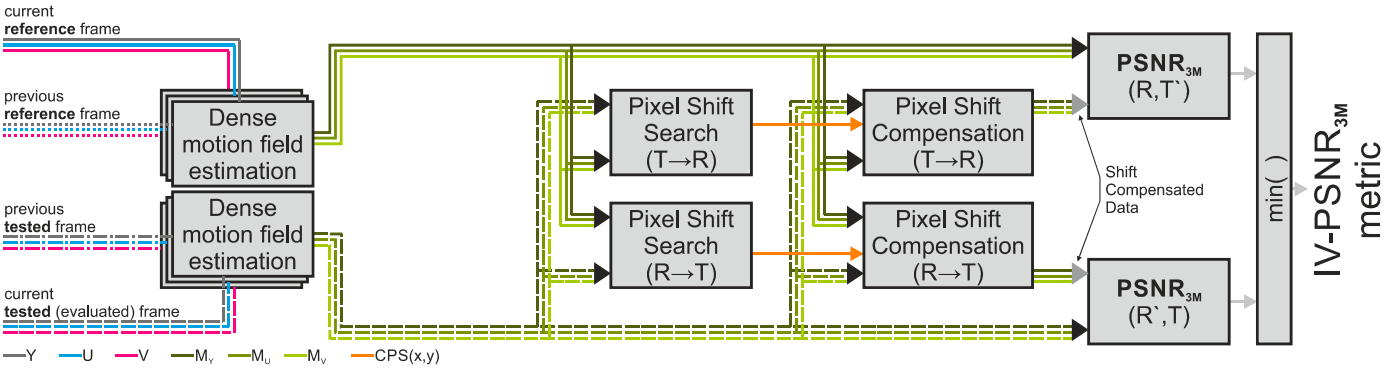

Figure 6. Schematic diagram of the proposed IV-PSNR$_{3M}$ metric. Arrow color notation: grey—luma component (Y); blue—first chroma component (U); pink—second chroma component (V); all shades of green—three motion field components (M_Y, M_U, and M_V); orange—estimated corresponding pixel shift (CPS) between two compared pictures (one value per pixel); light gray—metric values; solid lines represent current frame reference sequence; dashed lines represent current frame tested sequence; dotted lines represent previous frames of reference and tested sequence.

B. MotionCheck (DOF-Assisted Corresponding Pixel Shift Search)

The MotionCheck (MC) approach integrates DOF into the core of the corresponding pixel shift mechanism of the IV-PSNR metric. While the final similarity between two sequences is still computed based on the original input sequences, the process of searching for the "most similar pixel within a colocated block" [3] is guided by the DOF information. Specifically, the search for the best-matching pixel in the neighborhood, a key step in IV-MSE calculation [3], uses the motion field images to identify optimal correspondences. This means that the initial stage of the IV-PSNR calculation leverages DOF to find the proper pixel displacement, while the subsequent quality measurement utilizes the pixel values from the original video sequences.

This method, IV-PSNR$_{MC}$ (IV-PSNR$_{YUV}$ with MotionCheck, MC, Figure 7), aims to make the quality assessment robust to temporal shifts and distortions without direct assessment of the motion itself.

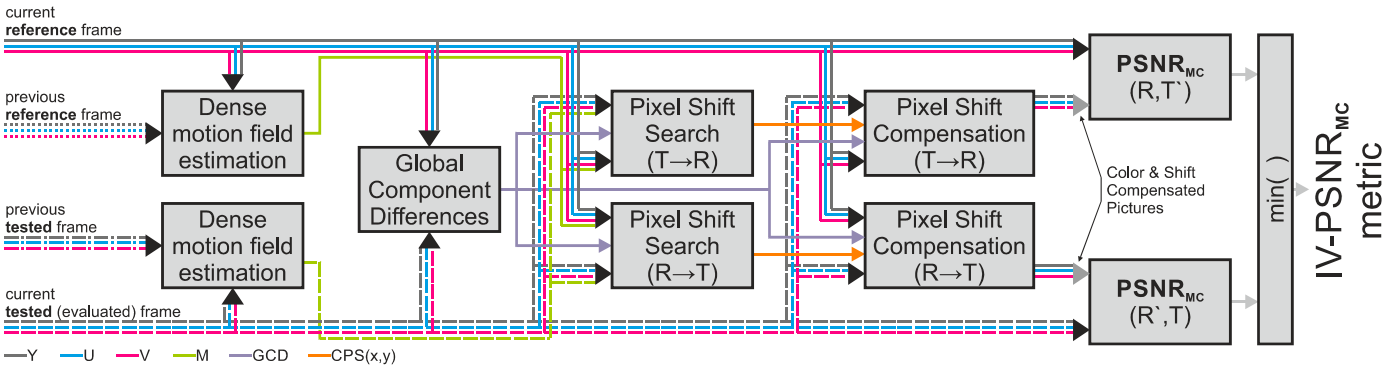

Figure 7. Schematic diagram of the proposed IV-PSNR$_{MC}$ metric. Arrow color notation: grey—luma component (Y); blue—first chroma component (U); pink—second chroma component (V); violet—global component differences (GCD) vector (one value per Y, U, and V channels); green—motion field component (M); orange—estimated corresponding pixel shift (CPS) between two compared pictures (one value per pixel); light gray—metric values; solid lines represent current frame reference sequence; dashed lines represent current frame tested sequence; dotted lines represent previous frames of reference and tested sequence.

3. Experimental Setup

The proposed approach, together with all the anchors, was evaluated using the results of the "MPEG Call for Proposals on 3DoF+ Visual" [22]. Ideally, the proposed metric should be evaluated on multiple diverse datasets to fully assess its robustness and generalization capabilities. However, the availability of suitable databases is a major and well-recognized limitation in immersive video quality assessment. To the best of our knowledge, there is currently no publicly available immersive video quality assessment (VQA) dataset containing MOS scores. Existing, state-of-the-art datasets fall into different categories which only partially relate to immersive video, e.g., omnidirectional video (simple, single-camera video 360 without any 3D reprojection) [23,24], static stereoscopic images [25], or volumetric, object-centric video (focused on isolated, low-resolution point-cloud objects instead of full scene-based immersive video) [26]. While other immersive video datasets exist (e.g., [27] or MIV CTC sequences [28]), they are not VQA datasets as they do not provide subjective quality scores.

Therefore, in this work, we evaluated the proposed temporal IV-PSNR extension on the only available immersive video database. Importantly, this dataset remains representative of the current state-of-the-art, as it contains multiview sequences encoded using seven miscellaneous immersive video coding techniques, including two simulcast scenarios and

five sophisticated algorithms combining the advantages of point cloud processing, different methods of view rendering, edge filtering, depth map refinement, and noise modelling—jointly reflecting various processing and coding artifacts, both spatial and temporal. All sequences were encoded at four different bitrates in the range between 6.5 and 25 Mbps. In total, mean opinion scores (MOS) were calculated for 280 test points [29].

According to the documentation in [29], a total of 49 participants took part in the subjective quality assessment. All viewers were screened for visual acuity and color blindness before the experiment. The evaluation was conducted using the absolute category rating (ACR) methodology, consistent with MPEG's standard testing procedures for video coding. Each participant evaluated the content in four sessions of approximately 12 min with mandatory breaks in between to avoid tiredness.

The stimuli were displayed on a 65-inch monitor with viewing conditions aligned with the ITU recommendations for video quality assessment (including viewing distance, room illumination, and background setup). The test set consisted of five diverse multiview video sequences, including both natural and computer-generated (CG) content, captured using perspective and omnidirectional cameras. The sequences varied in visual complexity, level of detail, motion characteristics, and resolution (ranging from FullHD to 4K). This dataset—with several changes introduced during years of development of the MPEG immersive video coding standard [28,30]—is still widely used in immersive video research and provides a representative variety of artifacts—spatial, temporal, geometric, and rendering-related—making it suitable for evaluating the performance and generalizability of objective quality metrics.

The proposal was compared with unmodified IV-PSNR [3], as it provides the highest correlation with subjective quality in immersive video applications (cf., Figure 1 and [3]). Moreover, we have included results for all the presented anchors in order to demonstrate the versatility of the proposed modification.

In the experiment, metrics were compared using two rank-based correlation coefficients: SROCC and KROCC [31] (Spearman and Kendall rank-order correlation coefficient, respectively—both computed with respect to subjective MOS scores). In total, nine variants of metrics were compared:

- Three state-of-the-art metrics: $PSNR_Y$ (PSNR calculated for luma component), $PSNR_{YUV}$ (weighted average of PSNR for three color components Y, U, and V; weights: 6:1:1—weight of luma six times higher than for both chroma components [3]), IV-$PSNR_{YUV}$ (the basic IV-PSNR, which is by default calculated for three color components with weights 4:1:1 [3]).
- IV-PSNR calculated on motion field sequences instead of textures (cf. the last row of Figure 2): IV-$PSNR_M$ and IV-$PSNR_{3M}$, with motion field calculated for luma (IV-$PSNR_M$), and motion fields calculated for all three color components (IV-$PSNR_{3M}$),
- IV-$PSNR_{MC}$, where the most similar pixel in the neighborhood is selected by comparing motion fields, but the quality itself is calculated based on textures.
- Three versions of four-component IV-PSNR, where motion vectors M are added as a fourth component, alongside color components Y, U, and V: IV-$PSNR_{YUVM}$ (with weights equal to 4:1:1:1, 4:1:1:2, and 4:1:1:4).

The main baseline used for comparison is the original IV-PSNR operating on YUV components (IV-$PSNR_{YUV}$), as it has been widely adopted in immersive video research and demonstrates the highest correlation with subjective quality (c.f., Figure 1 and [3]). Other variants, such as $PSNR_Y$, $PSNR_{YUV}$, and all the anchors defined in Section 2.2.2 are included as additional reference points to illustrate the contribution of individual components and to highlight the benefits of incorporating temporal motion information.

Parameters of the Farneback's algorithm were set as presented in Table 1.

Table 1. Used Farneback's algorithm parameters.

Parameter	Experiment	
	Farneback with Two Levels	**Farneback with Five Levels**
Pyramid scale	0.5	0.5
Levels	2	5
Window size	10	10
Iterations	2	2
Poly N	5	5
Poly sigma	1.2	1.2

For the RLOF algorithm, three experiments were performed. In each, a different interpolation method was used (RIC, EPIC, and GEO). For all other RLOF parameters, the default OpenCV values were used [32].

4. Experimental Results

This section presents the results of the experiments conducted to evaluate the proposed temporally aware quality metric for immersive video. Several optical flow estimation algorithms were tested to determine how motion-related information affects the correlation between the objective and subjective quality.

4.1. Farneback's DOF Algorithm

The first experiment was conducted using the dense optical flow algorithm proposed by Farneback [20]. Two configurations were tested—using two and five pyramid levels—in order to assess the impact of flow smoothness and motion detail on the metric's correlation with subjective quality.

Figure 8 presents the results of the experiments conducted using Farneback's algorithm with two levels of pyramid, as described in the previous chapter. They also show the results for $PSNR_Y$, $PSNR_{YUV}$, and IV-$PSNR_{YUV}$ for easier comparison. As we can see, the IV-$PSNR_{YUVM}$ method showed considerable improvement over original methods, especially the one weighted 4:1:1:4, where it outperformed IV-$PSNR_{YUV}$ by 0.031 and 0.023 for SROCC and KROCC, respectively. This indicates that DOF carries information that helps in making the IV-PSNR algorithm more robust. The sole use of DOF in IV-$PSNR_M$, as well as in IV-$PSNR_{3M}$, showed considerable improvement over the pixel-based $PSNR_{YUV}$ method. Lastly, the IV-$PSNR_{MC}$ technique showed almost no improvement in comparison to IV-$PSNR_{YUV}$.

Changing Farneback's algorithm levels from two to five resulted in a noticeable decrease in correlation (Figure 9). IV-$PSNR_{YUVM}$ has shown results for SROCC and KROCC drop below those of IV-$PSNR_{YUV}$, for example. This applies to every variant of the metric when compared to Farneback's algorithm using only two levels, except IV-$PSNR_{MC}$ and IV-$PSNR_{M,}$ which remained relatively the same (exact results for SROCC and a negligible loss for KROCC).

Overall, Farneback's algorithm with two pyramid levels yielded the highest correlation between the proposed metrics and subjective scores. Increasing the number of levels effected in a loss of fine motion detail, which negatively impacted metric performance. These results suggest that smoothing moderate flow provides the best trade-off between stability and motion detail, while strong smoothing may reduce the sensitivity of the metric to perceptually relevant temporal changes.

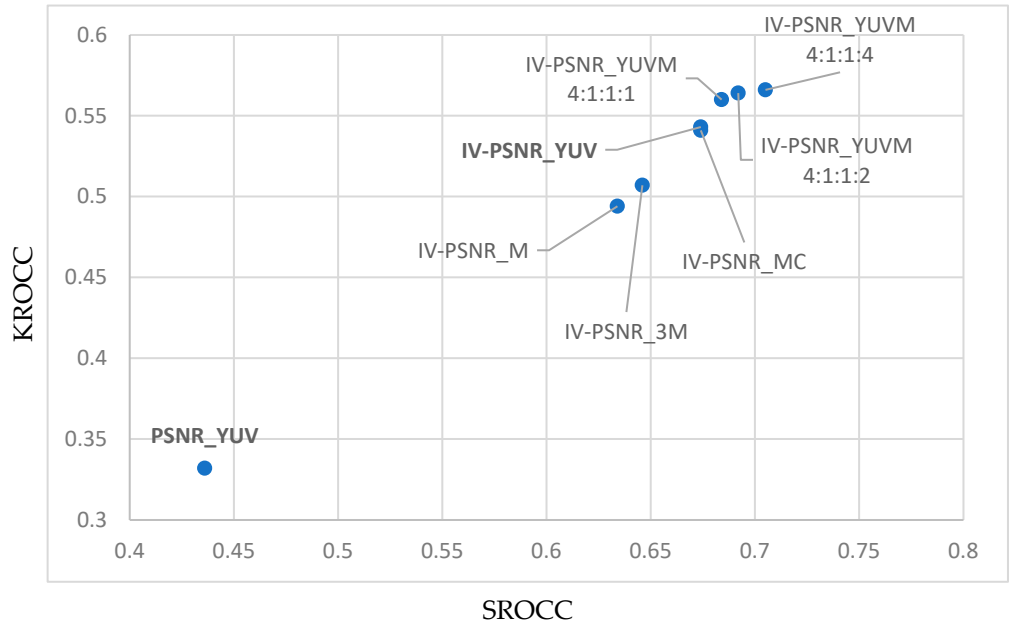

Figure 8. Correlation with subjective quality for the Farneback's algorithm with two levels; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

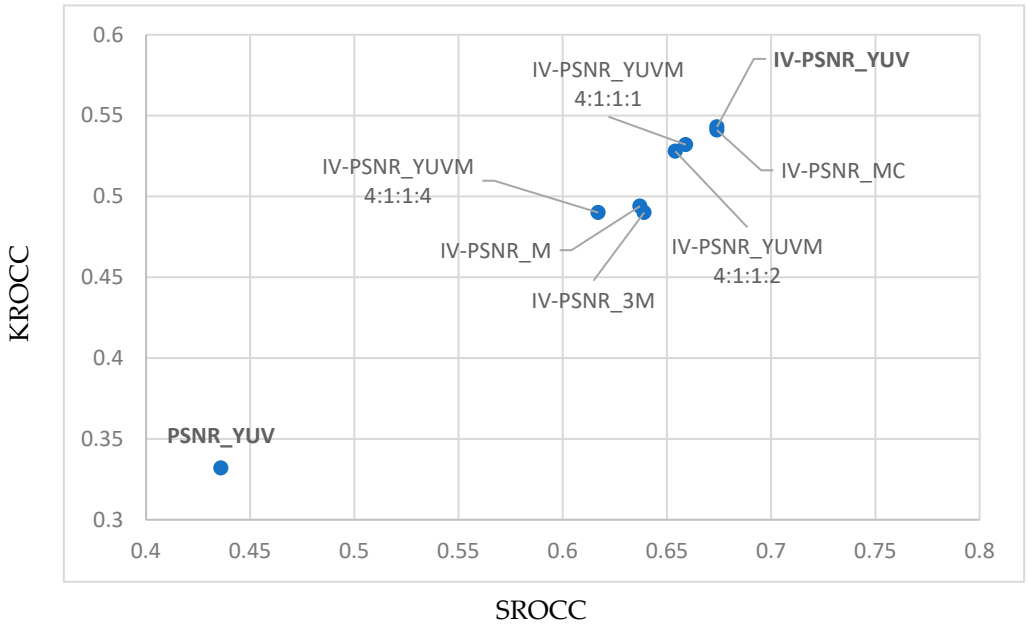

Figure 9. Correlation with subjective quality for the Farneback's algorithm with five levels; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

4.2. RLOF DOF Algorithm

To further analyze the influence of motion estimation quality on the proposed metric, a second experiment using the robust local optical flow (RLOF) algorithm [21] was conducted. In this experiment, three different interpolation methods were tested: RIC, EPIC, and GEO [32]. In general, RLOF offers higher robustness to noise and illumination changes than Farneback's method, making it a valuable reference for testing the stability of the metric across diverse flow characteristics.

The RIC interpolation, despite being the most advanced of the three available interpolation methods, showed the most significant loss of correlation with MOS. As shown in Figure 10, the usage of motion vectors only in IV-PSNR$_M$ provided similar results compared to Farneback's methods. Although both methods yield similar results in an IV-PSNR$_M$ scenario, RLOF with RIC interpolation performs considerably worse in all other cases (i.e., including other image components—IV-PSNR$_{YUVM}$).

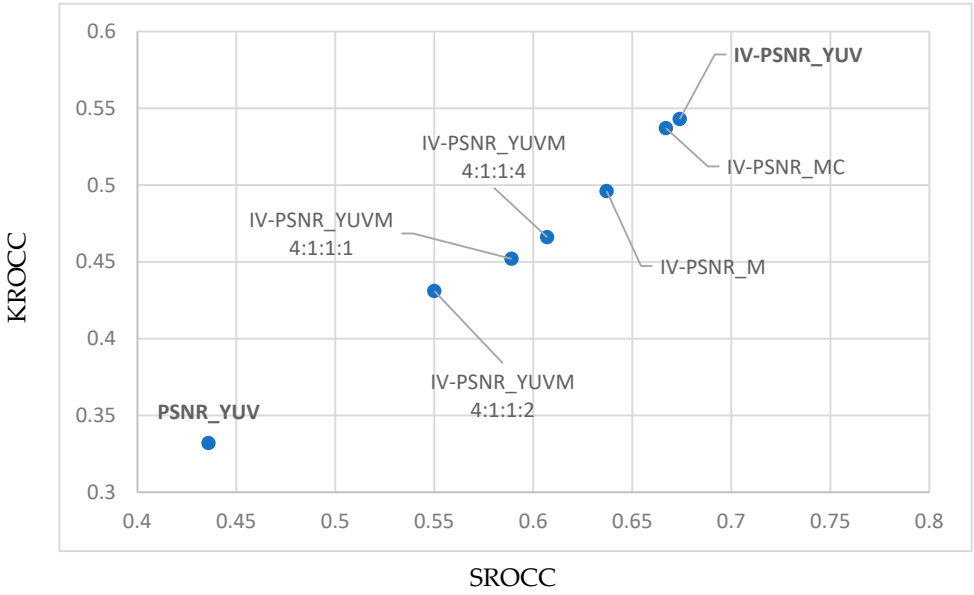

Figure 10. Correlation with subjective quality for the RLOF algorithm with RIC interpolation; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

EPIC interpolation method (Figure 11) offered better results than RIC, despite being less robust and faster. IV-PSNR$_{YUVM}$ with the weights 4:1:1:1 achieved similar results to IV-PSNR$_{YUV}$ (with a negligible gain for SROCC); however, higher weights showed a decline in achieved results for both SROCC and KROCC.

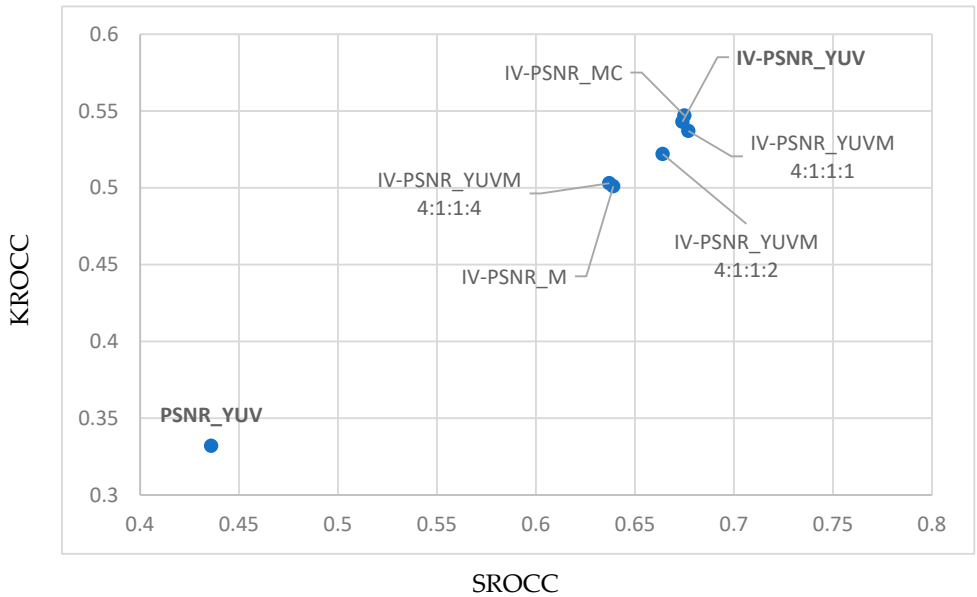

Figure 11. Correlation with subjective quality for the RLOF algorithm with EPIC interpolation; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

The GEO interpolation method was the simplest of the three interpolation methods, offering comparable results for IV-PSNR$_{YUVM}$ as well as IV-PSNR$_{M}$ to the EPIC interpolation method (Figure 12), and providing negligible SROCC gain for the 4:1:1:1 weighting scheme and losses for other scenarios.

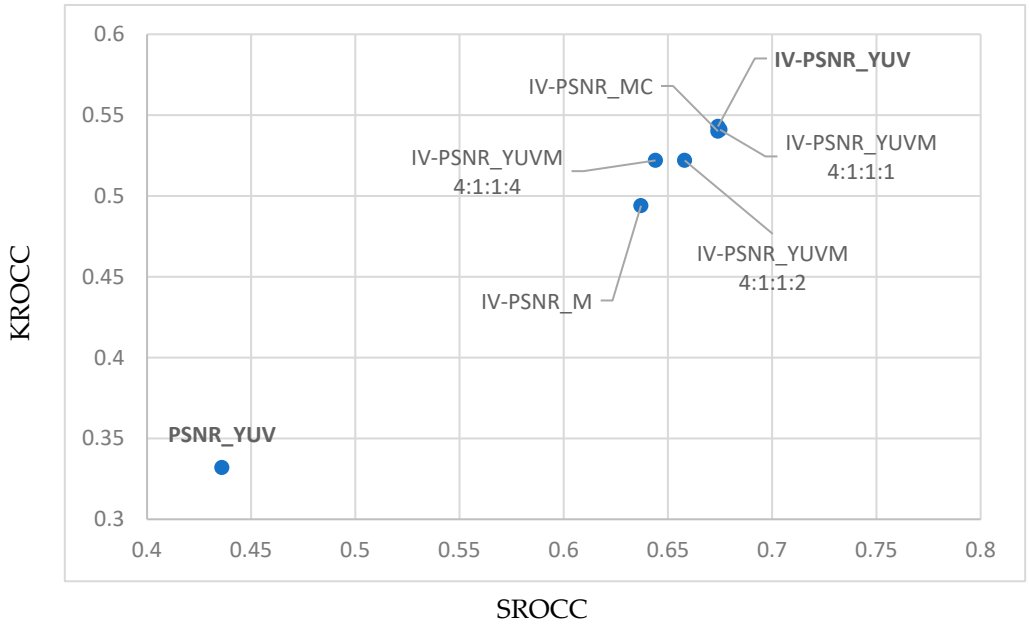

Figure 12. Correlation with subjective quality for the RLOF algorithm with GEO interpolation; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

The results obtained with RLOF indicate that interpolation significantly affects the correlation with subjective quality. While RIC interpolation led to noticeable degradation, the simpler EPIC and GEO methods achieved comparable or even better results. This implies that too complex flow smoothing may obscure motion details essential for proper temporal quality assessment.

4.3. Comparison of Methods

The two previous subsections demonstrate that incorporating temporal information into immersive video quality metrics can considerably improve their correlation with human perception. However, the extent of this improvement strongly depends on the chosen optical flow algorithm and its parameters.

To better understand how different variants perform, Figures 13–15 summarize the overall correlation results for all tested metrics.

For IV-PSNR$_{MC}$ (motion-assisted pixel search, Figure 14), the correlation with MOS is significantly higher, and the IV-PSNR$_{MC}$ calculated using all the DOF algorithms is comparable to IV-PSNR$_{YUV}$. Again, all DOF variants perform similarly—the spread between the best and the worst metric is smaller than 0.01, both for SROCC and KROCC. In other words, changing the pixel search method to a motion-based one does not produce a decisive improvement over the texture-only IV-PSNR$_{YUV}$.

For IV-PSNR$_{M}$ (motion vectors only, Figure 13), all tested DOF methods produce very similar results (the difference between the best and the worst configuration is equal to only 0.005 for SROCC and 0.007 for KROCC). Importantly, IV-PSNR$_{M}$ performs about 0.04 worse than the baseline IV-PSNR$_{YUV}$, indicating that motion-only measurement alone is insufficient to outperform the texture-based quality assessment.

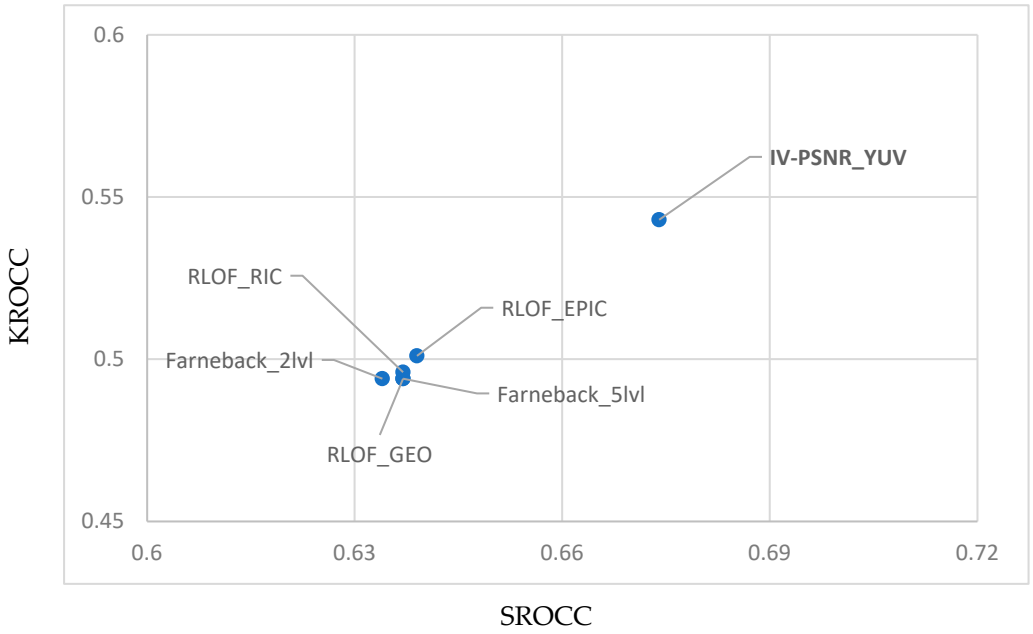

Figure 13. Correlation with subjective quality: comparison of the IV-PSNR$_M$ metric calculated using different DOF algorithms; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

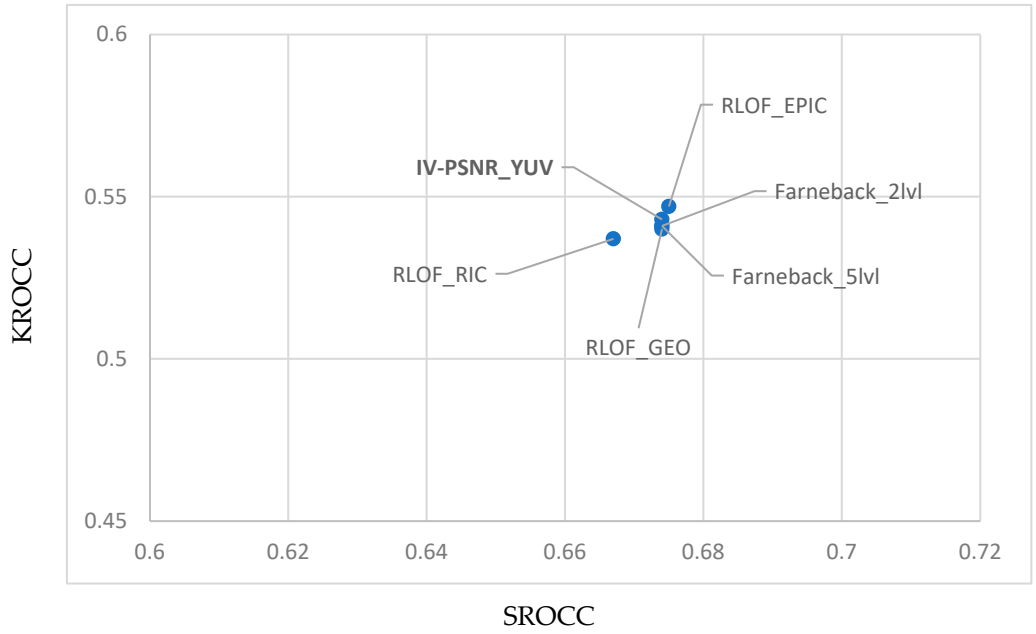

Figure 14. Correlation with subjective quality: comparison of IV-PSNR$_{MC}$ metric calculated using different DOF algorithms; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

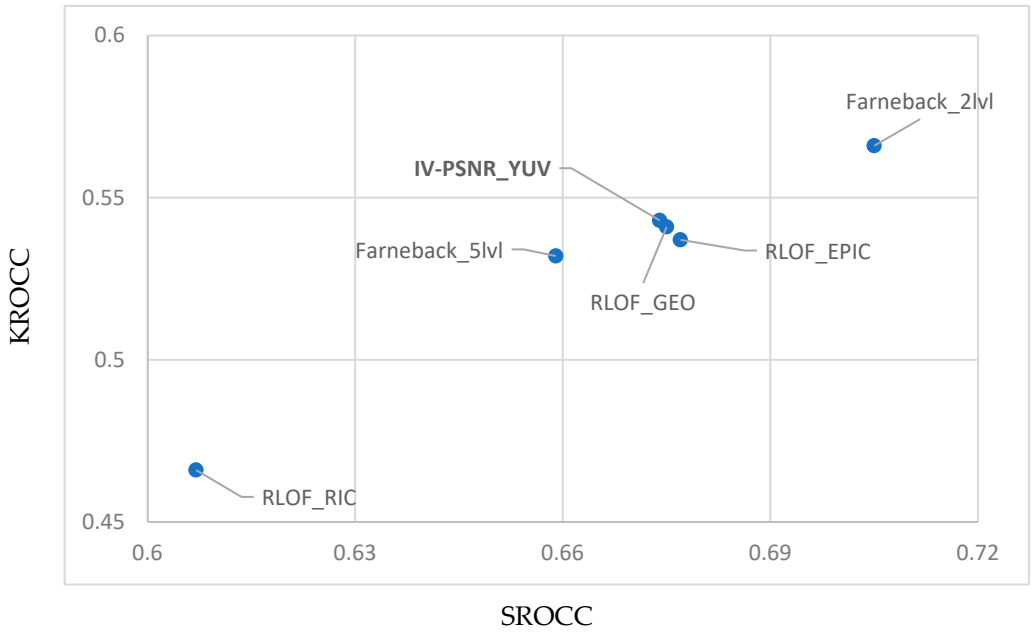

Figure 15. Correlation with subjective quality: comparison of IV-PSNR$_{YUVM}$ metric calculated using different DOF algorithms; each point corresponds to one tested metric; higher KROCC and SROCC indicate stronger monotonic agreement with MOS—and points closer to the upper-right corner represent better-performing metrics.

For the proposed IV-PSNR$_{YUVM}$ metric (Figure 15), the results are completely different, and adding motion vectors as a fourth component substantially changes the outcome depending on the DOF algorithm and its parameters. In particular:

- Farneback with two pyramid levels clearly outperforms IV-PSNR$_{YUV}$ (by 0.03 and 0.02 for SROCC and KROCC, respectively), i.e., this method provides the best correlation with MOS among all tested configurations,
- RLOF with GEO and EPIC interpolations produces results comparable to IV-PSNR$_{YUV}$ (no significant gain or loss),
- RLOF with RIC interpolation and Farneback with five pyramid levels perform significantly worse than IV-PSNR$_{YUV}$.

Taken together, these observations indicate two main conclusions. Firstly, the use of motion-based information alone (IV-PSNR$_{M}$) provides stable results, showing that motion analysis reliably reflects temporal aspects of perceived quality of immersive video.

Secondly, when motion information is combined with texture components, as in the proposed four-component IV-PSNR$_{YUVM}$, the metric outperforms both the baseline state-of-the-art IV-PSNR$_{YUV}$ as well as all other tested variants. While the observed gains over IV-PSNR$_{YUV}$ reflect the diminishing returns typical for high-performing metrics, they remain consistent. What should be highlighted, the degree of improvement depends on the quality of the optical flow estimation, with the two-level Farneback's algorithm yielding the most perceptually consistent results.

4.4. Computational Complexity Consideration

Computation of the dense optical flow is known to be a computationally demanding task. However, when a lightweight algorithm—such as Farneback's method applied in our implementation—the computational complexity becomes reasonably low.

Experimental assessment of computational complexity (evaluated by measuring processing time) was performed by running DOF estimation on a multicore CPU with 16 threads running in parallel. Measurements show that a single frame of FullHD

(1920 × 1080) sequence can be processed in ~382 ms (on average, c.f. Table 2), which demonstrates that the temporal analysis of the video can be incorporated without introducing substantial overhead.

Table 2. Average DOF calculation time for five test sequences from the "MPEG Call for Proposals on 3DoF+ Visual" [22] test set. Times averaged over all test points and all frames. Times measured on a PC machine equipped with Ryzen 9 5950X with 32 GB DDR 3200 MHz RAM.

Sequence Id	SA	SB	SC	SD	SE
Resolution	4096 × 2048	2048 × 2048	4096 × 4096	2048 × 1088	1920 × 1080
DOF calc. time per frame	1616 ms	780 ms	3326 ms	414 ms	382 ms
DOF calc. time per pixel	193 ns	186 ns	198 ns	186 ns	184 ns

As presented, the DOF calculation time linearly scales with the resolution of a sequence, requiring only ~190 ns per pixel.

Importantly, this runtime is small when compared to the typical processing required in immersive video pipelines, where computationally expensive operations such as depth estimation, multiview video coding, and view synthesis often dominate the total computational cost. In such a context, adding a 400 ms step per frame pair has a negligible impact on overall processing time.

When comparing the proposed IV-PSNR$_{YUVM}$ calculation to the baseline IV-PSNR$_{YUV}$, the increase in the computational time is noticeable, but relatively small. When averaging over all sequences, it is 35% slower.

Overall, by combining a computationally efficient DOF method with multithreaded processing, the proposed temporally aware IV-PSNR extension remains practical, fast, and easy to integrate into existing immersive video systems.

5. Conclusions and Future Work

In the paper, we have proposed an extension of the IV-PSNR metric [3], which includes an analysis of compared sequences in the temporal domain by measuring motion consistency, i.e., the similarity of the dense motion field in the reference and the tested sequence. The proposed IV-PSNR$_{YUVM}$ is based on adding motion vectors as a fourth component M, in addition to the three color components of the video: Y, U, and V.

The experimental results demonstrate that the proposed IV-PSNR$_{YUVM}$ increases the correlation between objective and subjective quality in immersive video quality assessment compared to the baseline IV-PSNR metric. Although the absolute gains are modest, reflecting the diminishing returns characteristic of improving an already high-performing baseline, the overall trend confirms that incorporating temporal motion consistency has a positive effect on monotonic agreement with quality perceived subjectively. Importantly, alternative motion-based extensions, such as IV-PSNR computed exclusively on motion fields or IV-PSNR with DOF-assisted corresponding pixel shift search, led to a noticeable degradation of the correlation. This implies that the proposed four-component IV-PSNR$_{YUVM}$ variant provides the most effective way of incorporating motion information into the IV-PSNR metric.

These findings validate the relevance of temporal motion consistency analysis for future objective quality assessment frameworks. Considering that the proposal enhances the original IV-PSNR metric, which itself already outperforms other state-of-the-art objective quality metrics, it can be stated that the proposal is valuable and highly practical to be used in real immersive video systems.

Finally, it should be clarified that the contribution of the paper is positioned not merely as an improvement in metric performance, but also as a methodologically sound and repro-

ducible research on temporal consistency analysis in IV-PSNR. We believe that bridging the conceptual gap of the "temporally blind" IV-PSNR metric constitutes a meaningful contribution to the immersive video quality assessment and could be a basis for future works on including temporal consistency analysis in immersive video quality assessment.

Future work will focus on further increasing the efficiency of the proposed metric. Since the four-component IV-PSNR$_{YUVM}$ has been shown to be the most promising direction, the work will focus on improving the DOF estimation by leveraging more advanced motion estimation techniques, i.e., more accurate deep-learning-based DOF algorithms (e.g., RAFT, FlowNet2). Furthermore, our future efforts will focus on several key areas to enhance the metric's generalizability and accuracy by investigating how temporal information can be incorporated more reliably across different motion patterns and content types, with the goal of improving the robustness and stability of the metric in challenging immersive and non-immersive scenarios. These steps aim to maximize the correlation between our objective quality metric and subjective human perception for immersive video applications.

Author Contributions: Conceptualization, J.S., B.S. and A.D.; methodology, A.D.; software, J.S. and B.S.; validation, B.S.; writing, J.S., B.S., T.G. and A.D. 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 original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Wien, M.; Boyce, J.M.; Stockhammer, T.; Peng, W.-H. Standardization status of immersive video coding. *IEEE J. Emerg. Sel. Top. Circuits Syst.* **2019**, *9*, 5–17. [CrossRef]
2. Boyce, J.M.; Dore, 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]
3. 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]
4. Wang, Z.; Bovik, A.; 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]
5. The Netflix Tech Blog, June 2016. Available online: https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652 (accessed on 20 December 2025).
6. Ghazanfari, S.; Garg, S.; Krishnamurthy, P.; Khorrami, F.; Araujo, A. R-LPIPS: An Adversarially Robust Perceptual Similarity Metric. *arXiv* **2023**. [CrossRef]
7. Pinson, M.H.; Wolf, S. A new standardized method for objectively measuring video quality. *IEEE Trans. Broadcast.* **2004**, *50*, 312–322. [CrossRef]
8. Soundararajan, R.; Bovik, A.C. Video Quality Assessment by Reduced Reference Spatio-Temporal Entropic Differencing. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *23*, 684–694. [CrossRef]
9. Seshadrinathan, K.; Bovik, A.C. Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos. *IEEE Trans. Image Process.* **2010**, *19*, 335–350. [CrossRef] [PubMed]
10. Moorthy, A.K.; Bovik, A.C. A motion compensated approach to video quality assessment. In Proceedings of the 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, USA, 1–4 November 2009; pp. 872–875.
11. Stankiewicz, O.; Domanski, M.; Dziembowski, A.; Grzelka, A.; Mieloch, D.; Samelak, J. A free-viewpoint television system for horizontal virtual navigation. *IEEE Trans. Multimedia* **2018**, *20*, 2182–2195. [CrossRef]

12. Mieloch, D.; Stankiewicz, O.; Domański, M. Depth map estimation for free-viewpoint television and virtual navigation. *IEEE Access* **2020**, *8*, 5760–5776. [CrossRef]
13. Sun, Y.; Lu, A.; Yu, L. Weighted-to-Spherically-Uniform Quality Evaluation for Omnidirectional Video. *IEEE Signal Process. Lett.* **2017**, *24*, 1408–1412. [CrossRef]
14. Dziembowski, A.; Nowak, W.; Stankowski, J. IV-SSIM—The Structural Similarity Metric for Immersive Video. *Appl. Sci.* **2024**, *14*, 7090. [CrossRef]
15. Brox, T.; Malik, J. Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 500–513. [CrossRef] [PubMed]
16. Liu, C. Introduction to Dense Optical Flow. In *Dense Image Correspondences for Computer Vision*; Hassner, T., Liu, C., Eds.; Springer: Cham, Switzerland, 2016.
17. Fortun, D.; Bouthemy, P.; Kervrann, C. Optical flow modeling and computation: A survey. *Comput. Vis. Image Underst.* **2015**, *134*, 1–21. [CrossRef]
18. Dosovitskiy, A.; Fischer, P.; Ilg, E.; Hausser, P.; Hazirbas, C.; Golkov, V.; Van Der Smagt, P.; Cremers, D.; Brox, T. FlowNet: Learning optical flow with convolutional networks. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2758–2766.
19. Teed, Z.; Deng, J. RAFT: Recurrent all-pairs field transforms for optical flow. In Proceedings of the 16th European Conference on Computer Vision–ECCV, Glasgow, UK, 23–28 August 2020; pp. 402–419.
20. Farnebäck, G. Two-frame motion estimation based on polynomial expansion. In Proceedings of the Scandinavian Conference on Image Analysis, Halmstad, Sweden, 29 June–2 July 2013; Springer: Berlin/Heidelberg, Germany, 2003.
21. Senst, T.; Eiselein, V.; Sikora, T. Robust Local Optical Flow for Feature Tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2012**, *22*, 1377–1387. [CrossRef]
22. ISO/IEC. Call for proposals on 3DoF+ visual. In *Document ISO/IEC JTC1/SC29/WG11 MPEG N18145*; International Organization for Standardization: Marrakech, Morocco, 2019.
23. Elwardy, M.; Zepernick, H.J.; Hu, Y.; Chu, T.M.C. ACR360: A Dataset on Subjective 360° Video Quality Assessment Using ACR Methods. In Proceedings of the 16th International Conference on Signal Processing and Communication System, ICSPCS, Bydgoszcz, Poland, 6–8 September 2023.
24. Duan, H.; Zhai, G.; Yang, X.; Li, D.; Zhu, W. IVQAD 2017: An immersive video quality assessment database. In Proceedings of the 2017 International Conference on Systems, Signals and Image Processing (IWSSIP), Poznań, Poland, 22–24 May 2017; pp. 1–5. [CrossRef]
25. Chen, M.J.; Su, C.C.; Kwon, D.K.; Cormack, L.K.; Bovik, A.C. Full-reference quality assessment of stereopairs accounting for rivalry. *Signal Process. Image Commun.* **2013**, *28*, 1143–1155. [CrossRef]
26. Cox, S.R.; Lim, M.; Ooi, W.T. VOLVQAD: An MPEG V-PCC Volumetric Video Quality Assessment Dataset. In Proceedings of the 14th ACM Multimedia Systems Conference (MMSys '23), Vancouver, Canada, 7–10 June 2023; pp. 357–362.
27. Yang, Z.; Pan, S.; Wang, S.; Wang, H.; Lin, L.; Li, G.; Wen, Z.; Lin, B.; Tao, J.; Yu, T. ImViD: Immersive Volumetric Videos for Enhanced VR Engagement. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 16554–16564.
28. 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.
29. Baroncini, V.; Baroncini, G. 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.
30. *ISO/IEC 23090-12:2025*; Information Technology—Coded Representation of Immersive Media. Part 12: MPEG Immersive Video. International Organization for Standardization: Geneva, Switzerland, 2025.
31. Chikkerur, S.; Sundaram, V.; Reisslein, M.; Karam, L.J. Objective video quality assessment methods: A classification, review, and performance comparison. *IEEE Trans. Broadcast.* **2011**, *57*, 165–182. [CrossRef]
32. OpenCV Documentation, 25 August 2025. OpenCV: Optical Flow Algorithms. Available online: https://docs.opencv.org/3.4/d2/d84/group__optflow.html (accessed on 20 December 2025).