INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC 1/SC 29/WG 4
MPEG VIDEO CODING

ISO/IEC JTC 1/SC 29/WG 4 m64721
October 2023, Hannover, Germany

Title: [INVR]EE2.1-Related: Report with New Natural INVR Video Contents: SKKU_VRroom
Source: Sungkyunkwan University (SKKU)
Authors: Jaeyeol Choi, Yeongil Ryu, Yihyun Choi, Jong-Beom Jeong, JunHyeong Park, Issac Yang, Eun-Seok Ryu

Abstract

This document presents two natural contents from different camera array. Both sequences capture scene centered around a person moving while wearing a head mounted display (HMD), consisting of 30 videos; One has camera arranged in 1-dimensional converging array, while the other is set in a parallel 3x10 array. Additionally, we validated the two contents by rendering non-trained viewpoints by MIV VWS and NeRF variants.

1 Introduction

For INVR, there are growing demands for natural datasets to compare the performance between various models. The new datasets captured at SKKU are introduced in this document. "SKKU_VRroom1D" and "SKKU_VRroom2D" both consist of 300 frames of 30 synchronized videos. The datasets enable the evaluation of the dynamic NeRF models' performance in effectively addressing the four challenges: the non-Lambertian effect, object movement, depth order, and high frequency details. Table 1 provides detailed scene configurations for these elements.

Table 1. The main components of the scene for effective measurement

Challenges	Components	Details
(1) non-Lambertian effect		A mirror is positioned within the scene. The reflection in the mirror varies in each view. Around the view intended for test set, moving person can also be observed.
(2) Object movement		A person wearing an HMD device moves his head and hands and walk back and forth in the scene. This allows if artifact occur when representing a dynamic scene.
(3) Depth order		From the cameras' perspective, each object is at different distance. This allows to access how well depth information is rendered during 3D reconstruction.
(4) High frequency texture		Flowers, composed of three different colors with intricate textures, are placed in a vase filled with water. The vase and the doll are used to determine how well model represents high-frequency components.

2 Datasets

In this section, we provide a detailed description of the camera system used for acquisition, video time synchronization and the camera calibration process.

The camera model used is the Intel RealSense L515. This camera produced 24-bit RGB raw video file with a resolution of 1920x1080 (a film aspect ratio of 1.78:1). While there are minor variations between individual devices, the focal length of the camera is 1.88mm, with a field of view (FOV) of 69° horizontally and 42° vertically.

30 cameras are allocated six per a computer, referred to as an edge device. The model used for edge device is the Intel NUC 13 Extreme Kit. Among the five edge devices connected to cameras, one serves as a server. At the start of recording, the server sends a start message to the clients along with time information. As each system has time discrepancies, post-processing was done based on the server's global time stamp to compensate for these differences and achieve time synchronization. Subsequently, a video editing software was used to finely adjust the color temperature and saturation.

Position and rotation of cameras were estimated using colmap[1]. The estimated camera parameters are provided in the attachments of this document as 'poses_bounds.npy', 'transforms.json' and MIV camera parameter format is also provided.

2.1 "SKKU_VRroom1D" Content

The "SKKU_VRroom1d" content was captured with 30 cameras arranged in an elliptical formation, all positioned at the same level of height. As shown in the left image of Figure 1, all cameras are pointed towards the center of the scene. Table 2 provides a basic description of the content. The right image of Figure 1 depicts the actual capturing environment. It describes the spacing between the cameras, the height of the cameras from the ground, the distance from the scene, and camera numbers.

Table 2. The description of "SKKU_VRroom1D " content

Item	Specification
Number of frames	300
Number of views	30 (v01~v30)
Recommended test view	v11
Format	mp4 (H.264 codec)
Resolution	1920 x 1080
FPS	30.0
Total size	373 MB

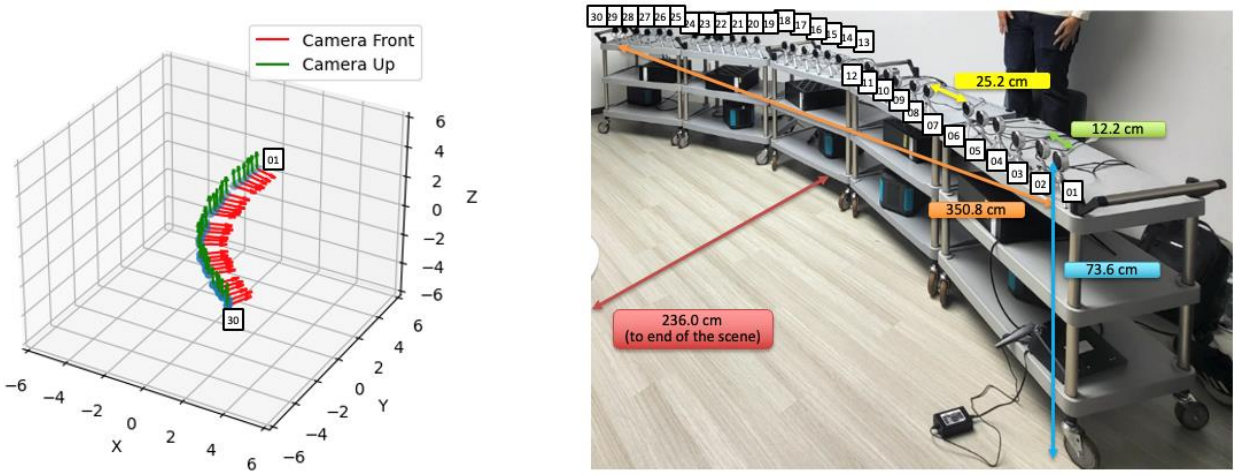

Figure 1. The visualization of camera rotation and position (left), and the acquiring environment of "VRroom1D" (right).

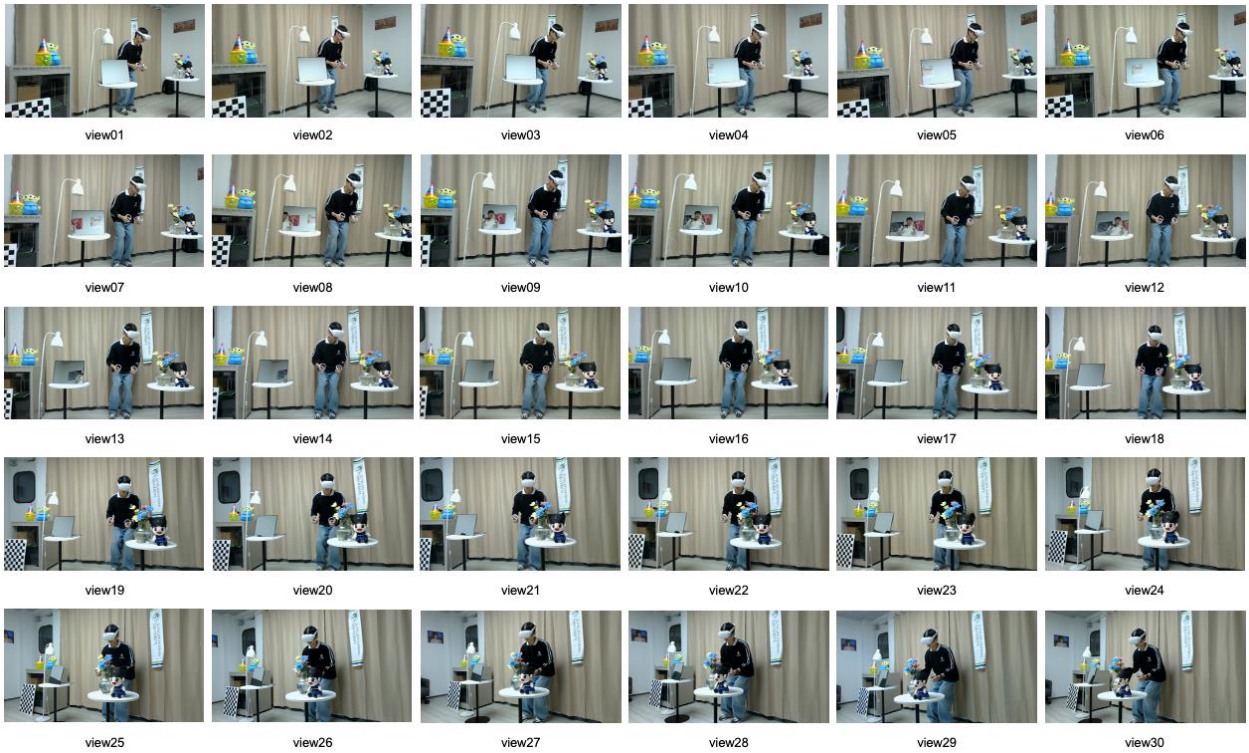

Figure 1. The snapshot example of "SKKU_VRroom1D" content (frame #72)

2.2 "SKKU_VRroom2D" Content

The "SKKU_VRroom2D" content was captured with 30 cameras facing the scene, arranged in three levels of height with 10 cameras on each level. The left image of Figure 3 visualizes the position of each camera, with all of them being in a nearly forward-facing orientation. Table 3 provides a basic description of the content, and the right image of Figure 3 shows the actual camera structure and physical conditions for camera array.

Table 3. The description of "SKKU_VRroom2D" content

Item	Specification
Number of frames	300
Number of views	30 (v01~v30)
Recommended test view	v14
Format	mp4 (H.264 codec)
Resolution	1920 x 1080
FPS	30.0
Total Size	374 MB

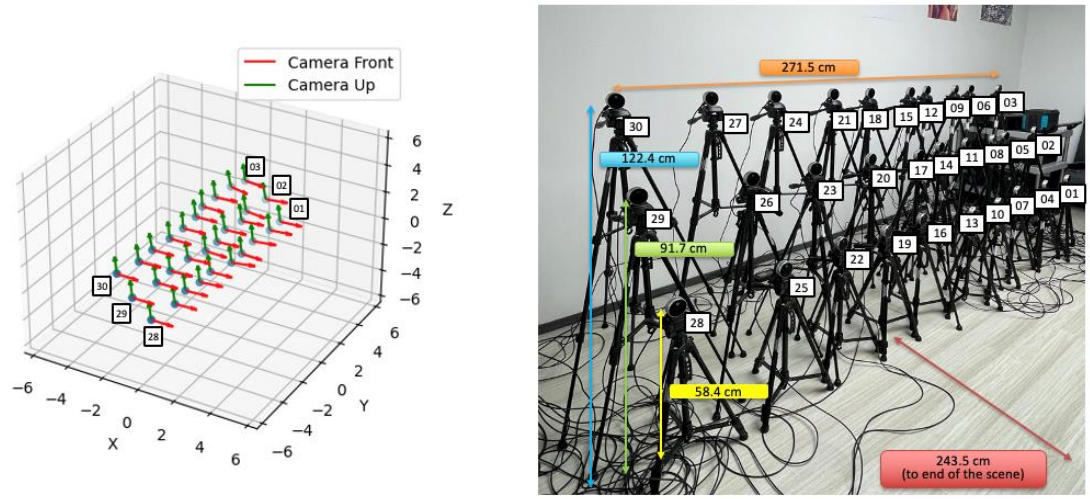

Figure 3. The visualization of camera rotation and position (left), and the acquiring environment of "VRroom2D" (right).

Figure 4. The snapshot example of "SKKU_VRroom2D" content (frame #8)

3 EE2.1 for "SKKU_VRroom" Datasets

The first 100 frames were used at this experiment. For "SKKU_VRroom1D", view11 was chosen as the test view, and for "SKKU_VRroom2D", view 14 served as the test view, with the remaining views used for training. The selected test views are relatively central and encompass the most elements described in Table 1 within the viewport.

The two sequences were verified through four 3D INVR models: NeRF[2][3], Instant-NGP[4], K-Planes[5], and Mixvoxels[6]. Instant-NGP is a model that enhances speed by simplifying the neural network structure and using multiresolution hash encoding. K-Planes is a model that represents high-dimensional data by decomposing it into multiple planes that can be linearly calculated. MixVoxels employs static and dynamic voxels to represent a scene. Since NeRF and Instant-NGP can only represent static scenes, a separate model was used for each frame. For these, rather than randomly

initializing the neural network's weights each frame, the training time was shortened by continuing the training using the weights learned from the previous frame.

For comparison with MIV, TMIV v17[7] was used as a main anchor mode. For depth estimation, the immersive video depth estimation (IVDE) was employed, and for view synthesis, the view weighting synthesizer (VWS) was employed. To align the experimental conditions with INVR, the MIV encoding was done without test view. Four atlases were used during the encoding process. Considering that compression is not applied to INVR models for this experiment, the video encoding process was not included in TMIV experiment.

Table 4 presents the objective quality assessment results of various models for "SKKU_VRroom1D" dataset. Based on PSNR, MixVoxels and K-Planes showed good results, while based on SSIM, MixVoxels and TMIV demonstrated favorable outcomes. Table 5 displays the results for "SKKU_VRroom2D", which, overall showed lower results compared to VRroom1D.

Table 4. Experimental results for **"SKKU_VRroom1D"** content test view v11

		TMIV	**NeRF**	**Instant-NGP**	**K-Planes**	**MixVoxels**
Objective Quality	**PSNR**	24.30	26.92	25.19	26.53	27.28
	SSIM	0.8793	0.8714	0.8547	0.8709	0.9361
Model (atlas) size (100 frames)		17.12GB	1.44 GB	3.51 GB	323.9 MB	530.5 MB

Table 5. Experimental results for **"SKKU_VRroom2D"** content test view v14

		TMIV	**NeRF**	**Instant-NGP**	**K-Planes**	**MixVoxels**
Objective Quality	**PSNR**	18.25	21.41	20.30	21.53	21.22
	SSIM	0.7643	0.8155	0.7813	0.8133	0.8790
Model (atlas) size (100 frames)		17.12GB	1.44 GB	3.51 GB	323.9 MB	531MB

The attached folder contains test view rendering results from the experiments above. For "SKKU_VRroom1D", v11 videos were rendered for each model that excluded v11 while training. In addition, for "SKKU_VRroom2D" v14 were rendered for each model that excluded v14 while training. Some results of rendering dynamic scene, subject to the camera path variations, have also been attached.

Figure 5 displays the results of rendering the first frame for the test view from each model that trained by "SKKU_VRroom1D". Unlike object quality measurement, Instant-NGP was observed to generally present a more natural appearance at subjective quality assessment. Both NeRF and Instant-NGP had 10,000 training iterations for the first frame, yet Instant-NGP showed more improved appearance. MixVoxels also demonstrated sharp results, excluding the area of the mirror. For MIV, some fragmentation was observed at group boundaries.

Ground Truth

MIV

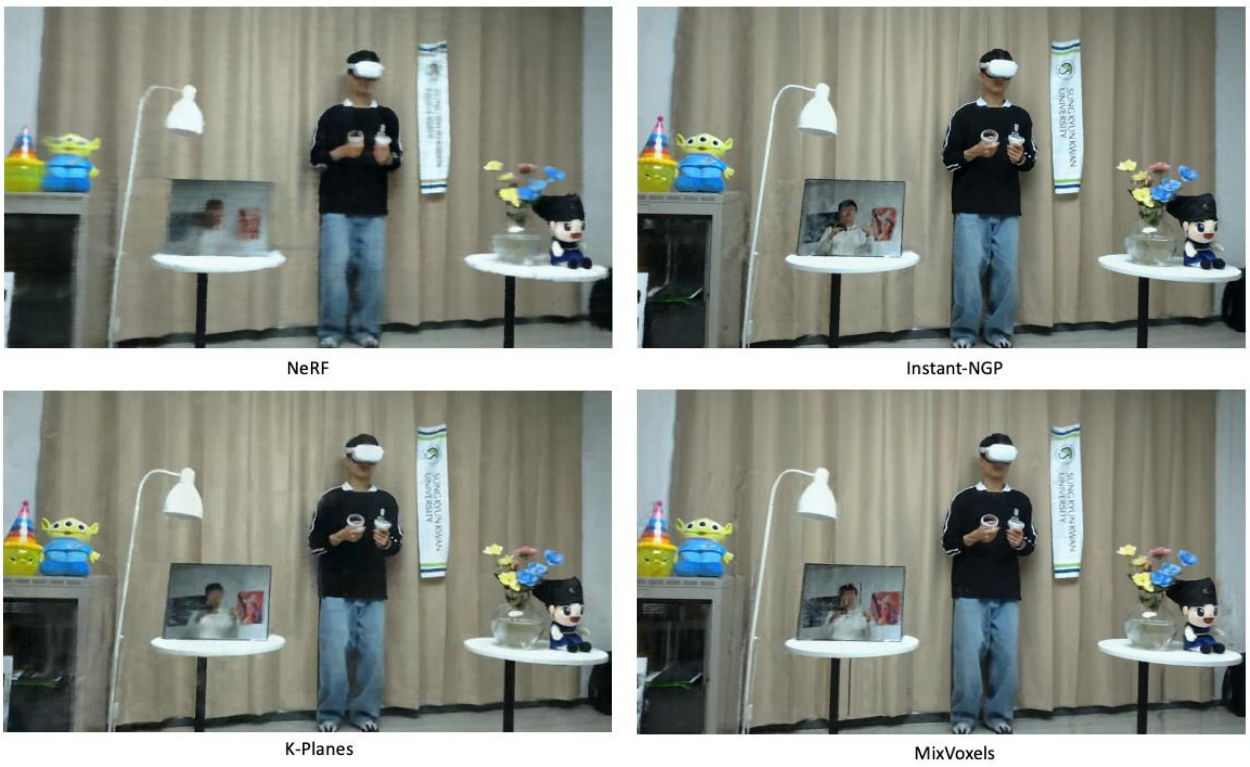

Figure 5. Visual quality of rendered test view v11 of "SKKU_VRroom1D"

Figure 6 shows the detailed view of the vase area in images rendered from each model of "SKKU_VRroom1D". It is observed that Instant-NGP relatively well stores the high-frequency details of the flowers and refraction of the stem in the water.

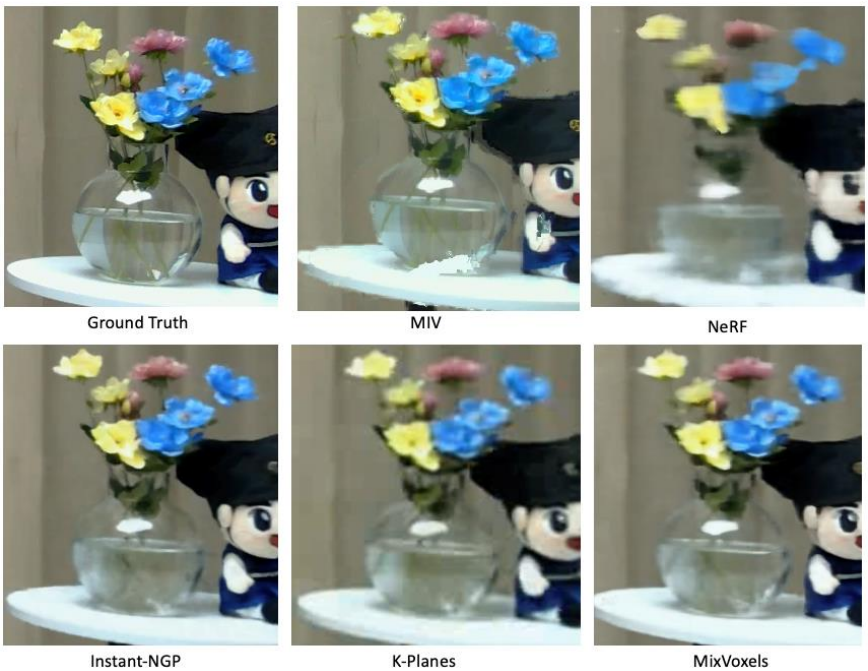

Figure 6. Detailed subjective quality comparison of vase area

Figure 7 presents the results concerning the mirror section of the first frame's rendered image. MIV provided a relatively sharp image within the mirror, but it exhibited artifact in border or the mirror.

Figure 7. Detailed subjective quality comparison of mirror area

4 Recommendation

"SKKU_VRroom" datasets can be usefully employed to evaluate various aspects of the performance of INVR models. We recommend "SKKU_VRroom" datasets to be used as one of INVR contents.

5 References

[1] J. L. Sch¨onberger and J.-M. Frahm, “Structure-from-motion revisited,” in Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis,” in European Conference on Computer Vision (ECCV), 2020.
[3] Yen-Chen, Lin, NeRF-pytorch, https://github.com/yenchenlin/nerf-pytorch for NeRF Software, GitHub repository, 2020.
[4] T. M¨uller, A. Evans, C. Schied, and A. Keller, “Instant neural graphics primitives with a multiresolution hash encoding,” ACM Transactions on Graphics (ToG), vol. 41, no. 4, pp. 1–15, 2022.
[5] S. Fridovich-Keil, G. Meanti, F. R. Warburg, B. Recht, and A. Kanazawa, “K-planes: Explicit radiance e lds in space, time, and appearance,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12 479–12 488, 2023.
[6] F. Wang, S. Tan, X. Li, Z. Tian, Y. Song, and H. Liu, “Mixed neural voxels for fast multi-view video synthesis,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 19 706–19 716.
[7] "Test model 17 for MPEG immersive video", Standard ISO/IEC JTC 1/SC 29/WG 4, MPEG/n0376, 2023.