arXiv:2403.00604v2 [eess.IV] 7 Mar 2024

ODVISTA: AN OMNIDIRECTIONAL VIDEO DATASET FOR SUPER-RESOLUTION AND QUALITY ENHANCEMENT TASKS

Ahmed Telili[1], *Ibrahim Farhat*[1], *Wassim Hamidouche*[1], *Hadi Amirpour*[2]

[1] Technology Innovation Institute P.O.Box: 9639, Masdar City, Abu Dhabi, UAE
[2]Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria

ABSTRACT

Omnidirectional or 360-degree video is being increasingly deployed, largely due to the latest advancements in immersive virtual reality (VR) and extended reality (XR) technology. However, the adoption of these videos in streaming encounters challenges related to bandwidth and latency, particularly in mobility conditions such as with unmanned aerial vehicles (UAVs). Adaptive resolution and compression aim to preserve quality while maintaining low latency under these constraints, yet downscaling and encoding can still degrade quality and introduce artifacts. Machine learning (ML)-based super-resolution (SR) and quality enhancement techniques offer a promising solution by enhancing detail recovery and reducing compression artifacts. However, current publicly available 360-degree video SR datasets lack compression artifacts, which limit research in this field. To bridge this gap, this paper introduces omnidirectional video streaming dataset (ODVista), which comprises 200 high-resolution and high-quality videos downscaled and encoded at four bitrate ranges using the high-efficiency video coding (HEVC)/H.265 standard. Evaluations show that the dataset not only features a wide variety of scenes but also spans different levels of content complexity, which is crucial for robust solutions that perform well in real-world scenarios and generalize across diverse visual environments. Additionally, we evaluate the performance, considering both quality enhancement and runtime, of two handcrafted and two ML-based SR models on the validation and testing sets of ODVista.
Dataset URL: https://github.com/Omnidirectional-video-group/ODVista

***Index Terms*—** Omnidirectional video, 360-degree video, super-resolution, video streaming, virtual reality, machine learning.

1. INTRODUCTION

Advancements in immersive video technologies have paved the way for users to engage in a virtual environment that mirrors real-life scenarios, thereby enhancing user engagement and a sense of belonging in a digital space. Various visual media formats, including omnidirectional video (ODV), volumetric videos, and light fields, are common and effective methods for facilitating an immersive viewing experience. In particular, ODV, also known as 360-degree video, has gained widespread popularity due to the availability of acquisition and display devices, as well as standardization efforts to ensure interoperability. However, the adoption of these videos in streaming encounters challenges related to bandwidth and latency, particularly in mobility conditions such as unmanned aerial vehicles (UAVs). Adaptive resolution and compression aim to maintain quality while minimizing latency in such scenarios. Nevertheless, the process of downscaling and encoding may result in quality degradation and the introduction of compression artifacts.

On the other hand, studies have demonstrated significant advancements in image and video super-resolution (SR) tasks through the adoption of deep learning-based methods. These methods, particularly those utilizing convolution neural network (CNN) [1], vision transformer (ViT) [2, 3], generative adversarial network (GAN) [4], and recurrent neural network (RNN) [5], have pushed the boundaries of what can be achieved in terms of image clarity and detail enhancement. The deep learning models are trained on vast datasets of low-resolution and high-resolution image pairs, enabling them to learn complex mappings between the two. Additionally, in the typical ODV video streaming pipeline, as illustrated in Fig. 1, SR can be integrated to enable effective upsampling of videos from lower resolutions to higher resolutions. In some cases, the original video can be intentionally downscaled to a lower resolution to preserve bandwidth. At the receiving end, the SR algorithm is applied to upscale the video back to a superior resolution. This process significantly enhances the viewing experience by providing higher resolution video without the need for increased bandwidth for direct high-resolution video streaming.

In the literature, there are datasets and training methodologies that enable SR models to produce high-resolution content with notable accuracy. However, the shortage of high-quality ODV video datasets limits the advancement of the accuracy of different SR models. Several proposals have introduced diverse datasets featuring distinct properties. Table 1 summarizes existing datasets with different characteristics. In the 2023 NTIRE challenge on 360° SR, Cao *et al.* [6] introduced a significant video dataset named *ODV360*. This dataset contains high-resolution (2K) 360° video content, featuring a total of 210 videos. The collection includes 90 videos from YouTube [7] and existing public 360° video datasets, alongside 120 videos directly recorded with Insta 360° cameras. In [8], Xu *et al.* proposed a dataset of 48 ODV sequences, each showcasing a wide variety of content that allows for categorization based on the video content. These sequences have been sourced from YouTube and other public domains under a free-use license. Subsequently, the original videos were edited to create short clips, with lengths ranging from 20 to 60 seconds. The video resolutions range from 3K (2880×1440) up to 8K (7680×3840), ensuring a broad range of details. Although this dataset is proposed for subjective quality assessment purposes, it can be very useful for SR tasks due to its variety and high-quality content. Li *et al.* in [9] proposed a dataset that features 600 ODV sequences, including 60 high-quality reference sequences with diverse content, sourced from raw formats and YouTube's virtual reality (VR) channel at bitrates exceeding 15 Mbps. These reference sequences span various content categories like nature, shows, and sports, with resolutions ranging from 4K to 8K. Additionally, these sequences are edited to lengths of 10 to 23 seconds at frame rates of 24-30 frames per second (fps) and are organized into 10 groups to enhance diversity and facilitate subjective analysis, ensuring varied resolution and content across groups.

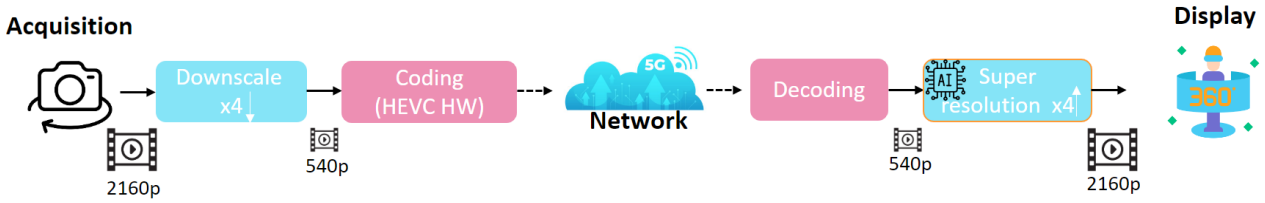


Fig. 1: Super-resolution integration in a typical streaming pipeline.

Table 1: Summary of existing 360-degree video datasets.

Database	Year	#Count	#Total	Resolutions	#Frames	Distortion type	Standard (Encoder)
ODV360 [6]	2023	210	630	2K	100	scaling (×2, ×4)	✗
VQA dataset [8]	2017	48	48	3K, 8K	600-1800	✗	✗
VQA-ODV [9]	2018	600	600	4K	240-690	✗	✗
ODVista (ours)	2024	200	1600	2K, 4K	100	scaling (×2, ×4) & compression	HEVC [10] (hevc_nvenc)

#Count: Total number of unique contents. #Total: Total number of video sequences, including reference and distorted videos.

While these datasets exhibit commendable variety across multiple dimensions, including content diversity, resolution, frame rate, and, in certain cases, a broad spectrum of bitrates, they fall short in a crucial area: compression distortion. This aspect plays a crucial role in developing SR models for streaming applications, as it introduces a loss of quality resulting from the video compression process. This distortion adds an extra layer of degradation, secondary to downscaling, affecting the visual fidelity of content in practical scenarios such as live streaming and video-on-demand (VoD) services. Addressing this challenge is essential for enhancing SR model performance in real-world streaming environments. The absence of the compression factor in existing datasets limits the ability to fully improve upon how SR algorithms can adapt to the artifacts introduced by compression. In this study, we propose omnidirectional video streaming dataset (ODVista), a comprehensive ODV dataset designed specifically to address SR challenges in the context of video streaming. The main contributions of this paper can be summarized as follows:

- Introducing ODVista, a novel dataset (Tables 1) featuring a diverse array of scenes with both scaling (2× and 4×) and compression distortions, aimed at facilitating the development of advanced SR models for streaming scenarios.
- Adopting a balanced sampling strategy based on spatial and temporal complexity, to ensure a balanced distribution and reduces outliers, enhancing model robustness.
- Propose a novel evaluation metric that can effectively capture the trade-off performance of SR techniques, specifically in balancing the enhancement of quality and runtime processing.
- Evaluating the efficacy of various SR approaches, including conventional and machine learning (ML)-driven methods, establishing a benchmark for future research.

The rest of the paper is organized as follows. Section 2 presents the proposed dataset with video collection and characteristics details. Section 3 evaluates SR algorithms, both traditional and ML-based, on the *ODVista* dataset to showcase its effectiveness across different baseline models. Finally, Section 4 concludes the paper.

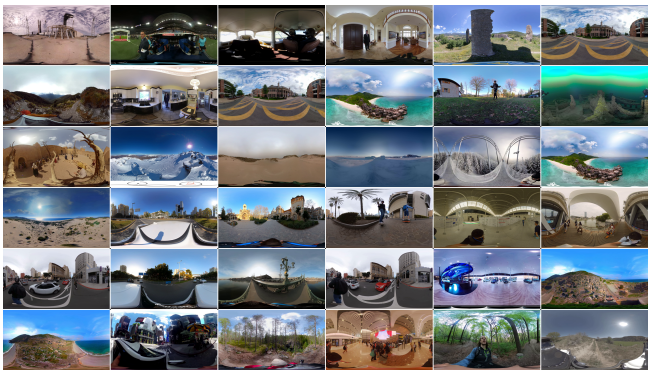
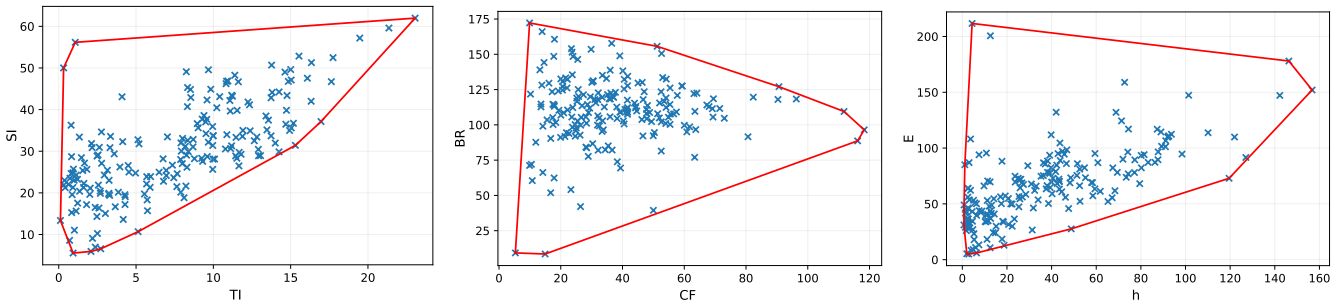


Fig. 2: Sample frames of the proposed ODVista dataset.

2. ODVISTA DATASET

2.1. Video collection

In the realm of VR and extended reality (XR) research, a notable gap has been identified: the limited availability of diverse high-quality ODV datasets that are both comprehensive and open to the academic community. To address this limitation, we meticulously collected the ODVista dataset, which includes 200 high-quality omnidirectional videos. These videos are equally divided into 100 videos with 2K (1080p) resolution and 100 videos with 4K (2160p) resolution, carefully collected from YouTube [7] and the ODV360 dataset [6], ensuring the inclusion of only high-quality sequences. All videos are licensed under creative commons attribution (CC) for academic and research purposes. To ensure homogeneity, all sequences were subjected to a scene segmentation process, guaranteeing that each sequence comprised only a single scene. This segmentation was accomplished using the PySceneDetect tool [11]. As a final step, all sequences were temporally cropped to 100 frames, if not originally formatted as such, resulting in a dataset that offers a comprehensive view of different scenes and scenarios. All these video sequences are stored in equirectangular projection (ERP) format. The proposed dataset includes a variety of indoor and outdoor scenes, as well as dynamic sports content. This variety is crucial to simulate realistic scenarios. Sample frames from the ODVista dataset are shown in



(a) SI vs. TI (b) BR vs. CF (b) h vs. E

Fig. 3: Source content distribution in paired feature space with corresponding convex hulls. Left column: TI versus SI, middle column: CF versus BR and right column: h versus E.

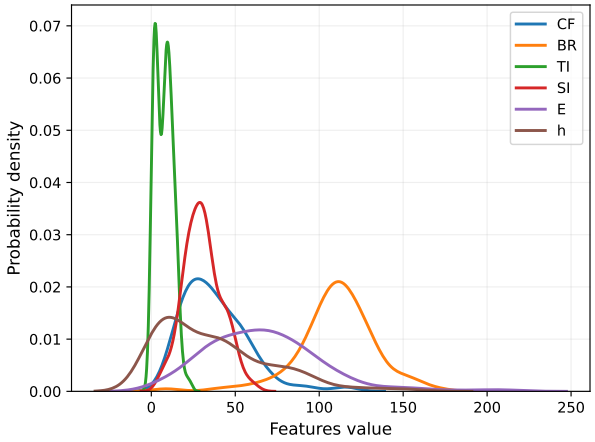


Fig. 4: Feature distribution comparisons among the proposed dataset.

Fig. 2, highlighting this diversity.

2.2. Dataset characterization

As a means of characterizing the content diversity of the videos in databases, Winkler *et al.* [12] initially proposed three video descriptors: spatial activity, temporal activity, and colorfulness. In our work, to achieve a more comprehensive analysis of content diversity, we expanded the set of these descriptors to include six low-level features, namely spatial information (SI), temporal information (TI), brightness (BR), colorfulness (CF), and two features derived from video complexity analyzer (VCA) [13], which are spatial complexity (E) and temporal complexity (h). Each of these features is calculated separately for each frame in the dataset. Subsequently, the computed values are averaged to obtain an overall (mean) representation. Scatter plots with convex hulls of paired features, illustrating the feature coverage of the proposed database, are shown in Fig. 3. Moreover, the fitted kernel distribution of each selected feature is illustrated in Fig. 4. Firstly, we observe that video sequences cover a wide range in the spatiotemporal domain, with values ranging from 5 to 62 for SI and 0 to 24 for TI, highlighting the diversity of the dataset. Furthermore, it is evident that the proposed dataset exhibits an extensive range of spatiotemporal complexities. The majority of the videos have low complexity as they contain only one scene, while a smaller portion consists of very complex sequences. Additionally, the scatter plot comparing BR to CF reveals a diverse range of content types in the dataset, with sequences ranging from 10 to 174 in BR and from 0 to 120 in CF. The range of BR values suggests the presence of various lighting conditions and scenes, spanning a variety of indoor and outdoor scenes, while the range of CF values indicates a variety of color palettes and visual styles.

2.3. Dataset processing

To construct a robust SR dataset for ODV streaming scenarios, we employ two essential processing techniques: downscaling and compression.

Downscaling. To generate the necessary low resolution (LR) and high resolution (HR) video pairs for SR tasks, all sequences undergo downsizing using a Lanczos-3 filter [14], as implemented by FFmpeg [15], at two different scales specifically, $\alpha = 2$ and $\alpha = 4$. This enables two distinct tracks for SR: $2\times$ and $4\times$.

Compression. Following the downscaling process, LR sequences undergo compression. We utilize a hardware-based implementation of the high-efficiency video coding (HEVC)/H.265 standard, embeded on an NVIDIA RTX A2000 8GB graphics card. The encoding process is carried out in random access (RA) at four distinct low bitrates (0.25 Mbps, 0.5 Mbps, 1 Mbps, and 2 Mbps) to accurately simulate bandwidth constraints encountered in mobility scenarios, such as UAVs. NVIDIA's implementation (NVENC [16]) offers various presets for live streaming environments. Our evaluations determined that the "low latency high quality" (llhq) preset provides the best compromise, ensuring high-quality output while meeting the real-time constraints critical to streaming applications in energy-aware devices and dynamic bandwidth environments. Consequently, 8 different bitstreams are encoded for each source content, resulting in a total of 1600 encoded video sequences.

Data splitting. In contrast to the conventional approach of random splitting employed by the majority of datasets, we divide our data into distinct sets, 80% for training, 10% for validation, and 10% for testing, using the stratified sampling technique. This stratification relies on a K-means clustering approach, focusing on spatial and temporal complexity features (E and h), specifically the mean and average across frames. The optimal number of clusters (k) is determined using the Elbow method [17]. This splitting strategy ensures a balanced distribution, effectively minimizing outliers, and

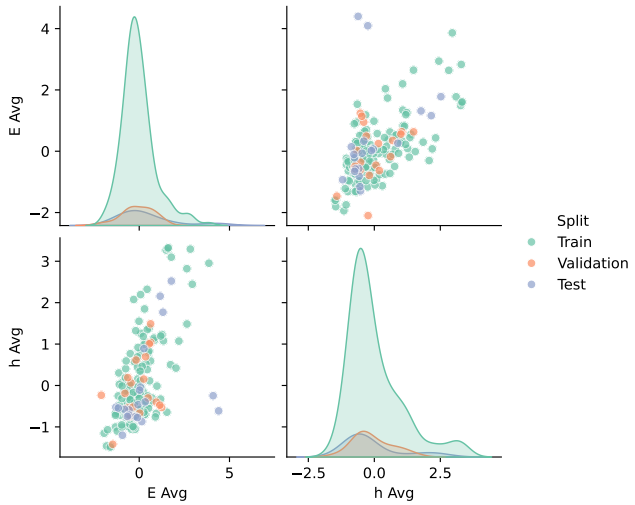


Fig. 5: Distributions of spatial complexity and temporal complexity across train, validation and test splits.

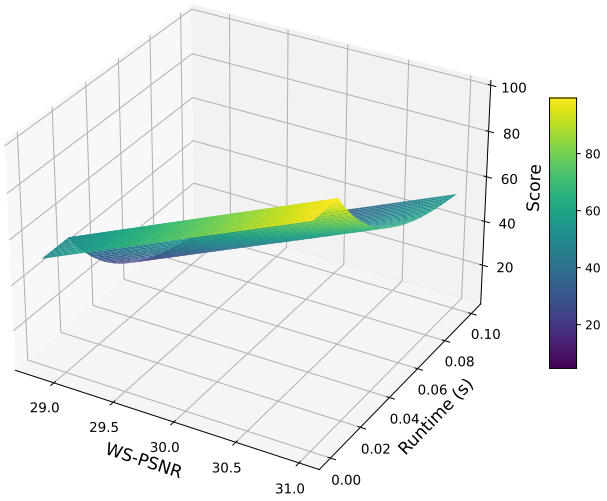


Fig. 6: 3D Visualization of score metric Q variation with WS-PSNR (dB) and runtime (s), with β = 0.5, WS-PSNR$_{min}$ = 28.8 dB and WS-PSNR$_{max}$ = 31 dB.

enables the reliable and robust development of SR methods. Fig. 5 illustrates the distribution of spatial and temporal complexity across different splits, indicating well-balanced partitions.

3. BENCHMARK

In this section, we evaluate the performance of various SR techniques, including both conventional and ML-based methods, on the proposed ODVista dataset.

3.1. Baselines

3.1.1. Conventional methods

Bicubic interpolation [18]. Bicubic interpolation is a commonly used method for image scaling in conventional SR techniques. It calculates the values of new pixels by applying a weighted average to the 16 surrounding pixels within a 4×4 neighborhood.

Lanczos filter [14]. Lanczos filter is an advanced conventional SR method. It uses a sinc-based kernel, called Lanczos kernel, for estimating pixel values. Unlike bicubic interpolation, which uses a 4×4 pixel grid, the Lanczos filter can take into account a larger number of surrounding pixels, with the exact number depending on the kernel size (e.g., Lanczos-3, Lanczos-4).

In our benchmark, we used the implementation provided by OpenCV [19] for both conventional methods. Specifically, for the Lanczos filter, we employ a kernel size of 4.

3.1.2. AI-based methods

FSRCNN [20]. Fast super-resolution convolutional neural network (FSRCNN) is a CNN-based method, evolved from SRCNN [21], designed for real-time SR applications. It features a more compact architecture that directly processes LR inputs, thereby reducing computational complexity. FSRCNN's notable performance is attributed to the use of deconvolution layers towards the end of the network, enabling it to enlarge the image size and reduce processing time in a single step.

SwinIR [22]. SwinIR, short for Image Restoration using Swin Transformer, is a transformer-based model designed for tasks such as SR, image denoising, and compression artifact reduction. Derived from the Swin Transformer architecture [23], SwinIR processes images at various scales through a unique shifted windowing scheme for self-attention. This design allows SwinIR to detect and interpret complex image patterns and dependencies over long distances more efficiently than traditional convolutional methods.

3.2. Evaluation metrics

To assess the performance of the baseline models on the proposed dataset, we evaluate both the quality of the upscaled video and the complexity of the SR process. Quality assessment is carried out using the objective quality metrics specifically designed for 360-degree videos, namely weighted spherical peak signal-to-noise ratio (WS-PSNR) and weighted spherical structural similarity index measure (WS-SSIM). The complexity of the model performing SR is measured by the inference time on a PC fitted with an Intel® Xeon 8280 CPU @ 2.70GHz × 56, 128GB RAM, and a 48GB VRAM NVIDIA RTX 6000 Ada graphics card. In particular, the runtime of ML-based models is estimated on the graphics processing unit (GPU) device. In order to evaluate the trade-off between quality enhancement and runtime, we propose a new scoring metric that takes into account both quality enhancement and runtime, as follows:

$$Q = (\beta \times \hat{Q} + (1 - \beta) \times C) \times 100, \tag{1}$$

where β is a weighting parameter (set to 0.5 in our evaluation), $\hat{Q}$ is the normalized WS-PSNR score, and C is the runtime evaluation score. The normalized quality score $\hat{Q}$ is computed as:

$$\hat{Q} = \frac{\text{WS-PSNR} - \text{WS-PSNR}_{min}}{\text{WS-PSNR}_{max} - \text{WS-PSNR}_{min}}, \tag{2}$$

where WS-PSNR$_{\text{min}}$ represents minimum value reached by the least performing model (i.e., Bicubic) and WS-PSNR$_{\text{max}}$ defines the the-

Table 2: Performance comparison of evaluated super-resolution methods.

Scale	Baseline	Validation set		Test set		Runtime/ 2k frame (s) ↓	Runtime/ 4k frame (s) ↓	Q ↑
		WS-PSNR (dB) ↑	WS-SSIM ↑	WS-PSNR (dB) ↑	WS-SSIM ↑			
2×	SwinIR✠	29.664	0.8437	29.761	0.8250	1.5232	7.5360	27.29
	FSRCNN✠	29.113	0.8321	29.280	0.8149	0.0015	0.0009	66.51
	Bicubic♣	28.829	0.8060	28.743	0.8117	✗	✗	✗
	Lanczos 4♣	28.880	0.8064	28.797	0.8110	✗	✗	✗
4×	SwinIR✠	28.811	0.8313	29.065	0.8099	0.4458	1.5155	29.79
	FSRCNN✠	28.018	0.8107	28.317	0.7912	0.0013	0.0015	61.10
	Bicubic♣	27.585	0.7982	27.790	0.7831	✗	✗	✗
	Lanczos 4♣	27.860	0.7995	27.795	0.7814	✗	✗	✗

✠ ML-based SR methods, ♣ Handcrafted-based SR methods.

oretical maximum WS-PSNR values we consider in our evaluation (30 dB and 31 dB for $\alpha = 4$ and $\alpha = 2$, respectively). The runtime evaluation metric assigns a full score to models that achieve a processing time of 0.016 seconds or less per 2K frame, as this speed enables a smooth 60 fps, essential for high-quality live streaming. To reinforce this standard, we apply penalties in our evaluation criteria for runtimes that exceed 0.016 seconds. The slower the model, the larger the penalties as follows:

$$C = \begin{cases} 1 & \text{runtime } \leq 0.016, \\ e^{B\times(0.016-\text{runtime})} & \text{otherwise, with } B = 30. \end{cases} \quad (3)$$

Fig. 6 illustrates the variation of the score Q based on runtime and WS-PSNR. It is evident that the highest scores are achieved by models operating in real time (60 fps) while delivering maximum WS-PSNR. Then, the score is penalized when there is an increase in runtime beyond real-time or when the model does not significantly outperform the quality of the baseline model.

3.3. Results and analysis

In this section, we assess the performance of the four baseline models on the validation and test sets of our proposed dataset, ODVista. Table 2 presents the results of the four baseline models introduced in Section 3, considering WS-PSNR, WS-SSIM, runtime, and Q score. It is evident that ML-based models significantly enhance the quality of the output videos. Specifically, there is an improvement of 0.78 dB and 0.23 dB in terms of WS-PSNR for the SwinIR and FSRCNN models, respectively, compared to the best-performing handcrafted model, Lanczos 4, in the 2× scaling configuration. This improvement is even more pronounced, particularly for the SwinIR model, achieving a 0.95 dB higher WS-PSNR compared to Lanczos 4 in the 4× scaling configuration. These results are corroborated by the WS-SSIM metric.

However, the quality improvements brought by the SwinIR model come at the expense of higher complexity, requiring an average of 0.44 seconds to process one 2K resolution frame and 1.51 seconds for a 4K frame in the 4× scale, falling short of maintaining real-time processing. This latency is even higher (1.52 seconds to process one 2K resolution and 7.53 seconds for a 4K resolution) in the 2× scaling configuration, mainly caused by the higher resolution of the input video compared to the 4× scale. On the contrary, the FSRCNN model exhibits a noteworthy balance between enhancing video quality and runtime efficiency. This model demonstrates the capacity to uphold real-time processing, surpassing 30 fps in both scaling configurations, even at a 4K video resolution. The proposed metric, denoted as Q, underscores the superiority of the FSRCNN model, securing the top rank (best trade-off between quality enhancement and runtime), and is subsequently trailed by the SwinIR model.

4. CONCLUSION

In this paper, we introduce the ODVista dataset, designed to address compression and scaling distortions in ODV. The proposed dataset comprises 200 high-quality and high-resolution videos. Each video has been scaled by two distinct factors and encoded across four different low-bitrate ranges to accurately simulate real-world scenarios characterized by limited bandwidth conditions. The ODVista dataset is characterized by its diversity, incorporating a range of indoor and outdoor scenes, covering various visual contents and distortion levels, making it well-suited for developing robust SR models. Additionally, we utilized a stratified sampling technique to ensure balanced training, validation, and test sets, improving the representativeness of the dataset and facilitating efficient model training and evaluation. Furthermore, we provide a comprehensive benchmark to evaluate both conventional and ML-based SR methods on the proposed dataset, enhancing its utility for the research community. As future work, we plan to expand the proposed dataset by including other videos captured using the Insta360 Pro 2 camera, further enriching the dataset. Furthermore, we aim to compress the dataset using additional video codecs that represent different video coding standards and formats, such as versatile video coding (VVC)/H.266 and AV1.

5. REFERENCES

[1] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Learning a deep convolutional network for image super-resolution," in *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, Eds., Cham, 2014, pp. 184–199, Springer International Publishing.

[2] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," in *2021 IEEE/CVF International Con-*

ference on Computer Vision Workshops (ICCVW), 2021, pp. 1833–1844.

[3] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong, "Activating more pixels in image super-resolution transformer," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 22367–22377.

[4] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi, "Photo-realistic single image super-resolution using a generative adversarial network," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 105–114.

[5] So Sasatani, Yutaro Iwamoto, and Yen–Wei Chen, "Frame attention recurrent back-projection network for accurate video super-resolution," in *2022 IEEE International Conference on Consumer Electronics (ICCE)*, 2022, pp. 01–05.

[6] Mingdeng Cao, Chong Mou, Fanghua Yu, Xintao Wang, Yinqiang Zheng, Jian Zhang, Chao Dong, Gen Li, Ying Shan, Radu Timofte, et al., "Ntire 2023 challenge on 360deg omnidirectional image and video super-resolution: Datasets, methods and results," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1731–1745.

[7] "YouTube," https://www.youtube.com/, 2024-01-30.

[8] Mai Xu, Chen Li, Yufan Liu, Xin Deng, and Jiaxin Lu, "A subjective visual quality assessment method of panoramic videos," in *2017 IEEE International Conference on Multimedia and Expo (ICME)*, 2017, pp. 517–522.

[9] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang, "Bridge the gap between vqa and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proceedings of the 26th ACM International Conference on Multimedia*, New York, NY, USA, 2018, MM '18, pp. 932–940, ACM.

[10] Gary J. Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[11] "Intelligent scene cut detection and video splitting tool," https://pyscenedetect.readthedocs.io/en/latest/, 2024-01-30.

[12] Stefan Winkler, "Analysis of Public Image and Video Databases for Quality Assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 6, pp. 616–625, 2012.

[13] Vignesh V Menon, Christian Feldmann, Hadi Amirpour, Mohammad Ghanbari, and Christian Timmerer, "VCA: Video Complexity Analyzer," in *Proceedings of the 13th ACM Multimedia Systems Conference*, 2022, pp. 259–264.

[14] Claude E Duchon, "Lanczos Filtering in One and Two Dimensions," *Journal of Applied Meteorology and Climatology*, vol. 18, no. 8, pp. 1016–1022, 1979.

[15] "FFmpeg," https://www.ffmpeg.org/, 2024-01-30.

[16] "Nvidia video codec sdk," https://developer.nvidia.com/video-codec-sdk, 2024-01-30.

[17] Trupti M Kodinariya, Prashant R Makwana, et al., "Review on determining number of cluster in k-means clustering," *International Journal*, vol. 1, no. 6, pp. 90–95, 2013.

[18] Robert Keys, "Cubic convolution interpolation for digital image processing," *IEEE transactions on acoustics, speech, and signal processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[19] "openCV," https://opencv.org/, 2024-01-30.

[20] Chao Dong, Chen Change Loy, and Xiaoou Tang, "Accelerating the super-resolution convolutional neural network," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 391–407.

[21] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang, "Image super-resolution using deep convolutional networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 2, pp. 295–307, 2015.

[22] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 1833–1844.

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10012–10022.