
# SVD: Spatial Video Dataset

MohammadHossein Izadimehr
University of Klagenfurt
Klagenfurt, Austria
seizadimehr@edu.aau.at

Milad Ghanbari
University of Klagenfurt
Klagenfurt, Austria
mighanbari@edu.aau.at

Guodong Chen
Northeastern University
Boston, USA
chen.guod@northeastern.edu

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

Xiaoshuai Hao
Beijing Academy of Artificial Intelligence
Beijing, China
xshao@baai.ac.cn

Mallesham Dasari
Northeastern University
Boston, USA
m.dasari@northeastern.edu

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

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

## Abstract

Stereoscopic video has long been the subject of research due to its ability to deliver immersive three-dimensional content to a wide range of applications. The dual-view format inherently provides binocular disparity cues that enhance depth perception and realism, making it indispensable for fields such as telepresence, 3D mapping, and robotic vision. Until recently, however, end-to-end pipelines for capturing, encoding, and viewing high-quality stereoscopic video were neither widely accessible nor optimized for consumer-grade devices. Today's smartphones, such as the iPhone Pro and modern Head-Mounted Displays (HMDs) like the Apple Vision Pro (AVP), offer built-in support for stereoscopic video capture, hardware-accelerated encoding, and seamless playback on devices like the AVP and Meta Quest 3, which require minimal user intervention. Apple refers to this streamlined workflow as spatial Video. Making the full stereoscopic video process available to everyone has made new applications possible. Despite these advances, there remains a notable absence of publicly available datasets that include the complete spatial video pipeline on consumer platforms, hindering reproducibility and comparative evaluation of emerging algorithms.

In this paper, we introduce SVD, a spatial video dataset comprising 300 five-second video sequences, i.e., 150 captured using an iPhone Pro and 150 with an AVP. Additionally, 10 longer videos with durations ranging from 2 min 29 s to 5 min have been recorded. The SVD dataset is publicly released to facilitate research in codec performance evaluation, subjective and objective Quality of Experience assessment, depth-based computer vision, stereoscopic video streaming, and other emerging 3D applications such as neural rendering and volumetric capture. Link to the dataset: https://cd-athena.github.io/SVD/.

## CCS Concepts

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

## Keywords

stereoscopic video, spatial video, dataset, HEVC, QoE

**ACM Reference Format:**
MohammadHossein Izadimehr, Milad Ghanbari, Guodong Chen, Wei Zhou, Xiaoshuai Hao, Mallesham Dasari, Christian Timmerer, and Hadi Amirpour. 2025. SVD: Spatial Video Dataset. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, Dublin, Ireland, 7 pages. https://doi.org/10.1145/3746027.3758246

## 1 Introduction

Immersive media technologies [1] are redefining how digital content is experienced by delivering more realistic and visually compelling representations of scenes. Advances in virtual [2, 3], augmented [4], and mixed reality [5] have driven the development of high-resolution HMD [6, 7], spatial audio integration, and improved stereoscopic rendering. These technologies enable engaging experiences across domains such as entertainment, education, and visual communication, where realism and a strong sense of presence are essential.

A key component of immersive media is stereoscopic video, which enhances realism by replicating the way human vision perceives depth through binocular disparity. In practice, this involves capturing two slightly offset views of a scene – one corresponding to the left eye and one to the right – using a dual-lens or two-camera rig that is carefully calibrated to maintain known baseline distance and optical parameters. During capture, precise synchronization and geometric calibration ensure that corresponding pixels in each view lie on the same epipolar line, facilitating accurate disparity estimation. During playback, specialized display technologies present each view to the appropriate eye. The human visual system then


*MM '25, Dublin, Ireland*

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3758246

**Table 1: Overview of stereoscopic video datasets.**

| Dataset Name | Year | Resolution | Description |
|---|---|---|---|
| KITTI Stereo 2012 [8] | 2012 | 1226×370 | Outdoor driving scenes |
| KITTI Stereo 2015 [9] | 2015 | 1242×375 | Outdoor driving scenes with dynamic scenes with objects |
| SceneFlow [10] | 2016 | 960×540 | Synthetic stereo sequences |
| MPI-Sintel [11] | 2012 | 1024×436 (24fps) | Synthetic scenes with complex motion and visual effects |
| RMIT3DV HD 3D Video [12] | 2012 | 1920×1080 (25fps) | Thirty-one diverse urban scenes |
| EPFL MMSPG 3DVQA [13] | 2010 | 1920×1080 (25fp) | Six high-quality visual variations |
| Stereo Video Database [14] | 2010 | 1920×1080 (25fps) | Stereo cinema post-production |
| NAMAD3D [15] | 2012 | 1920×1080(25fps) | Natural 3D scenes with twin-lens camera |
| SVSR-Set [16] | 2022 | 1920×1080 (30fps) | Indoor/outdoor with varied motion and lighting |
| SVD (Ours) | 2025 | 1920×1080 (30fps) | Indoor and outdoor, captured with iPhone Pro |
| | 2025 | 2200×2200 (30fps) | Indoor and outdoor, captured with AVP |

fuses these two images, leveraging small interocular differences to reconstruct a coherent depth map and evoke a convincing sense of three-dimensional space [1].

Despite its clear benefits for depth perception, stereoscopic video production has historically been constrained by increased capture complexity, the need for rigorous calibration, higher data rates to accommodate dual streams, and display hardware requirements that have, until recently, limited its adoption in consumer and broadcast contexts.

Recently, this barrier has been significantly lowered through consumer devices that support native stereoscopic video workflows. Smartphones such as the iPhone Pro now offer built-in dual-camera setups for spatial video capture, while headsets like the AVP and Meta Quest 3 provide native playback support. These devices also include hardware-accelerated encoding, enabling efficient compression using state-of-the-art codecs like High Efficency Video Coding (HEVC). Apple has introduced the term *spatial video* to describe this tightly integrated pipeline from capture to playback, which allows users to create and experience 3D content with minimal technical effort.

While there are many well-established 2D video datasets [17–19], the availability of high-quality stereoscopic video datasets has remained limited. This scarcity is largely due to the challenges associated with stereo video capture, the lack of accessible stereoscopic displays, and the need for optimized stereo video encoders. However, with recent advances in capture technologies and wider availability of immersive displays, these barriers have significantly diminished. To drive research in stereoscopic video processing, we introduce the Spatial Video Dataset (SVD) as a comprehensive collection of high-quality stereoscopic video clips captured using the latest iPhone 16 Pro and AVP devices. The dataset comprises 150 short 5 s videos from each device totalling to 300 spatial video sequences, along with 10 long-form sequences ranging from 2 min 29 s to 5 min captured with both, covering a diverse range of indoor and outdoor environments, varied motion dynamics, and unique capture scenarios. SVD is specifically designed to support a broad spectrum of applications, including stereoscopic image and video coding, streaming, Quality of Experience (QoE) assessment, and stereoscopic image and video quality evaluation, providing researchers with a powerful resource for advancing immersive media technologies.

## 2 Related Work

In this section, we introduce relevant existing stereoscopic video datasets from the literature. The KITTI Stereo 2012 dataset [8] serves as a key benchmark for stereo vision in autonomous driving. It contains stereo videos of road scenes captured from a calibrated pair of cameras mounted on a car. It includes 194 training and 195 test scenes with resolutions of 1226×370, captured in outdoor environments with high-resolution stereo cameras. The KITTI Stereo 2015 dataset [9] builds upon its predecessor by adding 200 training and 200 test scenes with a resolution of 1242×375 in dynamic environments with moving objects, enhancing its relevance for real-world driving scenarios.

SceneFlow [10] provides a dataset containing synthetic stereoscopic videos with a resolution of 960×540. The RMIT3DV HD 3D Video database [12] is a comprehensive dataset designed to represent diverse content and visual conditions for various research applications. It comprises 31 stereoscopic video sequences filmed across multiple locations at RMIT University and Melbourne CBD, with durations ranging from 17 s to 2.5 min. All videos are recorded using a stereoscopic camera (Panasonic AG-3DA1) in 1920×1080 resolution with 10-bit YUV 4:2:2 encoding at 25 fps, ensuring high visual fidelity and uncompressed quality. This dataset is particularly valuable for studies involving stereo video quality assessment, disparity estimation, and 3D visual analysis, providing high-resolution, uncompressed stereoscopic content for reliable experimental evaluation.

The MPI-Sintel dataset [11], derived from the open-source animated film Sintel, is a widely used benchmark originally developed for optical flow evaluation, but also highly relevant for stereoscopic research. It includes stereo video sequences rendered at a resolution of 1024×436 with rich visual effects such as motion blur, specular reflections, and atmospheric conditions, closely mimicking real-world scenes. Despite being synthetic, its image and motion statistics align well with those of natural videos, making it a credible proxy for stereo vision tasks. With dense ground truth, multiple rendering passes, and long sequences, MPI-Sintel provides a flexible and reproducible resource for benchmarking stereo matching, disparity estimation, and depth-aware video analysis.

The EPFL MMSPG HD 3D Video Database (3DVQA) [13] comprises six stereoscopic video scenes, each lasting 10 s and capturing a variety of colors, textures, motion, and depth variations. Recorded with a resolution of 1920×1080 at 25 fps, the videos are stored in

AVCHD format and compressed with MPEG-4 Advanced Video Coding (AVC) at 24 Mbps. Despite its compression, 3DVQA preserves high visual quality, making it an ideal resource for stereo video quality assessment, disparity estimation, and depth-aware encoding techniques. Its controlled yet diverse visual content supports reliable benchmarking in 3D video quality evaluation and computer vision research.

The Stereo Video Database [14] is specifically designed as a test resource for research and development in stereo cinema post-production. It features a diverse collection of sequences shot in both indoor and outdoor environments under controlled and uncontrolled lighting conditions, capturing various real-world scenarios. The footage includes both steadicam and tripod-based shots, providing different levels of motion dynamics. The experimental setup employs a dual-camera rig with two Iconix HD-RH1 cameras mounted on an Inition 'bolt' side-by-side rig. Data is recorded using Flash XDR units in 4:2:2 XDCAM format with the xd5e codec at a bitrate of 100 Mbps. All sequences are captured in 1920×1080 resolution at 25 fps, ensuring high-quality stereoscopic content suitable for post-production analysis and stereo depth processing.

In NAMAD3D [15], the sequences were captured using a Panasonic AG-3DAIE twin-lens camera, which features two synchronized lenses with a 60 mm separation, closely matching the human interpupillary distance for natural-looking 3D content. The sequences are recorded in 1920×1080 at 25 fps. When feasible, uncompressed dual SDI streams were sent to a Clearview Extreme system for high-quality recording, applied to sequences like Barrier gate, Hall, News report, Phone call, Soccer, Tree branches, and Umbrella. In cases where streaming to Clearview was impractical, like for Basket, Boxer, and Lab, the content was saved directly onto SD cards in AVC High-Profile format at a maximum bitrate of 24 Mbps (average 21 Mbps).

The SVSR-Set [16] dataset consists of 71 stereo videos captured with a ZED 2 stereoscopic camera. Videos are recorded in 1920×1080 resolution at 30 fps for a duration of 20 s and are available in SVO and AVI formats, respectively. The dataset includes a wide range of indoor and outdoor settings, with variations in motion levels and illumination conditions. To ensure accuracy, the camera was subjected to a detailed calibration process to correct potential shifts in its internal parts. The calibration file, generated once and reused for all recordings, contains the exact locations of the left and right cameras and their optical properties.

## 3 Spatial Video Dataset

In this section, we introduce our SVD, which contains stereoscopic video sequences captured using both the iPhone Pro and the AVP. We recorded a diverse subset of spatial video sequences with each device, covering a variety of indoor and outdoor scenarios to ensure content variability across lighting conditions, environments, and motion characteristics. Specifically, we captured 150 short video clips of 5 s each, along with 10 longer sequences, tailored to streaming-oriented use cases. A grid of the first frames of 35 randomly selected videos recorded with the iPhone Pro setup is shown in Fig. 1, providing a visual summary of the diversity within our dataset.

We begin by detailing the camera configurations and recording capabilities of the two devices, highlighting their roles in enabling high-quality spatial video capture without the need for external calibration or rigs. We then describe the set of low-level features extracted from the dataset, including spatial and temporal complexity, colorfulness, and luminance statistics, providing a quantitative characterization of the visual content.

### 3.1 Camera Configurations

Stereoscopic video capture has traditionally required complex and carefully calibrated hardware configurations [20, 21]. Conventional stereoscopic rigs often employ two physically separate cameras mounted on a rail or a rigid rig. These configurations typically required manual alignment, synchronization, and post-processing to ensure temporal and geometric consistency between the two video streams. Moreover, ensuring perfect lens matching, exposure control, and white balance between the cameras was necessary to avoid visual discomfort or depth perception errors during playback.

While effective in controlled studio environments, these traditional systems were bulky, expensive, and impractical for casual or mobile content capture. Their complexity created a barrier to the broader adoption of stereoscopic video, particularly among non-professional users.

Recent innovations in consumer electronics have dramatically simplified stereoscopic video capture. Modern devices such as the iPhone Pro and the AVP integrate dual-camera systems and advanced computational photography pipelines that enable spatial video recording without the need for external rigs or manual calibration.

*3.1.1 iPhone Pro Camera System.* The iPhone Pro features native spatial video recording using its precisely calibrated wide and ultra-wide rear cameras, which are spaced 19.2 mm apart to produce depth cues suitable for small screens and HMDs. Apple's spatial video system integrates real-time depth estimation, optical stabilization, and synchronized exposure control to ensure high-quality stereo capture. Videos are recorded in 1080p at 30 fps in standard dynamic range (SDR) and encoded in HEVC with stereoscopic metadata, enabling seamless playback on devices like the AVP. For spatial video capture on the iPhone Pro, the concept of "hero eye" refers to the primary camera (the wide (1x) lens) that records the main view. This lens provides a higher-quality image, while the ultra-wide (0.5x) lens captures a secondary view that is cropped and scaled to match the primary perspective.

*3.1.2 Apple Vision Pro Camera System.* The AVP represents a significant advancement in immersive media, offering both playback and recording capabilities for spatial video. Equipped with a stereoscopic 3D main camera system featuring 18 mm lenses with an $f/2.0$ aperture, the AVP captures spatial videos at a resolution of $2200 \times 2200$ per eye at 30 fps in SDR. Spatial videos on the AVP are encoded using the Multiview High-Efficiency Video Coding (MV-HEVC) format. This format stores stereoscopic views in separate layers (*i.e.*, one for each eye) within a single video file, accompanied by spatial metadata that enables immersive playback experiences. Table 2 compares the spatial video recording capabilities of the iPhone Pro and the AVP.

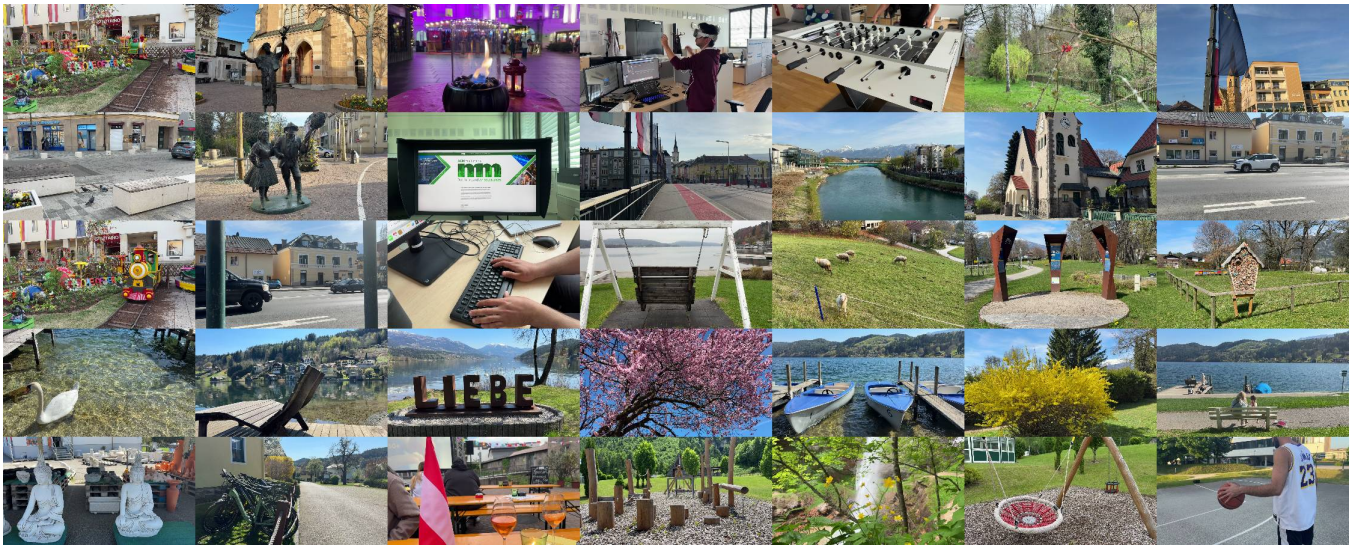



**Figure 1: Grid of the first frames from 35 randomly selected spatial videos recorded using the iPhone Pro.**

**Table 2: Spatial Video Recording: iPhone 16 Pro vs. AVP.**

| Feature | iPhone 16 Pro | Apple Vision Pro |
|---|---|---|
| Resolution & Frame Rate | 1920×1080 px @ 30fps (SDR) | 2200×2200 px @ 30fps (SDR) |
| Video Format | MV-HEVC | MV-HEVC |
| Horizontal Field of View (FOV) | 63.4° | 71.6° |
| Baseline (Interaxial Distance) | 19.2mm | 63.8mm |
| Hero Eye Concept | Yes ("hero eye" from Wide camera); | No (both eye streams equal quality) |
| Recording Orientation Requirement | Landscape | - |

## 3.2 Low-Level Video Features

For each video, we extract a comprehensive set of low-level features on a per-frame basis and include them alongside the original video in our released dataset. These features are widely used in video analysis and objective quality assessment, and cover spatial, temporal, stereo-view, and perceptual dimensions.

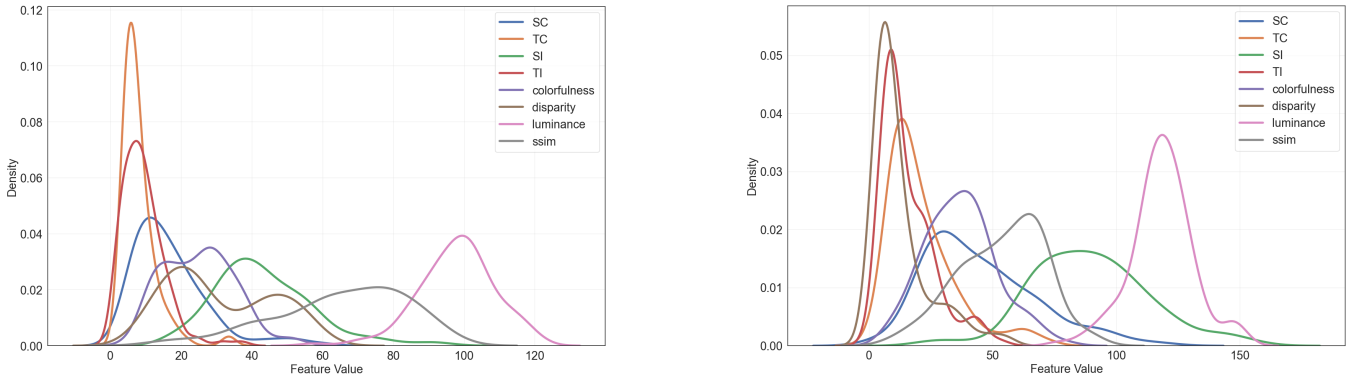
*3.2.1 Spatial Complexity.* Spatial complexity is a fundamental aspect of video content that significantly impacts both perceptual quality and compression efficiency. Scenes with high spatial detail, such as textures, edges, and fine patterns, are more challenging to compress without introducing visible artifacts, while simpler, smoother areas are easier to encode efficiently. For this reason, spatial complexity is widely used in video quality assessment and adaptive encoding strategies.

In our dataset, we quantify spatial complexity [22] using two complementary features: *(i)* Spatial Information (SI) and *(ii)* Spatial Complexity (SC). SI measures edge strength by applying a Sobel filter to each frame, capturing local contrast and sharpness. SI is standardized in ITU-T P.910 and is strongly correlated with spatial complexity [23]. SC, on the other hand, is derived from the Enhanced Video Complexity Analyzer (EVCA) framework [24] and operates in the Discrete Cosine Transformation (DCT) domain, capturing frequency-based spatial variation and block-level detail. It computes the spatial complexity by applying a weighted sum to the DCT coefficients of each block, where higher-frequency components are given greater emphasis to reflect the contribution of fine textures and detailed patterns within the frame.

*3.2.2 Temporal Complexity.* Temporal complexity reflects the amount of motion and dynamic change within a video, which significantly affects both perceived quality and compression performance. Videos with fast-moving objects, frequent cuts, or high activity between frames typically demand more resources for encoding and are more susceptible to motion-related artifacts.

To capture temporal complexity in our dataset, we use two complementary metrics: *(i)* Temporal Information (TI) and **(ii)** Temporal Complexity (TC). TI is defined in ITU-T P.910 and computed as the standard deviation of pixel-wise differences between consecutive frames, providing a frame-level measure of motion intensity. Higher TI values indicate stronger temporal variation, which is critical for tasks like motion-aware encoding, frame rate control, and adaptive streaming.

In addition to TI, we include TC, a motion-sensitive feature introduced in the EVCA framework [24]. Unlike TI, which operates in the pixel domain, TC is calculated in the DCT domain by computing the Sum of Absolute Differences (SAD) between the weighted DCT coefficients of corresponding blocks across consecutive frames. This weighting scheme emphasizes high-frequency components and, thus, captures subtle motion details and structural changes




**(a) AVP Features.** **(b) iPhone Features.**

**Figure 2: Comparison of feature distributions from AVP and iPhone sequences.**

more effectively. TC has been shown to correlate more strongly with perceptual temporal complexity than earlier pixel-domain metrics [24].

*3.2.3 Colorfulness.* Colorfulness is a perceptual attribute that reflects the intensity and diversity of colors within a video frame. It plays an important role in visual quality perception, content classification, and aesthetic evaluation. Videos with rich and varied colors tend to be perceived as more vivid and engaging, while those with dull or limited color ranges may appear flat or less appealing. In our dataset, we include a colorfulness metric introduced in [25], which combines the mean and standard deviation of red-green and yellow-blue color differences. The colorfulness feature is computed for each frame of both left and right views, allowing for the analysis of color consistency across stereo pairs.

*3.2.4 Luminance.* In addition to spatial, temporal, and color features, we also include luminance-based metrics to capture the overall brightness and contrast characteristics of each frame. Specifically, we compute the mean and variance of the luminance (Y) channel for both left and right views. These features provide insight into lighting conditions, exposure balance, and perceptual contrast within the video, which can influence both encoding efficiency and visual quality perception.

*3.2.5 Disparity.* Disparity refers to the horizontal offset between corresponding points in the left and right views of a stereoscopic video, and it provides a key cue for depth perception. To capture disparity information in our dataset, we compute dense disparity maps for each video frame using the Stereo Semi-Global Block Matching (StereoSGBM) algorithm [26], as implemented in OpenCV. This method balances local accuracy with global smoothness by aggregating matching costs along multiple paths, making it suitable for high-resolution stereo content.

*3.2.6 SSIM.* In addition to disparity, we compute the inter-view Structural Similarity Index (SSIM) [27] between the left and right views of each video frame to assess their perceptual correspondence. SSIM is a widely used image quality metric that evaluates luminance, contrast, and structural similarity, providing a more perceptually relevant comparison than pixel-wise differences. In our context, it serves as a complementary feature of disparity, offering a view-independent measure of stereo consistency. High SSIM values indicate strong structural alignment between the views, while lower values may signal mismatches, occlusions, or inconsistencies in stereo rendering.

Fig. 2 shows the *distribution* of the extracted low-level features for videos captured with the AVP and iPhone Pro devices using Kernel Density Estimation (KDE) plots. All features, including spatial complexity, temporal complexity, luminance, disparity, and SSIM (scaled by 100 for better visual representation), are computed on a per-frame basis for both the left and right views. The values are then averaged across all frames to produce a single representative feature vector per sequence. These KDE plots highlight the differences in content characteristics and capture profiles between the two devices, offering insights into the diversity and quality of the dataset.

Fig. 3 shows the Pearson *correlation* between corresponding low-level features extracted from the left and right views, along with the average SSIM values, for both the AVP and iPhone Pro devices. This analysis provides insight into the consistency of stereo content captured by each device. The results indicate that the AVP exhibits stronger correlations between views across most features, as well as higher SSIM scores, suggesting more consistent stereo alignment and better structural similarity. This highlights the superior stereo capture quality of the AVP compared to the iPhone Pro in our dataset. The lower consistency observed in the iPhone Pro recordings may be attributed to the "hero eye" concept, where the wide (1x) camera serves as the primary view and the ultra-wide (0.5x) camera is cropped and aligned post-capture. This asymmetric processing can introduce disparities in quality and content between the two views.

## 4 Potential Applications

The rich set of features and high-quality stereoscopic content included in our dataset enables a wide range of research and development applications across multimedia, computer vision, and immersive media domains. In the following, we outline several key areas where this dataset can be effectively leveraged.

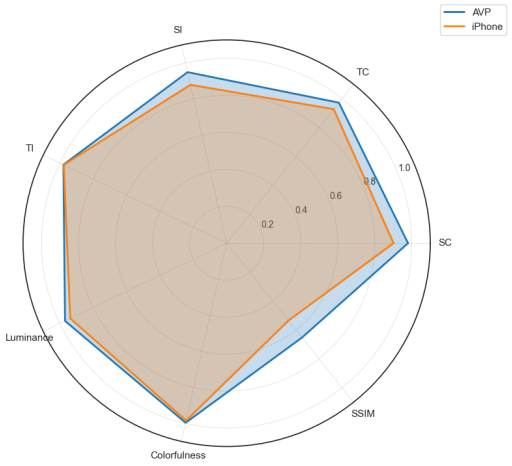



**Figure 3: Correlation of low-level features and average SSIM between left and right views for spatial videos recorded with the AVP and iPhone Pro.**

## 4.1 Codec Development and Comparison

This dataset serves as a practical benchmark for codec development and evaluation, particularly for stereoscopic and multiview content. Earlier standards, such as Multiview Video Coding (MVC) in AVC [28] and MV-HEVC in HEVC [29], introduced inter-view prediction to improve compression efficiency for stereo video. More recently, Apple adopted MV-HEVC for its Spatial Video format, and as of version 4.1, the x265 encoder added support for MV-HEVC, enabling optimized stereoscopic encoding within its efficient compression framework.

With rich diversity, including spatial and temporal complexities, colorfulness, luminance, disparity, and SSIM, our dataset allows for comprehensive codec comparisons in terms of rate-distortion performance, view consistency, and encoding speed. It also supports the evaluation of fast encoding algorithms and learning-based strategies for content-adaptive compression.

## 4.2 Monoscopic-to-Stereoscopic Video Conversion

Our dataset can be used to train and evaluate models that convert monoscopic (2D) videos into stereoscopic (3D) formats, which is an increasingly important task for supplying immersive content in AR/VR applications [30, 31]. As an inherently ill-posed problem, stereo conversion has evolved significantly with deep learning, progressing from early convolutional approaches to advanced diffusion-based models. These methods typically generate the right view from the left by estimating monocular depth and compensating for occluded regions through inpainting or generative synthesis. However, they often suffer from artifacts and lack control over structural accuracy. By offering high-quality stereo pairs, dense disparity maps, and perceptual similarity metrics such as SSIM, our dataset provides strong supervision and validation tools for improving the realism, consistency, and fidelity of stereoscopic view synthesis.

## 4.3 Video Quality Assessment

Our dataset is well-suited for conducting subjective quality assessments of stereoscopic video, thanks to its diversity in various features. This variability enables controlled experiments that evaluate how different content characteristics influence human perception of 3D video quality under various viewing conditions, including HMDs and stereoscopic monitors [32]. The results of such subjective studies can be used to develop and validate both full-reference and no-reference video quality metrics tailored for stereoscopic content [33].

## 4.4 Video Streaming

The longer video sequences in our dataset make it particularly suitable for streaming applications, enabling realistic evaluations of adaptive delivery strategies over time [34, 35]. These clips support research in content-aware bitrate ladder construction [36], where spatial, temporal, and disparity features can inform optimal quality tiers for stereoscopic video. The dataset also facilitates per-title encoding [37–41], allowing encoding parameters to be tailored to individual content characteristics for improved compression efficiency and visual quality. Furthermore, it enables studies on QoE in 3D streaming, including the effects of bitrate fluctuations, depth artifacts, and inter-view inconsistencies. By combining objective features with potential subjective evaluations, the dataset offers a comprehensive foundation for developing and testing adaptive streaming algorithms for stereoscopic and immersive video services.

# 5 Conclusions

We presented SVD, a publicly available spatial video dataset designed to support a broad range of research in stereoscopic and immersive media technologies. Captured using consumer-grade devices (iPhone Pro and AVP), the dataset includes both short and long-form high-quality stereoscopic video sequences, covering a wide range of real-world scenes. Alongside the raw videos, we provide a rich set of low-level features including spatial and temporal complexity, colorfulness, luminance, disparity, and inter-view SSIM, enabling in-depth analysis across multiple application domains.

SVD is specifically tailored for tasks such as codec development and benchmarking, monoscopic-to-stereoscopic video synthesis, video quality assessment (both subjective and objective), and adaptive streaming. Its inclusion of diverse content types, extended sequence durations, and per-frame metrics makes it an ideal resource for training, evaluating, and comparing algorithms in both traditional and emerging 3D video processing tasks. The dataset is available at https://cd-athena.github.io/SVD/.

# 6 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] I. Wohlgenannt, A. Simons, and S. Stieglitz, "Virtual Reality," *Business & Information Systems Engineering*, vol. 62, pp. 455–461, Oct. 2020.

[3] C. Anthes, R. J. García-Hernández, M. Wiedemann, and D. Kranzlmüller, "State of the art of virtual reality technology," in *2016 IEEE Aerospace Conference*, pp. 1–19, Mar. 2016.

[4] J. Carmigniani, B. Furht, M. Anisetti, P. Ceravolo, E. Damiani, and M. Ivkovic, "Augmented reality technologies, systems and applications," *Multimedia Tools and Applications*, vol. 51, pp. 341–377, Jan. 2011.

[5] M. Speicher, B. D. Hall, and M. Nebeling, "What is Mixed Reality?," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, (New York, NY, USA), pp. 1–15, Association for Computing Machinery, May 2019.

[6] B. Dunphy, G. Young, G. Dinan, and N. Murray, "Integrating Head Mounted Displays into Live Broadcasting Workflows: Implications and Possibilities from an Industry Perspective," in *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences Workshops*, (Stockholm Sweden), pp. 131–136, ACM, June 2024.

[7] R. Cheng, N. Wu, M. Varvello, E. Chai, S. Chen, and B. Han, "A First Look at Immersive Telepresence on Apple Vision Pro," in *Proceedings of the 2024 ACM on Internet Measurement Conference*, (Madrid Spain), pp. 555–562, ACM, Nov. 2024.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, "YouTube-8M: A Large-Scale Video Classification Benchmark," Sept. 2016. arXiv:1609.08675 [cs].

[18] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC Dataset for Video Compression Research," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–5, Sept. 2019. ISSN: 2473-3628.

[19] H. Amirpour, V. V. Menon, S. Afzal, M. Ghanbari, and C. Timmerer, "VCD: Video Complexity Dataset," in *Proceedings of the 13th ACM Multimedia Systems Conference*, (Athlone Ireland), pp. 234–239, ACM, June 2022.

[20] D. Lee and I. Kweon, "A novel stereo camera system by a biprism," *IEEE Transactions on Robotics and Automation*, vol. 16, pp. 528–541, Oct. 2000.

[21] S. Tzavidas and A. Katsaggelos, "A multicamera setup for generating stereo panoramic video," *IEEE Transactions on Multimedia*, vol. 7, pp. 880–890, Oct. 2005.

[22] H. Amirpour, K. Shoeffmann, M. Ghanbari, and C. Timmerer, "DeepVCA: Deep Video Complexity Analzer," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2024.

[23] International Telecommunication Union (ITU-T), "P.910: Subjective video quality assessment methods for multimedia applications," Recommendation P.910, International Telecommunication Union – Telecommunication Standardization Sector, 2021.

[24] H. Amirpour, M. Ghasempour, L. Qu, W. Hamidouche, and C. Timmerer, "EVCA: Enhanced Video Complexity Analyzer," in *ACM MMSys 2024*, MMSys '24, pp. 285–291, Apr. 2024.

[25] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video*. Boston, MA: Springer US, 2002.

[26] 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).

[27] 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.

[28] A. Vetro, T. Wiegand, and G. J. Sullivan, "Overview of the Stereo and Multiview Video Coding Extensions of the H.264/MPEG-4 AVC Standard," *Proceedings of the IEEE*, vol. 99, pp. 626–642, Apr. 2011.

[29] 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.

[30] 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.

[31] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381, June 2024. ISSN: 2575-7075.

[32] 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.

[33] 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.

[34] G. Chen, S. Wang, J. Chakareski, D. Koutsonikolas, and M. Dasari, "Spatial Video Streaming on Apple Vision Pro XR Headset," in *Proceedings of the 26th International Workshop on Mobile Computing Systems and Applications*, HotMobile '25, (New York, NY, USA), pp. 115–120, Association for Computing Machinery, Feb. 2025.

[35] 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.

[36] V. V. Menon, H. Amirpour, M. Ghanbari, and C. Timmerer, "OPTE: Online Per-Title Encoding for Live Video Streaming," in *ICASSP 2022*, pp. 1865–1869, May 2022. ISSN: 2379-190X.

[37] H. Amirpour, C. Timmerer, and M. Ghanbari, "PSTR: Per-Title Encoding Using Spatio-Temporal Resolutions," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, July 2021.

[38] H. Amirpour, M. Ghanbari, and C. Timmerer, "DeepStream: Video Streaming Enhancements using Compressed Deep Neural Networks," *Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[39] A. Telili, W. Hamidouche, H. Amirpour, S. A. Fezza, C. Timmerer, and L. Morin, "Convex Hull Prediction Methods for Bitrate Ladder Construction: Design, Evaluation, and Comparison," *ACM Transactions on Multimedia Computing, Communications, and Applications*, p. 3723006, Mar. 2025.

[40] V. V. Menon, H. Amirpour, M. Ghanbari, and C. Timmerer, "Perceptually-Aware Per-Title Encoding for Adaptive Video Streaming," in *ICME 2022*, pp. 1–6, IEEE, July 2022.

[41] V. V. Menon, H. Amirpour, M. Ghanbari, and C. Timmerer, "CODA: Content-aware Frame Dropping Algorithm for High Frame-rate Video Streaming," in *2022 Data Compression Conference (DCC)*, pp. 475–475, IEEE, Mar. 2022.