Perceptual Quality Assessment of Spatial Videos on Apple Vision Pro

Afshin Gholami
University of Klagenfurt
Klagenfurt, Austria
afshingh@edu.aau.at

Sara Baldoni
University of Padova
Padova, Italy
sara.baldoni@unipd.it

Federica Battisti
University of Padova
Padova, Italy
federica.battisti@unipd.it

Wei Zhou
Cardiff University
Cardiff, UK
ZhouW26@cardiff.ac.uk

Christian Timmerer
University of Klagenfurt
Klagenfurt, Austria
christian.timmerer@aau.at

Hadi Amirpour
University of Klagenfurt
Klagenfurt, Austria
hadi.amirpour@aau.at

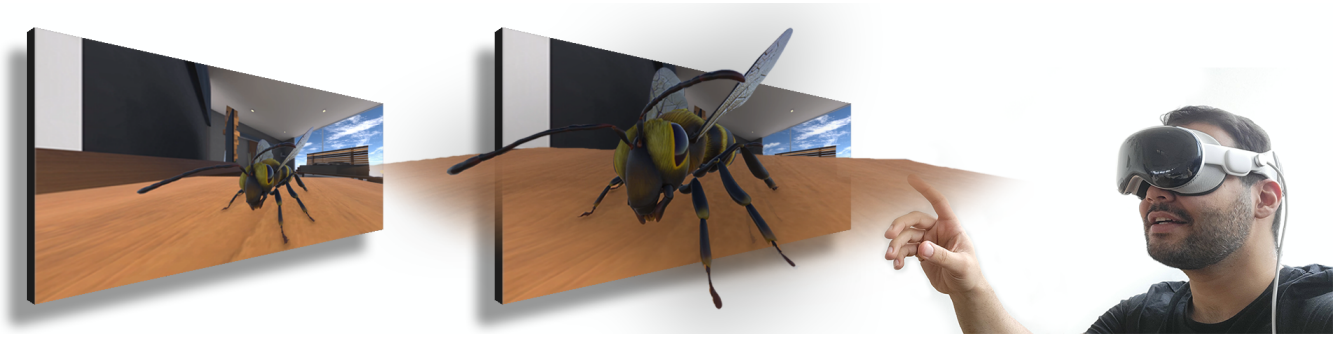


Figure 1: Quality evaluation of 2D and stereoscopic videos on the Apple Vision Pro (AVP).

Abstract

Immersive stereoscopic (3D) video experiences have entered a new era with the advent of smartphones capable of capturing stereoscopic videos, advanced video codecs optimized for multiview content, and Head Mounted Displays (HMDs) that natively support stereoscopic video playback. In particular, Apple's recent introduction of *spatial video* capture on the recent iPhone Pro series and immersive playback on the Apple Vision Pro (AVP) has accelerated the mainstream adoption of stereoscopic content. In this work, we evaluate the quality of spatial videos encoded using optimized x265 software implementations of Multiview HEVC (MV-HEVC) on the AVP and compare them with their corresponding 2D versions through a subjective test.

To support this study, we introduce *SV-QoE*, a novel dataset comprising video clips rendered with a twin-camera setup that replicates the human inter-pupillary distance. Our analysis reveals that spatial videos consistently deliver a superior Quality of Experience (QoE) when encoded at similar bitrates, with the benefits becoming more pronounced at higher bitrates. Additionally, renderings at closer distances exhibit significantly enhanced video quality and depth perception, highlighting the impact of spatial proximity on immersive viewing experiences.

We further analyze the impact of disparity on depth perception and examine the correlation between Mean Opinion Score (MOS) and established objective quality metrics such as PSNR, SSIM, MS-SSIM, VMAF, and AVQT. Additionally, we explore how video quality and depth perception together influence overall quality judgments. The complete dataset, including videos and subjective scores, is publicly available at https://github.com/cd-athena/SV-QoE.

CCS Concepts

• **Information systems** → **Multimedia streaming**.

Keywords

stereoscopic, MV-HEVC, QoE, spatial video, depth

ACM Reference Format:
Afshin Gholami, Sara Baldoni, Federica Battisti, Wei Zhou, Christian Timmerer, and Hadi Amirpour. 2025. Perceptual Quality Assessment of Spatial Videos on Apple Vision Pro. In *Proceedings of the 3rd International Workshop on Interactive eXtended Reality (IXR '25), October 27–28, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3746269.3760422

IXR '25, Dublin, Ireland

ACM ISBN 979-8-4007-2051-2/2025/10
https://doi.org/10.1145/3746269.3760422

1 Introduction

Immersive media has transformed how users engage with digital content, extending beyond traditional viewing to provide highly interactive and engaging experiences [1, 2]. Advances in Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR) have led

to the development of immersive environments that enhance user presence and interaction. As display hardware continues to evolve, with innovations such as high-resolution HMDs, eye-tracking, and spatial audio, immersive media is becoming increasingly realistic and accessible. Additionally, improvements in content creation tools are pushing the boundaries of digital storytelling, gaming, training simulations, and remote collaboration [3–5]. With the growing demand for lifelike and interactive experiences, immersive media is reshaping entertainment, education, and professional applications [6].

Among the various forms of immersive media, stereoscopic videos have regained popularity thanks to their ability to enhance realism through depth perception. Stereoscopic imaging works by capturing two slightly different perspectives of a scene (*i.e.*, one for the left eye and one for the right eye), mimicking the natural disparity of human vision [7]. This disparity allows the brain to interpret depth, resulting in a 3D perception of the content. Typically, stereoscopic videos are recorded using a dual-camera setup. During playback, these two images are displayed separately to each eye, either through passive polarization, active shutter glasses, or direct display on head-mounted devices [8].

Widespread adoption of this format has historically been limited by challenges across the entire multimedia pipeline, including content creation, efficient encoding, and display compatibility. Recent advancements have significantly lowered these barriers, making stereoscopic video a more accessible and scalable format for immersive media. The integration of stereoscopic video capture into consumer-grade devices, such as the iPhone Pro, has enabled effortless content creation without the need for specialized camera setups. On the encoding side, optimized HEVC-based [19] compression software, such as x265, ensures that high-quality stereoscopic video can be efficiently stored and streamed while maintaining perceptual quality. Furthermore, the emergence of HMDs with native support for stereoscopic video playback, such as the AVP and Meta Quest 3, has provided a dedicated ecosystem for consuming stereoscopic content. These developments have effectively bridged the gap between content creation, encoding, and rendering, thus enabling stereoscopic videos to become a viable and accessible format for next-generation multimedia applications.

While in Apple's definition, *spatial* videos refer to videos recorded with the iPhone Pro or AVP [18], encoded with their MV-HEVC [20] codec and displayed on the AVP, in this paper, we define *spatial* videos as stereoscopic content encoded using the MV-HEVC format and designed for seamless playback on the AVP.

Despite the advances in spatial video capture and display, there remains a significant gap in the research on spatial videos for immersive platforms, as most studies focus primarily on traditional 2D videos. While objective and subjective quality evaluation methodologies are well established for conventional video formats, their applicability to spatial videos, particularly on HMDs, remains less explored. Factors such as depth perception, binocular disparity, compression artifacts, and motion cues all influence the perceptual quality of spatial video in ways that differ from 2D content [21, 22]. Another key challenge in this domain is the lack of publicly available datasets for spatial video quality assessment. Existing datasets primarily focus on 2D content and do not adequately capture the perceptual nuances of spatial formats. Moreover, most of the stereoscopic datasets are limited to standard resolutions and frame rates, lacking high-resolution (*e.g.*, 4K) and high frame-rate (*e.g.*, 60fps) content, which are increasingly common in modern immersive video applications.

To address these limitations, (*i*) we introduce a novel dataset, *SV-QoE*, specifically designed for spatial video quality assessment, featuring diverse high-resolution 4K and high-frame-rate 60fps video sequences encoded at multiple quality levels (available at https://github.com/cd-athena/SV-QoE.) Furthermore, (*ii*) this paper bridges the research gap by conducting a comprehensive subjective evaluation of both 2D and spatial video quality on the AVP. We systematically assess viewer responses across three distinct quality levels, ensuring that all content is encoded using the same optimized open-source x265 codec at similar bitrates. Our experimental design involves controlled subjective tests, where participants experience a series of video sequences in both 2D and spatial formats. To comprehensively assess the viewing experience, participants evaluate three key aspects:

(1) video quality (Q1),
(2) depth perception (Q2), and
(3) overall quality (Q3).

(*iii*) We also evaluate the correlation between disparity and perceived depth, shedding light on how disparity cues affect depth perception and contribute to overall quality in stereoscopic content. (*iv*) Additionally, we examine the correlation between well-known objective quality metrics (*i.e.*, PSNR, SSIM, VMAF, and AVQT) and subjective video quality scores, offering a deeper understanding of the reliability and applicability of existing objective models in evaluating 2D and spatial video content on AVP. (*v*) Finally, we analyze the impact of video quality and depth perception on overall quality, providing insights into how these factors influence the immersive viewing experience.

2 Related Work

The stereoscopic video pipeline comprises several key modules: creation, encoding, delivery, rendering, and quality evaluation. The creation phase involves capturing left and right views using dual-camera systems, which require precise calibration and synchronization to ensure consistency. Existing stereoscopic video datasets have been summarized in Table 1. These datasets typically lack high-resolution and high frame-rate content, which limits their applicability to modern immersive viewing scenarios.

Recently, deep learning methods have also been applied to generate stereoscopic videos. For example, Zhang *et al.* [23] introduced a novel framework for converting 2D videos into stereo videos. Their approach employs depth-warping and blend-inpainting techniques, incorporating a mask-based hierarchical feature update refiner and a disparity expansion strategy to improve inpainting accuracy and reduce foreground bleeding. During the encoding stage, advanced compression methods are utilized to maintain high perceptual quality while reducing bitrate. In particular, inter-view redundancy between the left and right views is exploited to enhance compression efficiency, as explored in MV-HEVC [20] and its optimized variants [24, 25]. The delivery phase focuses on reliably transmitting this content over varying network conditions

Table 1: Overview of stereoscopic video datasets.

Dataset Name	Year	Resolution	Description
KITTI Stereo 2012 [9]	2012	1226×370	Outdoor driving scenes
KITTI Stereo 2015 [10]	2015	1242×375	Outdoor driving scenes with dynamic scenes with objects
SceneFlow [11]	2016	960×540	Synthetic stereo sequences
MPI-Sintel [12]	2012	1024×436 (24fps)	Synthetic scenes with complex motion and visual effects
RMIT3DV HD 3D Video [13]	2012	1920×1080 (25fps)	Diverse urban scenes
EPFL MMSPG 3DVQA [14]	2010	1920×1080 (25fp)	High-quality visual variations
Stereo Video Database [15]	2010	1920×1080 (25fps)	Stereo cinema post-production
NAMAD3D [16]	2012	1920×1080(25fps)	Natural 3D scenes with twin-lens camera
SVSR-Set [17]	2022	1920×1080 (30fps)	Indoor/outdoor with varied motion and lighting
SVD [18]	2025	1920×1080 (30fps)	Spatial videos taken by iPhone and Apple Vision Pro
SV-QoE (Ours)	2025	3840×2160 (60fps)	High resolution, high frame-rate synthetic sequences captured from two distances

by employing adaptive streaming protocols [26] and network optimization strategies, as explored by Chen *et al.* [27]. Finally, the quality evaluation module employs both subjective assessments and objective metrics to gauge the perceptual impact of distortions unique to stereoscopic content [21].

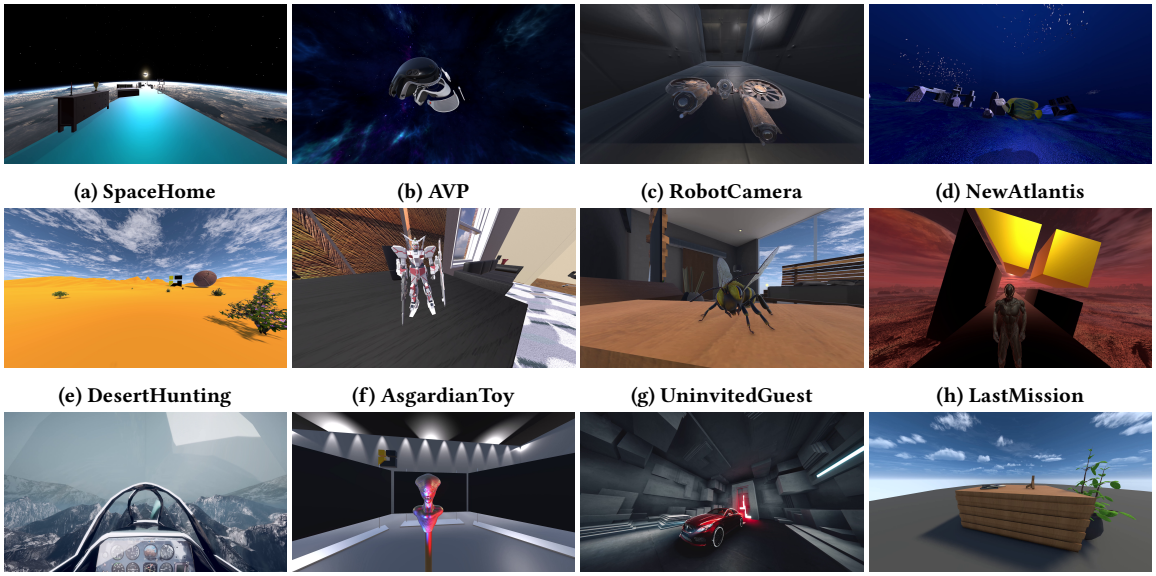
The quality evaluation of stereoscopic multimedia involves multiple parameters, including video quality, depth perception, and overall satisfaction. Goldmann *et al.* [28] investigated the impact of acquisition distortions, such as the baseline distance between left and right cameras, on the perceived quality of stereoscopic images and videos. Their findings indicate that as the camera baseline distance increases, perceived quality decreases, highlighting the critical role of acquisition setup in maintaining high-quality stereoscopic content. Zhou *et al.* [21] presented a comprehensive study on the visual quality assessment of 3D-HEVC compressed stereoscopic videos. They analyzed the impact of video compression and depth quality on the overall QoE. They develop a No Reference (NR) bitstream-level objective quality assessment model that extracts key features from 3D-HEVC bitstreams, such as quantization parameters and prediction residuals, to predict perceived video quality. Chen *et al.* [29] introduced a depth perception quality metric and extended it to an NR stereoscopic video quality assessment. Wan *et al.* [30] analyzed the impact of coding artifacts on depth perception in stereoscopic 3D videos, revealing that compression distortions introduced by the Advanced Video Coding (AVC) standard can significantly alter depth quality. Their subjective experiments showed that coding artifacts affect different spatial frequency components unequally, with high-pass and band-pass components being more crucial for depth perception than low-pass components. They also found that horizontal orientation structures play a dominant role in depth perception, and distortions in these components lead to more noticeable depth degradation.

Compared to 2D images, a 3D image consists of two 2D images – left and right views – introducing additional challenges in objective quality assessment. When the left and right views exhibit different types and levels of artifacts, asymmetric distortions occur, making it more complex to evaluate the quality of 3D images [31]. The simple average of the predicted quality scores from both views does not account for the binocular processing mechanisms of the Human Visual System (HVS) [32, 33]. To address these challenges, various 3D image quality assessment (IQA) methods have been developed, incorporating specific 3D characteristics. Notable approaches include the cyclopean model [34], weighted SSIM (W-SSIM), and weighted FSIM (W-FSIM) [35]. For 3D omnidirectional image quality assessment (OIQA), a multi-viewport-based model has been introduced [36], while the stereoscopic omnidirectional image quality evaluator (SOIQE) was designed based on predictive coding theory [37]. Galkandage *et al.* [38] introduced a novel HVS model inspired by physiological findings characterizing the motion-sensitive response of complex cells in the primary visual cortex. The proposed full-reference stereoscopic video quality assessment method leverages this model to enhance the prediction accuracy of perceived video quality.

3 Dataset Creation

In this paper, we present *SV-QoE*, a dataset that includes 12 scenes created using Unity engine 6 (Version: 6000.0.28f1) to showcase a variety of artistic styles, ensuring accurate depth representation and minimizing capturing distortions. Each scene was recorded as a 10-second clip. The Unity engine camera was configured with a perspective projection to ensure realistic depth perception, with a 137° horizontal Field of View (FoV) for a wide yet natural perspective, chosen to closely match the effective horizontal viewing range of the human visual system and modern VR headsets, thereby providing an immersive experience without introducing noticeable geometric distortion.

To capture both 2D and spatial video content, we employed a three-virtual camera setup. The central camera, located at the origin (0,0,0), was used to create the 2D content. Meanwhile, a stereo pair was placed 65 mm apart to replicate the human interocular distance, capturing the left and right views for immersive spatial video. All videos were rendered in 4K resolution at 60 fps. Sample frames from these videos are presented in Figure 2. To incorporate variations in object distance, five scenes (*AsgardianToy*, *AVP*, *CommaDotStudio*, *NewAtlantis*, and *UninvitedGuest*) were captured twice to represent both 'near' and 'far' perspectives. The 'near' perspective features objects positioned close to the camera, emphasizing fine spatial details and depth, while the 'far' perspective captures the same scenes with objects located farther away, offering a broader and more distant view of the environment. To maintain a consistent compression standard across formats, both 2D and spatial videos



(a) SpaceHome (b) AVP (c) RobotCamera (d) NewAtlantis

(e) DesertHunting (f) AsgardianToy (g) UninvitedGuest (h) LastMission

(i) Bumblebee (j) CommaDotStudio (k) Mercedes (l) TrainingScene

Figure 2: Sample frames of the created *SV-QoE* dataset.

were encoded using the open-source x265 encoder (version 4.1)[1]. The x265 encoder was chosen for its efficiency in high-quality video compression and its support for both standard 2D encoding and multiview coding, allowing MV-HEVC compatibility for spatial content. A Constant Rate Factor (CRF) quality control approach was employed during encoding. For spatial videos, three quality levels (*i.e.*, low, medium, high) were generated, while for 2D videos, the CRF values were selected to closely match the corresponding bitrates of these levels. High quality corresponds to a VMAF [39] score of 95, medium to 85, and low to approximately 75, with VMAF being the average of the left and right views.

4 Testing Procedure

We conducted a subjective test using the 5-point Absolute Category Rating (ACR) [40] methodology to evaluate the perceived quality of both 2D and spatial video content. In this test, participants viewed a series of 10-second video sequences and rated each on three distinct aspects: video quality, perceived depth, and overall quality using a five-point categorical scale. In this study, video quality refers to the fidelity of the videos (*e.g.*, compression artifacts), while overall quality captures the way the viewer experiences the content as a combination of video quality with perceived depth and immersion. Prior to the main test, participants underwent a training session designed to familiarize them with the rating procedure and the different video formats they would encounter. In addition, a Snellen visual acuity test was administered to ensure that all participants had normal vision. A total of 30 participants (11 females and 19 males; average age: 31 ± 6 years) took part in the study, and the entire testing session took, on average, 31 minutes to complete. Each participant rated each video six times: three times in spatial video encoded at three quality levels, and three times in 2D encoded at the corresponding bitrates of the spatial videos. To mitigate ordering effects, the presentation order of the videos was randomized for each participant.

[1]https://bitbucket.org/multicoreware/x265_git/src/Release_4.1/

5 Evaluation and Results

In this section, we analyze the results of the subjective test. First, we identify and remove any outliers from the collected data. Outlier detection was conducted using statistical methods, including the Z-score and Interquartile Range (IQR) [41], to identify extreme values that deviate significantly from the dataset's central distribution. As a result, three outlier responses were excluded from the analysis, and the subsequent results are based on data from 27 subjects.

5.1 Integrated perceptual quality analysis

Figure 3 shows the integrated results from the subjective test. In terms of video quality, both 2D and spatial formats were perceived similarly by the participants. However, depth perception showed a significant difference between the two modalities, with spatial content providing a notably enhanced depth experience. Additionally, overall quality ratings were higher for spatial videos compared to 2D videos. ANOVA [42] results further confirmed these findings, *i.e.*, video quality showed no significant difference ($p = 0.479$), while depth perception and overall quality exhibited significant differences ($p < 0.0001$ for both).

5.2 Perceptual quality analysis across different quality levels

Figure 4 shows that the perceived video quality at medium and high quality levels is similar for 2D and spatial videos encoded at the same bitrate. However, all other metrics exhibit statistically significant differences (see Table 2). This outcome may be attributed to the use of the default player in AVP, where the combination of

Table 2: ANOVA significance test results for different quality levels. A significance level of $\alpha = 0.05$ was used.

Quality level	Video quality	Depth perception	Overall quality
Low	Significant ($p = 1.90 \times 10^{-4}$)	Significant ($p = 1.23 \times 10^{-20}$)	Significant ($p = 1.10 \times 10^{-2}$)
Medium	Not Significant ($p = 1.60 \times 10^{-1}$)	Significant ($p = 3.69 \times 10^{-30}$)	Significant ($p = 5.12 \times 10^{-9}$)
High	Not Significant ($p = 2.45 \times 10^{-1}$)	Significant ($p = 1.27 \times 10^{-24}$)	Significant ($p = 7.06 \times 10^{-11}$)

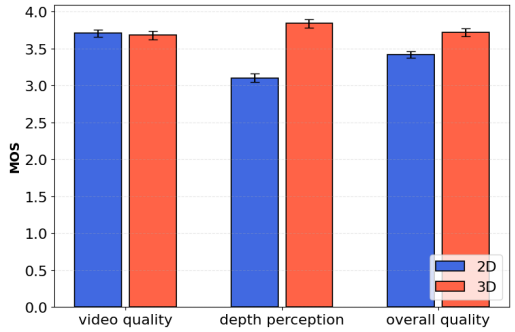


Figure 3: Integrated results: video quality is perceived similarly for both 2D and spatial formats, while depth perception is significantly enhanced in spatial content (3D), leading to higher overall quality ratings.

viewing distance [43] and player size might have made it difficult for users to notice artifacts at medium and high quality levels. This observation highlights the need for further investigation in future work.

Figure 5 shows sample frames from videos rendered at near and far object distances, while Figure 6 presents the average subjective ratings for five sequences, each captured with both near and far distances from the camera. The results consistently show that scenes featuring near object distances received significantly higher ratings in video quality, depth perception, and overall quality compared to their far-distance counterparts. When objects are positioned closer to the stereo cameras, the binocular disparity between the left and right views increases. This increased disparity enhances stereoscopic depth cues, making the 3D structure of the scene more salient and compelling for viewers. Consequently, the near-distance scenes produced a more immersive depth experience. These findings highlight the critical role of object proximity in shaping the viewing experience: as objects are rendered closer to the camera, spatial cues become more prominent, thereby enhancing perceived video quality, depth, and overall content appreciation. One potential explanation for the improved video quality is that closer objects occupy a larger portion of the screen and often reveal more visual detail, which may lead viewers to perceive the image as sharper or more vivid. However, this hypothesis warrants further investigation in future work.

5.3 Relationship between disparity and depth perception

To assess the strength of the relationship between stereoscopic disparity and depth perception (Q2), we first extracted disparity using the Stereo Semi-Global Block Matching (StereoSGBM) algorithm [44]. We then trained regression models on the full dataset and evaluated the Pearson Linear Correlation Coefficient (PLCC) and the coefficient of determination (R^2). These metrics respectively quantify the linear correlation and the proportion of variance in Q2 MOS explained by the average disparity. The results are presented in Table 3. All models exhibit a positive correlation, with performance steadily improving from simple linear models to more complex regressors. The highest accuracy is achieved by the Random Forest model, which yields a PLCC of 0.9184 and an R^2 of 0.8434, indicating a strong nonlinear relationship between disparity and depth perception.

Next, to evaluate out-of-sample predictive performance, we applied a leave-one-video-out cross-validation scheme and computed the Mean Absolute Error (MAE) for each model. MAE measures the average magnitude of prediction errors in the same units as Q2 MOS. Table 4 lists these MAE values, sorted from highest (worst) to lowest (best). The MAE results follow a consistent trend: as model complexity increases, prediction errors decrease. While linear models yield relatively high MAE values (*e.g.,* 0.3205 for linear regression), nonlinear models such as SVR (0.2411), KNN (0.2222), and Random Forest (0.2127) provide more accurate depth perception predictions.

5.4 Relationship between video quality and objective quality metrics

Similar to the previous section, we evaluate the strength of the relationship between widely used objective quality metrics—PSNR, SSIM [45], MS-SSIM [46], VMAF [39], and AVQT[2]—and subjective video quality (Q1). AVQT version 2 is specifically designed for spatial video content. For all other metrics, we computed the average scores between the left and right views.

We trained various regression models on the full dataset and computed PLCC and R^2. These metrics quantify the linear correlation and the proportion of variance in Q1 explained by each objective metric, respectively. The results are summarized in Table 5.

Among the metrics, AVQT exhibits the highest correlation with subjective video quality, particularly when modeled with nonlinear regressors. For instance, using Random Forest, AVQT achieves a PLCC of 0.9650 and an R^2 of 0.9231, outperforming all other metrics. VMAF also shows strong predictive power, especially with complex models like Random Forest (PLCC = 0.9575, R^2 = 0.9027). In contrast, traditional metrics such as PSNR, SSIM, and MS-SSIM show weaker correlations in simpler models, though their performance improves with increased model complexity. Notably, even with linear regression, AVQT (PLCC = 0.7302) and VMAF (PLCC = 0.6643) demonstrate stronger alignment with subjective quality than all other metrics.

[2]https://developer.apple.com/download/all/?q=avqt

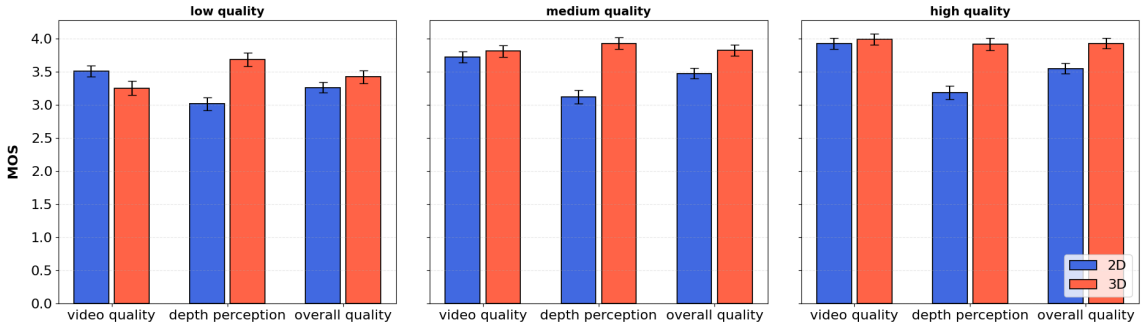
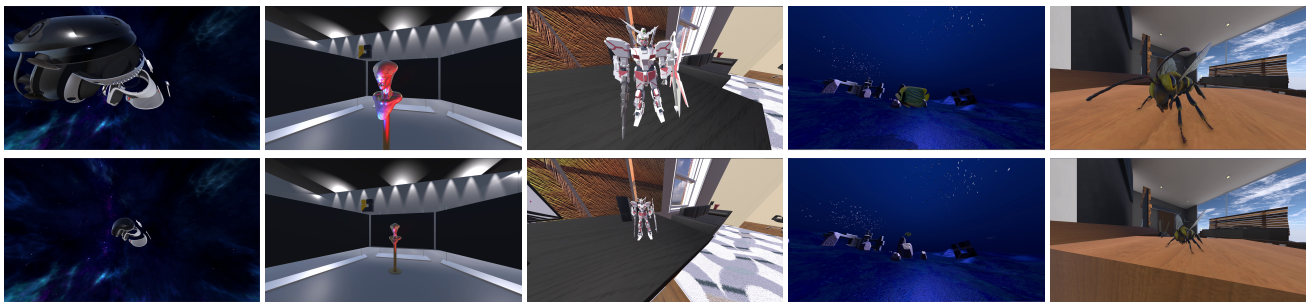


Figure 4: Results by quality level: at lower quality, 2D content exhibits higher video quality with similar overall quality, but immersive content provides enhanced depth perception. With increasing quality, video quality ratings converge, depth perception remains superior in immersive content, and overall QoE improves.



(a) AVP **(b) CommaDotStudio** **(c) AsgardianToy** **(d) NewAtlantis** **(e) UninvitedGuest**

Figure 5: Sample frames from videos rendered at near (upper row) and far (lower row) distances.

Table 3: PLCC and R^2 between disparity and depth perception.

Model	PLCC	R^2
Linear Regression	0.4788	0.2293
Polynomial Regression (Degree 2)	0.5533	0.3062
Polynomial Regression (Degree 3)	0.6921	0.4790
Support Vector Regressor (SVR)	0.7673	0.5810
K-Nearest Neighbors (KNN)	0.8648	0.7393
Random Forest	0.9184	0.8434

Table 4: Leave-One-Video-Out MAE Comparison.

Model	MAE
Linear Regression	0.3205
Polynomial Regression (Degree 2)	0.3121
Polynomial Regression (Degree 3)	0.2879
Support Vector Regressor (SVR)	0.2411
K-Nearest Neighbors (KNN)	0.2222
Random Forest	0.2127

Table 5: PLCC and R^2 between objective quality metrics and Q1 across regression models.

Model	**PSNR**		**SSIM**		**MS-SSIM**		**VMAF**		**AVQT**	
	PLCC	R^2	PLCC	R^2	PLCC	R^2	PLCC	R^2	PLCC	R^2
Linear Regression	0.4186	0.1753	0.4933	0.2434	0.3904	0.1524	0.6643	0.4413	**0.7302**	**0.5332**
Polynomial Regression (Degree 2)	0.4379	0.1917	0.5396	0.2912	0.4397	0.1934	0.7356	0.5412	**0.7968**	**0.6349**
Polynomial Regression (Degree 3)	0.4449	0.1980	0.5401	0.2917	0.4448	0.1979	0.7571	0.5732	**0.8034**	**0.6454**
Support Vector Regressor (SVR)	0.5778	0.3311	0.5634	0.3185	0.5067	0.2299	0.7587	0.5746	**0.8122**	**0.6492**
K-Nearest Neighbors (KNN)	0.5778	0.3311	0.5304	0.2793	0.5416	0.2891	0.7399	0.5445	**0.8284**	**0.6706**
Random Forest	0.9393	0.8573	0.9339	0.8278	0.9274	0.8263	0.9575	0.9027	**0.9650**	**0.9231**

We then evaluated generalization performance using leave-one-video-out cross-validation and reported the MAE for each model. As shown in Table 6, MAE provides a complementary view to PLCC and R^2, capturing prediction accuracy on unseen videos. The table highlights that AVQT consistently outperforms the other metrics across all regression models, achieving the lowest MAEs in every

Table 6: MAE between objective quality metrics and video quality (Q1) across regression models using leave-one-out cross-validation.

Model	PSNR	SSIM	MS-SSIM	VMAF	AVQT
Linear Regression	0.3475	0.3334	0.3521	0.3064	**0.2749**
Polynomial Regression (Degree 2)	0.3500	0.3249	0.3460	0.2921	**0.2462**
Polynomial Regression (Degree 3)	0.3591	0.3361	0.3493	0.2884	**0.2438**
Support Vector Regressor (SVR)	0.3484	0.3357	0.3528	0.3005	**0.2774**
K-Nearest Neighbors (KNN)	0.3743	0.3763	0.3773	0.3278	**0.2774**
Random Forest	0.4267	0.4385	0.3889	0.3626	**0.3042**

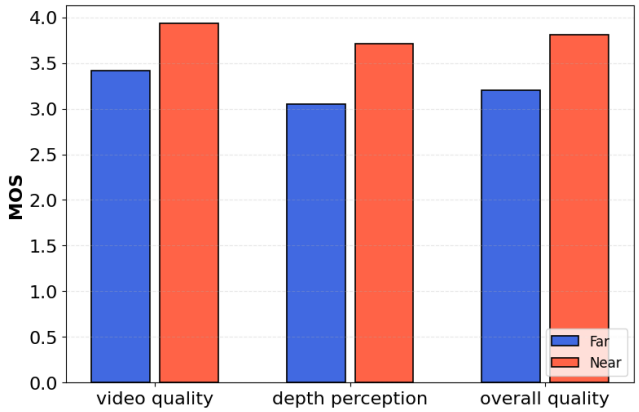


Figure 6: The average subjective results for five sequences captured from 'Near' and 'Far' distances.

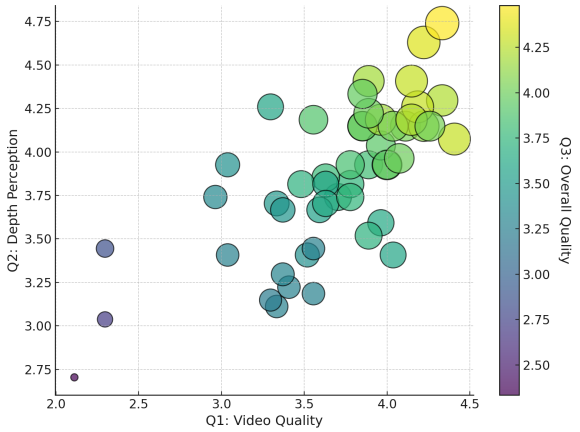


Figure 7: Scatter plot illustrating the relationship between video quality (Q1) and depth perception (Q2), with circle size and color representing overall quality (Q3). Larger and lighter circles indicate higher perceived overall quality.

case. For example, with Polynomial Regression (Degree 3), AVQT reaches an MAE of 0.2438, notably lower than VMAF (0.2884), MS-SSIM (0.3493), SSIM (0.3361), and PSNR (0.3591).

5.5 Relationship between video quality, depth perception, and overall quality

To explore how perceived video quality (Q1) and depth perception (Q2) jointly influence overall quality (Q3), we visualize their relationship in a scatter plot. As shown in Figure 7, Q1 is plotted on the x-axis, Q2 on the y-axis, and Q3 is encoded through both the color and size of each circle. The plot reveals a clear trend: higher values of Q1 and Q2 generally correspond to larger, lighter-colored circles, indicating higher overall quality (Q3). This visual pattern confirms that both video quality and depth perception contribute meaningfully to the overall viewing experience.

The relationship can be approximated using a linear model, expressed as:

$$Q3 = 0.4830 \cdot Q1 + 0.5234 \cdot Q2 - 0.0663. \quad (1)$$

This model explains 97.4% of the variance in Q3 ($R^2 = 0.974$), confirming that a linear combination of Q1 and Q2 effectively captures users' perception of overall quality. It also suggests that depth perception (Q2) has a slightly stronger influence on Q3 than video quality (Q1), though both factors are essential.

To further assess how well Q3 can be predicted using a combination of objective quality (AVQT) and depth-related cues (disparity), we trained several regression models and computed their PLCC and R^2. These results, summarized in Table 7, show that all models capture a meaningful relationship between these predictors and Q3. The Random Forest model yields the highest correlation (PLCC = 0.9822) and the greatest explained variance ($R^2 = 0.9532$), indicating its superior ability to model the nonlinear interactions between AVQT and disparity.

To evaluate how these models generalize to unseen content, we applied a leave-one-video-out cross-validation strategy and measured prediction accuracy using MAE. Table 8 presents these results. Again, Random Forest delivers the best performance with the lowest MAE of 0.2463, followed closely by KNN and SVR. These results confirm that leveraging both disparity and AVQT leads to robust predictions of overall quality across diverse video content.

Table 7: PLCC and R^2 for predicting Q3 using AVQT and disparity.

Model	PLCC	R^2
Linear Regression	0.6961	0.4845
Polynomial Regression (Degree 2)	0.7549	0.5699
Polynomial Regression (Degree 3)	0.8106	0.6571
Support Vector Regressor (SVR)	0.7792	0.5658
K-Nearest Neighbors (KNN)	0.8120	0.6479
Random Forest	0.9822	0.9532

Table 8: Leave-One-Video-Out MAE for predicting Q3 using AVQT and disparity.

Model	MAE
Linear Regression	0.2897
Polynomial Regression (Degree 2)	0.2732
Polynomial Regression (Degree 3)	0.3298
Support Vector Regressor (SVR)	0.2762
K-Nearest Neighbors (KNN)	0.2639
Random Forest	0.2463

6 Conclusion

This study provides a comprehensive evaluation of spatial video quality on HMDs, addressing critical gaps in immersive multimedia research. Our subjective assessment using the *SV-QoE* dataset –developed as part of this work– reveals that spatial videos consistently outperform their 2D counterparts in depth perception and overall quality, especially at higher bitrates The analysis further highlights the influence of viewing distance, with videos captured at closer ranges offering a notably enhanced QoE. Through a rigorous statistical analysis, we establish that both perceived video quality (Q1) and depth perception (Q2) are significant predictors of overall quality (Q3), with Q2 exerting a slightly stronger influence. These findings emphasize the importance of considering depth perception alongside traditional quality metrics when evaluating spatial content.

7 Acknowledgment

The financial support of the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development, and the Christian Doppler Research Association is gratefully acknowledged. Christian Doppler Laboratory ATHENA: https://athena.itec.aau.at/.

References

[1] J. Van Der Hooft, H. Amirpour, M. T. Vega, Y. Sanchez, R. Schatz, T. Schierl, and C. Timmerer, "A Tutorial on Immersive Video Delivery: From Omnidirectional Video to Holography," *IEEE Communications Surveys & Tutorials*, vol. 25, no. 2, pp. 1336–1375, 2023.

[2] W. Zhou, H. Amirpour, C. Timmerer, G. Zhai, P. L. Callet, and A. C. Bovik, "Perceptual Visual Quality Assessment: Principles, Methods, and Future Directions," 2025. Version Number: 1.

[3] J. J. Cummings, M. Tsay-Vogel, T. J. Cahill, and L. Zhang, "Effects of immersive storytelling on affective, cognitive, and associative empathy: The mediating role of presence," *New Media & Society*, vol. 24, pp. 2003–2026, Sept. 2022.

[4] J. Psotka, "Immersive training systems: Virtual reality and education and training," *Instructional Science*, vol. 23, pp. 405–431, Nov. 1995.

[5] S. C. Mallam, S. Nazir, and S. K. Renganayagalu, "Rethinking Maritime Education, Training, and Operations in the Digital Era: Applications for Emerging Immersive Technologies," *Journal of Marine Science and Engineering*, vol. 7, p. 428, Dec. 2019. Number: 12 Publisher: Multidisciplinary Digital Publishing Institute.

[6] S. C. Bronack, "The Role of Immersive Media in Online Education," *The Journal of Continuing Higher Education*, vol. 59, pp. 113–117, May 2011. Publisher: Routledge _eprint: https://doi.org/10.1080/07377363.2011.583186.

[7] N. Dodgson, "Autostereoscopic 3D displays," *Computer*, vol. 38, pp. 31–36, Aug. 2005.

[8] N. S. Holliman, N. A. Dodgson, G. E. Favalora, and L. Pockett, "Three-Dimensional Displays: A Review and Applications Analysis," *IEEE Transactions on Broadcasting*, vol. 57, pp. 362–371, June 2011.

[9] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, June 2012. ISSN: 1063-6919.

[10] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3061–3070, June 2015. ISSN: 1063-6919.

[11] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation," pp. 4040–4048, IEEE Computer Society, June 2016. ISSN: 1063-6919.

[12] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A Naturalistic Open Source Movie for Optical Flow Evaluation," in *Computer Vision – ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), (Berlin, Heidelberg), pp. 611–625, Springer, 2012.

[13] E. Cheng, P. Burton, J. Burton, A. Joseski, and I. Burnett, "RMIT3DV: Pre-announcement of a creative commons uncompressed HD 3D video database," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, pp. 212–217, July 2012.

[14] L. Goldmann, F. D. Simone, and T. Ebrahimi, "A comprehensive database and subjective evaluation methodology for quality of experience in stereoscopic video," in *Three-Dimensional Image Processing (3DIP) and Applications*, vol. 7526, pp. 242–252, SPIE, Feb. 2010.

[15] D. Corrigan, F. Pitié, V. Morris, A. Rankin, M. Linnane, G. Kearney, M. Gorzel, M. O'Dea, C. Lee, and A. Kokaram, "A Video Database for the Development of Stereo-3D Post-Production Algorithms," in *2010 Conference on Visual Media Production*, pp. 64–73, Nov. 2010.

[16] M. Urvoy, M. Barkowsky, R. Cousseau, Y. Koudota, V. Ricorde, P. Le Callet, J. Gutiérrez, and N. García, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *2012 Fourth International Workshop on Quality of Multimedia Experience*, pp. 109–114, July 2012.

[17] H. Imani, M. B. Islam, and L.-K. Wong, "A New Dataset and Transformer for Stereoscopic Video Super-Resolution," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 705–714, June 2022. ISSN: 2160-7516.

[18] M. H. Izadimehr, M. Ghanbari, G. Chen, W. Zhou, X. Hao, M. Dasari, C. Timmerer, and H. Amirpour, "SVD: Spatial Video Dataset," 2025. Version Number: 1.

[19] G. Tech, Y. Chen, K. Müller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the multiview and 3d extensions of high efficiency video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 35–49, 2016.

[20] G. Tech, Y. Chen, K. Muller, J.-R. Ohm, A. Vetro, and Y.-K. Wang, "Overview of the Multiview and 3D Extensions of High Efficiency Video Coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, pp. 35–49, Jan. 2016.

[21] Wei Zhou, Ning Liao, Zhibo Chen, and Weiping Li, "3D-HEVC visual quality assessment: Database and bitstream model," in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, (Lisbon, Portugal), pp. 1–6, IEEE, June 2016.

[22] R. Hussain, M. Chessa, and F. Solari, "Improving Depth Perception in Immersive Media Devices by Addressing Vergence-Accommodation Conflict," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, pp. 6334–6346, Sept. 2024. Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[23] J. Zhang, Q. Jia, Y. Liu, W. Zhang, W. Wei, and X. Tian, "SpatialMe: Stereo Video Conversion Using Depth-Warping and Blend-Inpainting," 2024. Version Number: 1.

[24] T. Guionnet, K. Jerbi, T. Burnichon, and M. Raulet, "MV-HEVC: How to optimize compression of immersive 3D content," in *Proceedings of the 3rd Mile-High Video Conference on zzz*, (Denver CO USA), pp. 87–87, ACM, Feb. 2024.

[25] W. Liu, J. Li, and Y. B. Cho, "A novel architecture for parallel multi-view HEVC decoder on mobile device," *EURASIP Journal on Image and Video Processing*, vol. 2017, p. 24, Dec. 2017.

[26] C. Timmerer, H. Amirpour, F. Tashtarian, S. Afzal, A. Rizk, M. Zink, and H. Hellwagner, "HTTP Adaptive Streaming: A Review on Current Advances and Future Challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications*, p. 3736306, May 2025.

[27] G. Chen, S. Wang, J. Chakareski, D. Koutsonikolas, and M. Dasari, "Spatial video streaming on apple vision pro xr headset," 2025.

[28] L. Goldmann, F. De Simone, and T. Ebrahimi, "Impact of acquisition distortion on the quality of stereoscopic images," in *Proceedings of the International Workshop on Video Processing and Quality Metrics for Consumer Electronics*, 2010.

[29] Z. Chen, W. Zhou, and W. Li, "Blind Stereoscopic Video Quality Assessment: From Depth Perception to Overall Experience," *IEEE Transactions on Image Processing*, vol. 27, pp. 721–734, Feb. 2018.

[30] W. Wan, D. Huang, B. Shang, S. Wei, H. R. Wu, J. Wu, and G. Shi, "Depth Perception Assessment of 3D Videos Based on Stereoscopic and Spatial Orientation Structural Features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, pp. 4588–4602, Sept. 2023.

[31] Y.-H. Lin and J.-L. Wu, "Quality Assessment of Stereoscopic 3D Image Compression by Binocular Integration Behaviors," *IEEE Transactions on Image Processing*, vol. 23, pp. 1527–1542, Apr. 2014.

[32] F. Qi, T. Jiang, S. Ma, and D. Zhao, "Quality of experience assessment for stereoscopic images," in *2012 IEEE International Symposium on Circuits and Systems*, (Seoul, Korea (South)), pp. 1712–1715, IEEE, May 2012.

[33] F. Battisti, M. Carli, P. Le Callet, and P. Paudyal, "Toward the assessment of quality of experience for asymmetric encoding in immersive media," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 392–406, 2018.

[34] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Processing: Image Communication*, vol. 28, pp. 1143–1155, Oct. 2013.

[35] J. Wang, A. Rehman, K. Zeng, S. Wang, and Z. Wang, "Quality Prediction of Asymmetrically Distorted Stereoscopic 3D Images," *IEEE Transactions on Image Processing*, vol. 24, pp. 3400–3414, Nov. 2015.

[36] J. Xu, Z. Luo, W. Zhou, W. Zhang, and Z. Chen, "Quality Assessment of Stereoscopic 360-degree Images from Multi-viewports," in *2019 Picture Coding Symposium (PCS)*, (Ningbo, China), pp. 1–5, IEEE, Nov. 2019.

[37] Z. Chen, J. Xu, C. Lin, and W. Zhou, "Stereoscopic Omnidirectional Image Quality Assessment Based on Predictive Coding Theory," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, pp. 103–117, Jan. 2020.

[38] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Full-Reference Stereoscopic Video Quality Assessment Using a Motion Sensitive HVS Model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, pp. 452–466, Feb. 2021.

[39] "VMAF: The Journey Continues. by Zhi Li, Christos Bampis | by Netflix Technology Blog | Netflix TechBlog."

[40] International Telecommunication Union, "Recommendation ITU-T P.910: Subjective video quality assessment methods for multimedia applications," Tech. Rep. P.910, International Telecommunication Union, 2008.

[41] A. S. Yaro, F. Maly, P. Prazak, and K. Malý, "Outlier Detection Performance of a Modified Z-Score Method in Time-Series RSS Observation With Hybrid Scale Estimators," *IEEE Access*, vol. 12, pp. 12785–12796, 2024.

[42] F. Z. H., "Calculation and Interpretation of Analysis of Variance and Covariance: By George W. Snedecor. Ames, Iowa: Collegiate Press, Inc. 105 pages. 1934. $1," *Agronomy Journal*, vol. 26, pp. 255–256, Mar. 1934.

[43] H. Amirpour, R. Schatz, C. Timmerer, and M. Ghanbari, "On the Impact of Viewing Distance on Perceived Video Quality," in *VCIP 2021*, pp. 1–5, IEEE, Dec. 2021.

[44] H. Hirschmuller, "Stereo Processing by Semiglobal Matching and Mutual Information," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, pp. 328–341, Feb. 2008. Publisher: Institute of Electrical and Electronics Engineers (IEEE).

[45] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Transactions on Image Processing*, vol. 13, pp. 600–612, Apr. 2004. Conference Name: IEEE Transactions on Image Processing.

[46] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale Structural Similarity for Image Quality Assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, (Pacific Grove, CA, USA), pp. 1398–1402, IEEE, 2003.