MULTI-SCALE STRUCTURAL SIMILARITY WITH GEOMETRIC ALIGNMENT FOR IMMERSIVE VIDEO QUALITY ASSESSMENT

Bartosz Sojka, Jakub Stankowski, Adrian Dziembowski

Institute of Multimedia Telecommunications, Poznan University of Technology, Poznań, Poland

ABSTRACT

Immersive video systems introduce unique distortions, such as pixel shifts caused by inter-view reprojection errors and color differences caused by different lighting conditions and camera sensor characteristics. These distortions are often inaccurately assessed by conventional quality metrics. While the Structural Similarity Index (SSIM) has been successfully adapted for immersive video (in IV-SSIM metric), a multi-scale approach – better correlating with the human visual system (HVS) – lacks a dedicated adaptation. This paper bridges this gap by proposing two multi-scale metrics robust to geometric distortions: IV-MS-SSIM, a low-complexity variant applying geometric alignment before the multi-scale analysis, and MS-IV-SSIM, a high-performance variant performing geometric alignment at each scale level. Experimental evaluation on the immersive video dataset demonstrates that both metrics outperform state-of-the-art metrics in terms of correlation with subjective quality. Furthermore, evaluation on the TID2013 dataset confirms that both metrics maintain high performance even for typical 2D images, showing high versatility of the proposed metrics.

Index Terms— Quality Assessment, Structural Similarity, Multi-Scale SSIM, Immersive Video

1. INTRODUCTION

Immersive video applications, ranging from free-viewpoint television (FTV) [1] to virtual reality systems using head-mounted displays (HMD) [2], fundamentally change the way the content is observed. Unlike traditional 2D video, these systems allow users to freely navigate within a 3D scene by synthesizing (rendering) [3] virtual viewports from a limited set of source views. Additionally, recent activities towards standardization of gaussian splat coding [4] also involves computation of selected viewport. However, the rendering process introduces unique distortions, such as pixel shifts caused by depth or calibration inaccuracies, as well as color differences caused by view-dependent lighting or sensor characteristics [5].

While the human visual system (HVS) is relatively tolerant to minor geometric shifts (e.g., an object shifted by a few pixels is often perceptually acceptable), conventional objective quality metrics like PSNR (peak signal-to-noise ratio) or SSIM (structural similarity index, [6]) are not, treating a slight geometric misalignment as a severe error [5], [8]. This discrepancy leads to a low correlation between objective and subjective quality.

Recently, deep learning (DL) based metrics have gained significant popularity [7]. While these networks provide high correlation on traditional 2D content, their direct application to immersive video remains heavily constrained. Due to the lack of explicit mechanisms for handling the rendering-induced artifacts, standard DL metrics fail to properly mimic the HVS in immersive video applications. Furthermore, their high computational complexity and typical reliance on GPU acceleration make them impractical for standard codec optimization loops, which require lightweight, real-time CPU-based metrics. To address the geometric distortions efficiently without relying on heavy neural network architectures, the "IV" variants of conventional metrics: IV-PSNR [5] and IV-SSIM [8] have been introduced. These metrics include a geometric alignment step of searching for corresponding pixels within a small local window in order to compensate for rendering-induced pixel shifts before calculating pixel-wise differences between two compared images.

However, both metrics operate on a single scale. Typically, the HVS perceives image quality in a multi-scale manner, depending on viewing distance and resolution [9]. The multi-scale SSIM (MS-SSIM, [9]) addresses this for 2D images, achieving better correlation with MOS than conventional, single-scale SSIM. For immersive video, however, it performs worse than IV-PSNR and IV-SSIM.

In this paper, we bridge this gap by proposing two novel multi-scale quality metrics: IV-MS-SSIM and MS-IV-SSIM. The first one is a low-complexity variant that applies geometric alignment once, before video scaling. In the second, high-precision approach, the geometric alignment is performed at each scale of the structural analysis. Experimental results demonstrate that both metrics significantly outperform state-of-the-art metrics in terms of correlation with MOS for immersive video, and maintaining the MS-SSIM's high accuracy for traditional, 2D images.

2. METHODOLOGY: TWO APPROACHES TO MULTI-SCALE SSIM FOR IMMERSIVE VIDEO

The proposed metrics extend the multi-scale structural similarity index (MS-SSIM) by including immersive-adapted

mechanisms introduced in [5] and [8]. We propose two approaches: IV-MS-SSIM – focused on computational efficiency, and MS-IV-SSIM – focused on maximum accuracy achieved through multi-scale geometric analysis.

2.1. The core "IV" component

The core "IV" module, shared by both proposed metrics, integrates two fundamental components: global color difference (GCD) compensation and geometric alignment based on corresponding pixel shift (CPS) compensation.

Let $R_c(i,j)$ and $T_c(i,j)$ denote the pixel value of the reference and test images, respectively, at spatial coordinates (i,j) for a specific color component c (e.g., Y, U, V).

First, to adjust the color characteristics of two compared images, a global color offset $\vartheta_c^{R\to T}$ is calculated, separately for each color component c:

$$\vartheta_c^{R\to T} = \frac{1}{N}\sum_{i,j}\left(R_c(i,j) - T_c(i,j)\right), \qquad (1)$$

where N is the total number of pixels.

In the second step, for every pixel (i,j) in the reference image, the algorithm searches for the most similar pixel within a local search window W centered at (i,j) in the test image. Its coordinates (i',j'), calculated as:

$$(i',j') = \underset{(x,y)\in W}{\operatorname{argmin}} |R_c(i,j) - T_c(i,j) - \vartheta_c^{R\to T}| \qquad (2)$$

are used for the calculation of pixel (i,j) of the geometrically aligned test image T_c':

$$T_c'(i,j) = T_c(i',j') + \vartheta_c^{R\to T}, \qquad (3)$$

in which each pixel contains the value from the best-matching neighbor in the original test image, corrected by the global color difference.

Finally, consistent with the design of IV-PSNR [5] and IV-SSIM [6], the proposed metrics employ a symmetrical evaluation scheme. Therefore, the color and geometry alignment (and the entire similarity calculation) are performed in two directions. In the first pass $R \to T$, the metric compares R with the aligned test image T'. In the second pass $T \to R$, the algorithm generates a geometrically aligned reference image R' in an analogous way as described earlier. The final quality score for both proposed metrics is defined as the minimum of two directional scores:

$$\text{score}(R,T) = \min(\text{score}^{R\to T}, \text{score}^{T\to R}), \qquad (4)$$

2.2. IV-MS-SSIM: Preprocessing approach

The first proposed metric, IV-MS-SSIM, separates the geometric compensation from the multi-scale analysis to minimize computational complexity. As presented in Fig. 1A, the process starts with applying the core IV algorithm (color and geometric alignment described in Section 2.1) to the full-resolution test image T relative to the reference image R, resulting in the generation of the aligned image T'.

Then, the standard MS-SSIM algorithm [7] is applied to the pair (R,T'):

$$\text{IV-MS-SSIM}^{R\to T} = \qquad (5)$$

$$= [l_M(R,T')]^{\alpha_M} \cdot \prod_{m=1}^{M} [c_m(R,T')]^{\beta_m} \cdot [s_m(R,T')]^{\gamma_m},$$

where $l_M(\cdot)$ denotes luminance similarity at the lowest scale M (by default equal to 5), $c_m(\cdot)$ and $s_m(\cdot)$ – contrast and structural similarity at scale m, respectively, and $\alpha_M, \beta_m, \gamma_m$ are the standard weights used in MS-SSIM [7].

The main advantage of the IV-MS-SSIM approach is its speed, as the computationally expensive corresponding pixel search is performed only once. Additionally, if one calculates more metrics at once (i.e., IV-SSIM and IV-MS-SSIM), the corresponding pixel search can be performed once, and its result can be reused by both metrics.

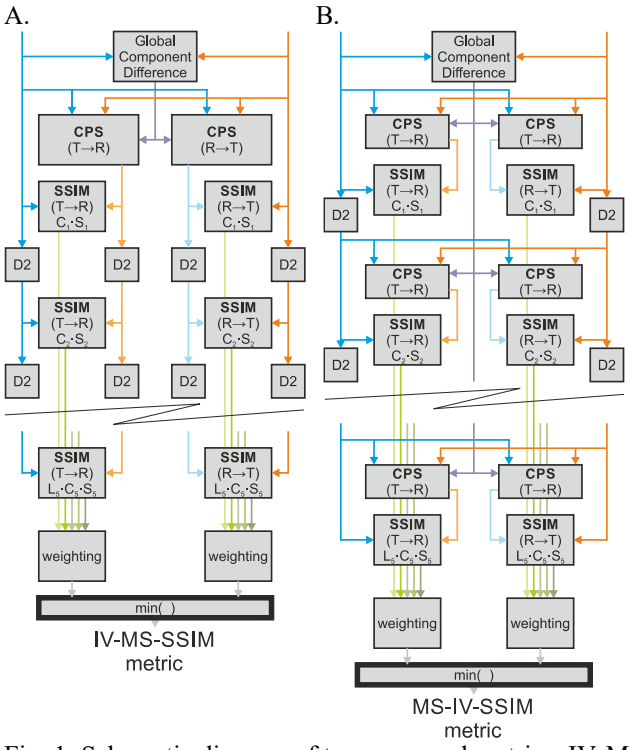


Fig. 1. Schematic diagram of two proposed metrics: IV-MS-SSIM (A) and MS-IV-SSIM (B); CPS – compensation of corresponding pixel shift, D2 – low-pass filtering + decimation of the image (by a factor of 2); arrow color notation: blue – reference image, orange – test image, grey – global component difference vector (one value per Y, U, and V channels), all shades of green – per-scale SSIM values.

2.3. MS-IV-SSIM: Per-scale approach

In the second metric, MS-IV-SSIM, the color alignment is performed as the preprocessing step, similarly to the variant described in Section 2.2. Geometric alignment in MS-IV-SSIM, however, is integrated directly into the multi-scale analysis. As presented in Fig. 1B, instead of aligning one of the images once at the beginning, images R and T are iteratively filtered and downsampled by a factor of 2, creating a set of image pairs (R_m, T_m) for each scale $m \in \{1, \dots, M\}$.

The geometric alignment is then performed independently at each scale level, producing a scale-specific

aligned image T'_m. Consequently, the structural similarity is calculated as:

$$\text{MS-IV-SSIM}^{R\to T} = \\ = [l(R_M, T'_M)]^{\alpha_M} \cdot \prod_{m=1}^{M} [c(R_m, T'_m)]^{\beta_m} \cdot [s(R_m, T'_m)]^{\gamma_m}. \quad (6)$$

It should be noted that the functions $l(\cdot)$, $c(\cdot)$, and $s(\cdot)$ in (6) differ from those in (5). While the notation in (5) implies internal downsampling to scale m, the functions in (6) represent the basic, single-scale similarity functions (the same as for baseline SSIM metric [6]) applied directly to the already downsampled image pairs (R_m, T'_m).

The per-scale approach allows the MS-IV-SSIM metric to better adapt to geometric distortions that vary across spatial frequencies, offering higher performance than the preprocessing approach in terms of correlation with subjective quality.

3. EXPERIMENTAL SETUP

To evaluate the performance of the proposed IV-MS-SSIM and MS-IV-SSIM, we compared them against 22 state-of-the-art objective quality metrics, including 18 classical algorithms, such as single- and three-component versions of PSNR, SSIM [6], and MS-SSIM [9], IV-PSNR [5], IV-SSIM [6], VMAF [10], VIF [11], LPIPS [12], SFF [13], FSIM [14], CS-PSNR [15], PSNR-HA and PSNR-HMA [16], MW-PSNR [17], and MP-PSNR [18], as well as four recent state-of-the-art DL-based metrics: TOPIQ [19], DMM [20], UVQ [21], and DOVER [22].

In total, two experiments have been conducted. The first experiment focused on the primary domain of the proposed metrics – immersive video. It utilized the dataset from "MPEG Call for Proposals on 3DoF+ Visual" [23], comprising five test sequences processed using seven different immersive compression methods. This 6DoF dataset is characterized by sparse camera setups, leading to massive inter-view baselines and complex rendering artifacts. Such conditions reflect the practical constraints of real-world immersive video acquisition, where the use of dense camera arrays is impractical. Consequently, this dataset presents a significantly different and harder challenge for geometric alignment than datasets with denser camera arrangements, e.g., light field arrays.

The second experiment aimed to assess the robustness and generalization capability of the proposed metrics on standard 2D content. For this purpose, we used the TID2013 dataset [24] comprising 25 reference and 3000 images with various distortions. This allowed for benchmarking IV-MS-SSIM and MS-IV-SSIM outside their intended use. This experiment serves as a mandatory sanity check for the proposed metrics. The corresponding pixel search component does not compute actual 3D geometric object displacement; instead, it is a purely signal-driven operation that identifies the most photometrically similar pixel within a local window. Consequently, even on traditional 2D images affected by standard noise or compression (without view synthesis), the local search frequently utilizes non-zero shifts to minimize localized errors. The second experiment was performed to ensure that the proposed geometric alignment remains backward-compatible with standard 2D distortions and does not degrade performance on traditional content.

The performance of all the objective metrics was evaluated against subjective MOS values using Spearman and Kendall rank order correlation coefficients (SROCC and KROCC, respectively [25]).

4. RESULTS AND DISCUSSION

4.1. Performance on immersive video (MIV CfP)

In the first experiment, the proposed metrics were evaluated in the immersive video scenario, using the MIV CfP database, which contains synthesized views affected by various rendering-induced geometric and color distortions. The correlation with subjective scores is presented in Fig. 2.

As can be observed, in an immersive video coding scenario, conventional metrics perform poorly. The conventional MS-SSIM metric, despite being a multi-scale metric, achieves correlation smaller than 0.65 and 0.5 in terms of SROCC and KROCC, respectively. This confirms that without geometric and color alignment, the effectiveness of the multi-scale analysis is significantly degraded. The introduction of the "IV" component in existing metrics significantly improves the correlation, increasing SROCC to over 0.7 and almost 0.8 for IV-PSNR and IV-SSIM, respectively.

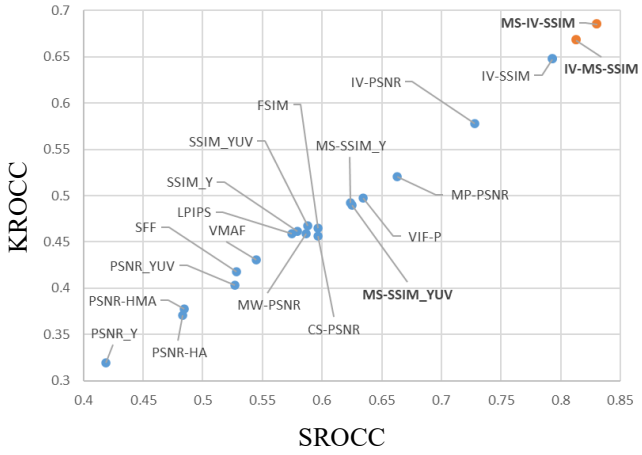


Fig. 2. SROCC and KROCC values for considered metrics; immersive video coding scenario.

The proposed metrics outperform all reference methods. IV-MS-SSIM outperforms the best-correlating state-of-the-art metric – IV-SSIM – by 0.02 in both SROCC and KROCC. MS-IV-SSIM metric increases the correlation even more, providing a further increase of 0.015. By performing alignment at each scale level, MS-IV-SSIM effectively models how the HVS perceives inter-view reprojection artifacts, distinguishing disruptive structural errors from acceptable geometric shifts.

Furthermore, all NL-based metrics were not able to properly mimic the HVS in immersive video application. They achieved SROCC and KROCC values below 0.4 and 0.3, respectively. This failure highlights a critical limitation of standard DL models: they lack explicit geometric alignment mechanisms and evaluate features at fixed spatial locations. Consequently, they are sensitive to negligible geometric shifts and rendering-induced artifacts inherent to 6DoF synthesis, treating them as massive structural errors. This proves that without a dedicated alignment, the advanced DL-based metrics are unusable for immersive video.

4.2. Performance on typical 2D images (TID2013)

In order to verify the versatility of the proposed metrics, they were additionally evaluated on the TID2013 dataset [24], containing distortions typical to 2D, non-immersive images.

Fig. 3 illustrates the SROCC and KROCC values of the proposed metrics compared to state-of-the-art algorithms. As can be observed, full-reference DL models (TOPIQ and DMM) and the baseline MS-SSIM (specifically, its three-component version – $\text{MS-SSIM}_{\text{YUV}}$) achieve high correlation with subjective scores, outperforming the majority of classical metrics.

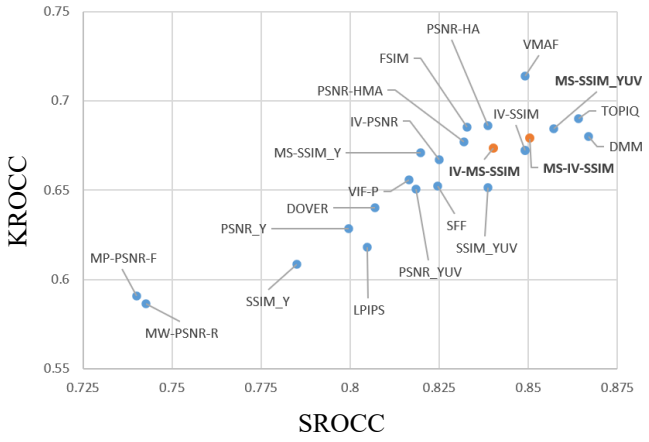


Fig. 3. SROCC and KROCC values for all considered metrics; non-immersive video scenario – TID2013 database.

Crucially, both IV-MS-SSIM and MS-IV-SSIM perform similarly to these high-performing references. This result confirms that the proposed metrics are versatile and can be used also in general, non-immersive video applications without performance degradation.

4.3. Complexity Analysis

The computational complexity has been verified on a system equipped with an AMD Ryzen 9 5950X 16-Core/32-Thread processor, 32 GiB DDR4/3200MT memory. For the calculation of all the metrics, we have used the QMIV 3.0 software [26], with appropriate modifications implementing the MS-IV-SSIM calculations. The time results of 16-threaded calculations for a single frame of a high-resolution (i.e., 4096x4096) have been shown in Fig. 4.

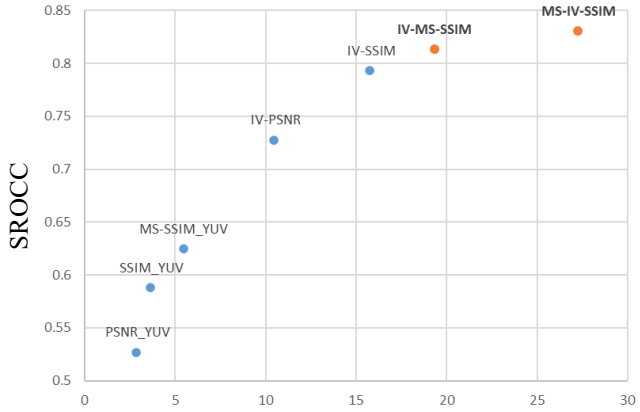


Fig. 4. Computational time vs. SROCC for typical and immersive-adapted metrics; immersive video coding scenario; implementation: QMIV 3.0 software [26].

Both IV-MS-SSIM and MS-IV-SSIM metrics show clear SROCC improvement over other metrics while still retaining reasonable calculation times within the 33 ms budget required for real-time 30 FPS applications, even for such a high-resolution video.

As presented, IV-MS-SSIM offers faster calculation times than MS-IV-SSIM. MS-IV-SSIM, however, is more robust, but needs to calculate the "IV" component for each scale level, which introduces overhead, increasing the calculation times.

Furthermore, we evaluated the execution times of the DL-based models within the same CPU environment to explicitly highlight the "performance vs. complexity" trade-off. While these metrics can offer high correlation on 2D images, their execution times for a single 4K frame are prohibitively long for practical applications: approximately 1.75 s for TOPIQ and 3.09 s for DMM. Similarly, no-reference models also severely violate the real-time budget (0.40 s for UVQ and 0.98 s for DOVER). Additionally, their low correlation with subjective quality for 6DoF content (SROCC < 0.4) validates the practical value of the proposed MS-IV-SSIM and IV-MS-SSIM metrics, which are fast, hardware-friendly, and accurate multi-scale metrics.

It is worth noting that 4096x4096 was the highest resolution available in the utilized datasets (specifically, in the MIV CfP dataset), representing the worst-case scenario for computational complexity.

4.4. Qualitative analysis

To provide deeper insight into the performance differences between conventional and immersive-adapted metrics, a qualitative analysis using spatial similarity maps was conducted. Fig. 5 presents a visual comparison of quality assessment for a synthesized view containing geometric distortions (sequence *Frog* [27], one of the sequences from the MIV CfP dataset).

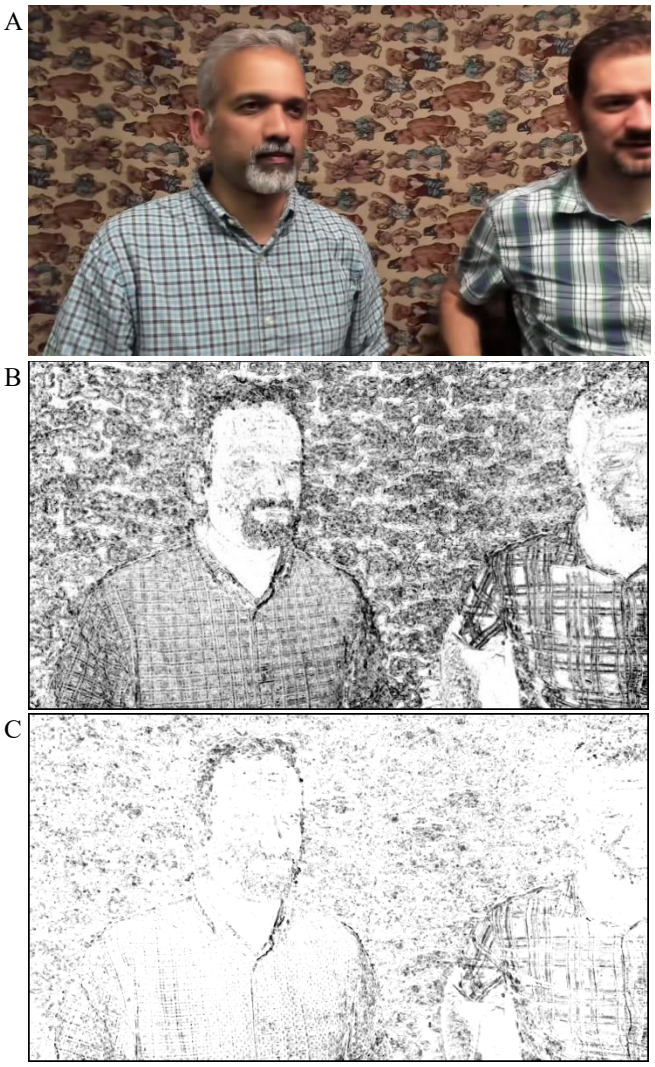


Fig. 5. Qualitative comparison of the *Frog* sequence [27]. A: synthesized view, B: similarity map of MS-SSIM, C: similarity map of IV-MS-SSIM. Both similarity maps were calculated on the highest scale (full resolution).

For visualization purposes, the similarity maps were computed at the single, highest scale (full resolution). Fig. 5A presents the evaluated synthesized view. Figs. 5B and 5C illustrate the structural similarity maps for the standard MS-SSIM and proposed IV-MS-SSIM, respectively. In these maps, pixel brightness shows the local quality score: brighter areas represent high similarity to the reference, while darker regions indicate larger errors (lower structural similarity).

As shown in Fig. 5B, the standard MS-SSIM similarity map is heavily affected by rendering-induced geometric shifts, which can be observed as dark contours along high-contrast edges, appearing not only at object boundaries but also within detailed textures, such as the grid pattern on the shirt. The metric interprets these minor misalignments as severe structural errors.

IV-MS-SSIM, on the other hand, demonstrates the effectiveness of the proposed immersive-adapted modification. As shown in Fig. 5C, the dark error contours are significantly reduced, following the behaviour of the human visual system, which is insensitive to minor shifts. However, it is crucial to note that the metric remains sensitive to excessive distortions. As can be observed on the right side of the face, where geometric misalignment exceeds the geometric compensation threshold, the map correctly retains low similarity scores (darker regions). This confirms that the metric effectively differentiates between severe structural degradations and negligible rendering offsets, ensuring a more accurate objective quality assessment.

5. CONCLUSIONS AND FUTURE WORKS

In this paper, we addressed the problem of objective quality assessment for immersive video applications, where traditional metrics fail due to the presence of rendering-induced geometric and color artifacts. We proposed two novel multi-scale metrics: IV-MS-SSIM and MS-IV-SSIM, which extend the commonly used MS-SSIM [7] metric by incorporating color and geometric alignment mechanisms.

Experimental evaluation on the immersive video dataset demonstrated that the proposed metrics significantly outperform state-of-the-art approaches. The high-performance variant, MS-IV-SSIM, achieved the highest correlation with subjective scores, confirming that performing the geometric alignment at each scale level efficiently mimics the human perception of rendered content. On the other hand, the low-complexity variant, IV-MS-SSIM, offers a trade-off between accuracy and computational cost, making it practical for time-constrained applications.

Furthermore, additional validation on the TID2013 dataset [24] confirmed the versatility of the proposed metrics, which maintain high performance for typical 2D images, statistically indistinguishable from the baseline MS-SSIM. These results ensure that the introduced immersive-adapted modifications of the MS-SSIM metric do not degrade the correlation when the evaluated content lacks rendering-induced distortions.

To support reproducible research and facilitate further development of immersive video quality assessment, an open-source implementation of the proposed metric is available [26]. This follows previous contributions regarding IV-PSNR and IV-SSIM software.

Future work will focus on extending the proposed metrics to the temporal domain by incorporating motion-related information to better assess the spatio-temporal consistency of synthesized videos, which is crucial for the overall Quality of Experience in immersive video systems.

6. ACKNOWLEDGMENT

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

7. REFERENCES

[1] M. Tanimoto, “Overview of FTV (free-viewpoint television),” *International Conference on Multimedia and Expo*, New York, USA, pp. 1552-1553, 2009.

[2] Anthes, C. et al., “State of the art of virtual reality technology,” *IEEE Aerospace Conference*, Big Sky, USA, 2016.

[3] S. Fachada et al., “Depth image based view synthesis with multiple reference views for virtual reality,” *3DTV Conference 2018*, Stockholm/Helsinki, Sweden/Finland, Jun. 2018.

[4] ISO/IEC, “Common test conditions for gaussian splat coding,” in ISO/IEC JTC1/SC29/WG04 MPEG VC N0751, 2025.

[5] Dziembowski, A. et al., “IV-PSNR—The objective quality metric for immersive video applications,” *IEEE Tr. on Circuits and Systems for Video Technology* 32, pp. 7575–7591, 2022.

[5] Wang, Z. et al., “Image quality assessment: From error measurement to structural similarity,” IEEE Transactions on Image Processing 13, pp. 600–612, 2004.

[7] J. Yang et al., “Deep Learning Based Image Quality Assessment: A Survey,”, Procedia Computer Science, Vol. 221, pp 1000-1005, 2023.

[8] Dziembowski, A. et al., „IV-SSIM—The Structural Similarity Metric for Immersive Video,” *Applied Sciences* 14 (16), 2024,

[9] Wang, Z. et al., “Multiscale structural similarity for image quality assessment,” *37th Asilomar Conference on Signals, Systems & Computers*, Pacific Grove, USA, Nov. 2003.

[10] Li, Z. et al., “Toward a practical perceptual video quality metric,” *Netflix Technology Blog; Technical Report*, 2016. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652; accessed on 31 January 2026.

[11] Sheikh, H.R. and Bovik, A.C., “Image information and visual quality,” *IEEE Tr. on Image Process*. 15, pp. 430–444, 2006.

[12] Zhang, R. et al., “The unreasonable effectiveness of deep features as a perceptual metric,” *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June 2018.

[13] Chang, H.W. et al., “Sparse feature fidelity for perceptual image quality assessment,” *IEEE Transactions on Image Processing* 22, pp. 4007–4018, 2013.

[14] Zhang, L. et al., “FSIM: A feature similarity index for image quality assessment,” *IEEE Transactions on Image Processing* 20, pp. 2378–2386, 2011.

[15] Shang, X. et al., “Color-sensitivity-based combined PSNR for objective video quality assessment,” *IEEE Tr. on Circuits and Systems for Video Technology* 29, pp. 1239–1250, 2019.

[16] Ponomarenko, N. et al., “Modified image visual quality metrics for contrast change and mean shift accounting,” *CADSM conference*, Polyana, Ukraine, 2011.

[17] Sandić-Stanković, D. et al., “DIBR-synthesized image quality assessment based on morphological multi-scale approach,” *EURASIP J. Image and Video Processing* 4, 2016.

[18] Sandić-Stanković, D. et al., “DIBR synthesized image quality assessment based on morphological pyramids,” *3DTV-Conference (3DTV-CON)*, Lisbon, Portugal, May 2015.

[19] C. Chen et al., “TOPIQ: A Top-Down Approach From Semantics to Distortions for Image Quality Assessment,” IEEE Transactions on Image Processing, vol. 33, pp. 2404-2418, 2024

[20] B. Chen et al., “Debiased Mapping for Full-Reference Image Quality Assessment,” IEEE Transactions on Multimedia, vol. 27, pp. 2638-2649, 2025.

[21] Y. Wang et al., “Rich Features for Perceptual Quality Assessment of UGC Videos”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13435-13444, 2021.

[22] H. Wu et al., “Exploring Video Quality Assessment on User Generated Contents from Aesthetic and Technical Perspectives,” 2023 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 20087-20097, Paris, France, 2023.

[23] ISO/IEC, “Call for proposals on 3DoF+ visual,” Doc. ISO/IEC JTC1/SC29/WG11 MPEG N18145, Marrakech, Morocco, 2019.

[24] Ponomarenko, N. et al., “Image database TID2013: Peculiarities, results and perspectives,” *Signal Processing and Image Communication* 30, pp. 57–77, 2015.

[25] Chikkerur, S. et al., “Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison,” *IEEE Transactions on Broadcasting* 57, pp. 165–182, 2011.

[26] ISO/IEC, “Software manual of QMIV 2”, Doc. JTC1/SC29/WG04 MPEG VC/N0580, Antalya, Turkey, 2024.

[27] Salahieh, B. et al., “Kermit Test Sequence for Windowed 6DoF Activities,” *Doc. ISO/IEC JTC1/SC29/WG11 MPEG/M43748*, Ljubljana, Slovenia, 2019.