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 m68240
July 2024, Sapporo

Title: Report on EE3: Thoughts on MIV DSDE Anchor Generation
Source: Sungkyunkwan University (SKKU)
Authors: Jong-Beom Jeong, Jun-Hyeong Park, Jaeyeol Choi, Yeong Gyu Kim, Eun-Seok Ryu (SKKU)

1 Introduction

During the last INVR meeting, EE3 was established to generate the anchor for INVR. EE3 aims to generate anchors by using the decoder-side depth estimation (DSDE) profile of MPEG immersive video (MIV) as the baseline, which encodes only textures and generates depth at decoder-side. This document provides experimental results for *VRroom1D*, and multiple discussion points for a better anchor generation.

2 Test Views Selection of *VRroom1D*

VRroom1D (M-NC1) was proposed by SKKU, which contains 30 forward facing views in natural environments[1]. Each view has 1920x1080 resolution, and cameras were aligned in 1-D arc shape where 12.2cm of interval was applied. In the last meeting, test views for *VRroom1D* were set to *v11* and *v27*, which contain reflected objects in the mirror. Note that high-frequency texture objects can be observed in all views.
During the last meeting, there was a request to increase the number of test views from 3 to around 4. This is because the VRroom1D sequence includes many views, totaling 30. Therefore, four views were newly selected as test views. Two views were selected that contain many objects within the mirror and additionally selected two more symmetrical views. Figure 1 shows the newly selected test views: *v07*, *v11*, *v20*, and *v24*, where snapshots are visualized in Figure 2. The views containing the mirror are *v07* and *v11*, which best depict the posters and people opposite the camera. The symmetrical views, *v20* and *v24*, represent objects such as the checkerboard, which is advantageous for identifying pixel shifts. In summary, training and test views are as follows:

- Test views: *v07*, *v11*, *v20*, *v24*
- Training views: All remaining views

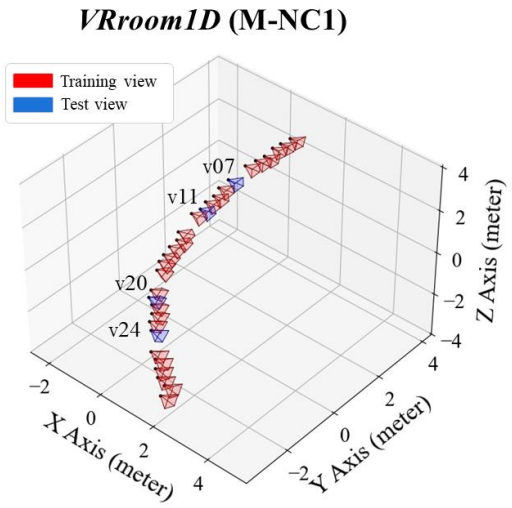
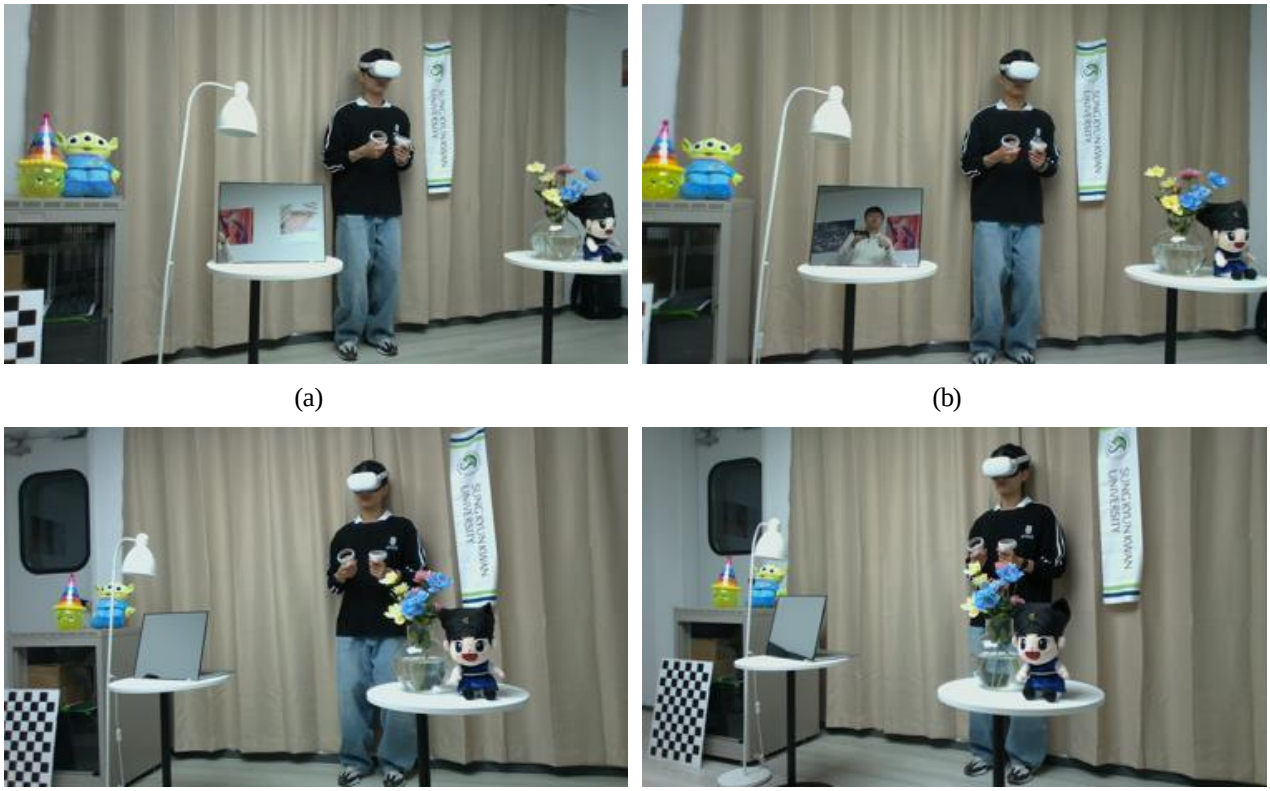


Figure 1. Training and test views visualization in *VRroom1D*



(a) (b)

(c) (d)

Figure 2. Snapshots of test views in *VRroom1D*, (a) *v07*, (b) *v11*, (c) *v20*, (d) *v24*

3 Test Conditions

During the offline discussion discussed in #issue2 of MPEG INVR Gitlab, it was decided to use version 18 of the test model for immersive video (TMIV). Experiments were conducted following the common test condition (CTC) of MIV that complies with this TMIV version. Compression was performed for a total of **97 frames** following the INVR CTC decided in the last meeting[3], and since both the group of pictures (GOP) size and intra period are 32, a total of 3 GOPs are included. As stated in the INVR CTC, *VRroom1D* begins encoding from **frame #16**. Since test views must be excluded during MIV encoding, these test views were removed from the ***sourceCameraIds*** and ***sourceCameraNames*** of the camera parameter.

This document introduces the results of two experimental conditions: **Partial** and **Full**. Table 1 shows the settings for both experimental conditions. In **Partial**, the number of atlases is limited to 4 according to the MIV DSDE CTC conditions, and each atlas has a resolution of approximately 2Kx4K. Considering that each view of the typical test sequences has a resolution of 2Kx1K, four views fit into each atlas. Therefore, following the MIV DSDE CTC conditions, a maximum of 16 views can be included in the atlas. However, since *VRroom1D* has a total of 26 training views, 16 views are included in the partial condition, and 10 are discarded. To address this, an additional experimental condition called **Full** was introduced to include all training views in the atlas. The number of atlases in the full condition is determined to include all training views, and in the *VRroom1D* experimental condition, a total of 7 atlases are used.

Four QPs were selected to Partial and Full, where similar bitrates are observed. Approximately, RP1 to RP4 shows 27, 12, 7.7, 4.8 Mbps, respectively. Following the INVR CTC, RGB-PSNR, SSIM, and LPIPS were measured.

Table 1. Experimental values for Partial and Full conditions

Item	**Partial**	**Full**
No. of atlases	4	7
No. of views in atlases	16	26
Included training views	v01, v03, v04, v06, v08,	All training views

in atlases	v10, v12, v13, v15, v18, v19, v21, v23, v25, v28, v30	
Excluded training views from atlases	v02, v05, v09, v14, v16, v17, v22, v26, v27, v29	None
QPs	22, 26, 30, 34	24, 30, 34, 39

4 Experimental Results and Discussions

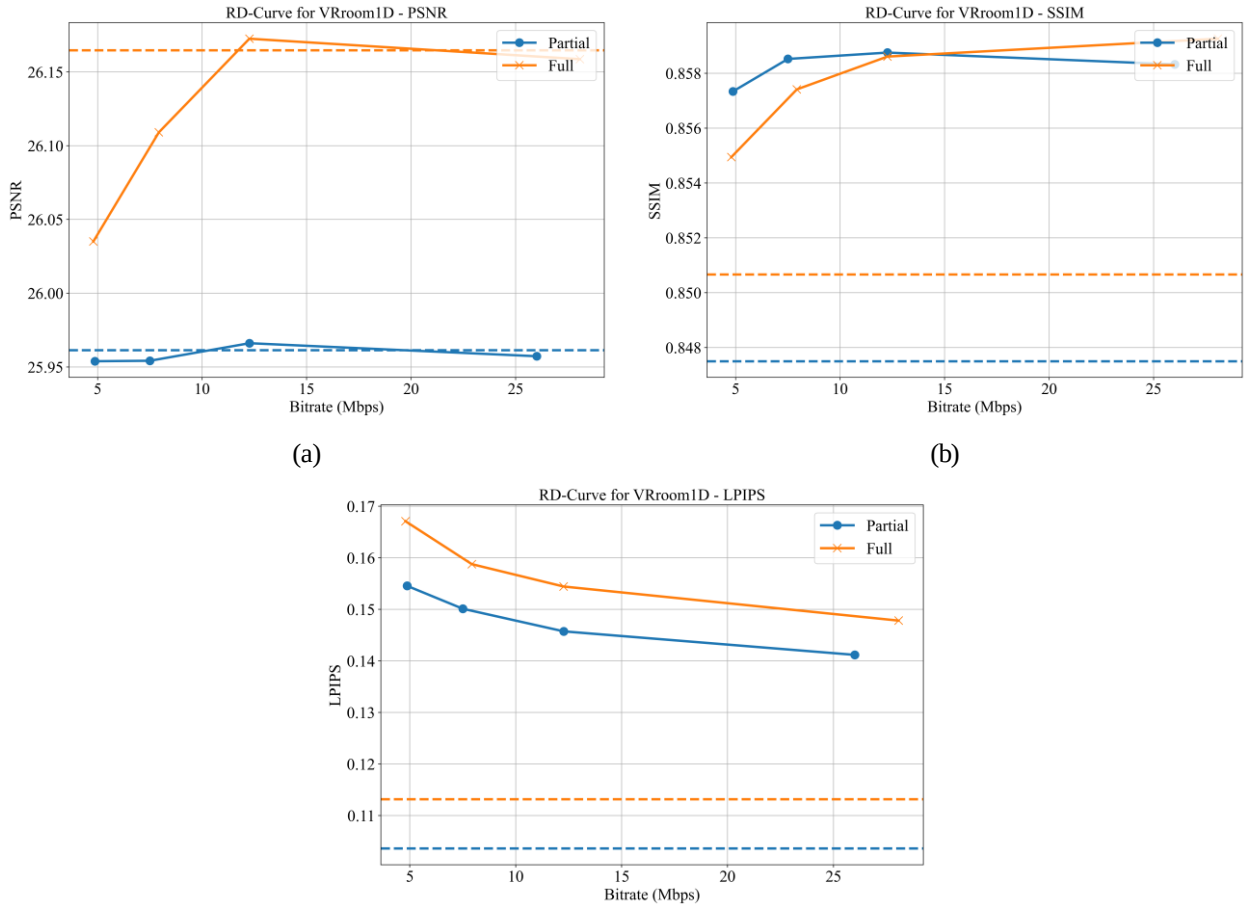


Figure 3. RD-curves of Partial and Full conditions, (a) RGB-PSNR ↑ , (b) SSIM ↑ , (c) LPIPS ↓ . Dotted lines represent values for RP0 (noncoded).

Figure 4. Synthesized test views of source, Partial and Full conditions, first row represents source, second row represents Partial (RP2, 12.26Mbps), third row represents Full (RP2, 12.26Mbps). First and second columns represent *v07*, third and fourth columns represent *v24*.

Figure 3 shows the RD-curve for Partial and Full conditions. The dashed line represents the results for RP0. At similar bitrates, increasing the number of atlases to include all training views is advantageous in terms of PSNR and SSIM. Although there are overlapping points between Partial and Full for SSIM, RP0 values are higher for Full. However, in the LPIPS, which measures the similarity of features between the source and reconstructed images, Partial showed better results.

Figure 4 shows the synthesized test views at RP2. The first row represents the source, the second row represents the Partial condition, and the third row represents the Full condition. In the Partial condition, relatively fewer training views were used, resulting in some holes and awkward color representation. However, since a relatively lower QP can be applied compared to Full, the texture quality is well-preserved. On the other hand, the Full condition, using all training views for synthesis, shows fewer holes compared to Partial. However, because more views were used in test view synthesis, some synthesis errors were observed.
Below are some discussion points based on the experimental results:

Discussion 1. Number of atlases

- To comply with the experimental constraints of the MIV DSDE profile, one can **(i)** use only some cameras within the entire space, **(ii)** use only odd or even-numbered cameras, or **(iii)** maintain 4 atlases and let the MIV encoder automatically select basic views, causing some training views to be dropped. None of these conditions are desirable as the INVR baseline. (i) restricts the viewing space, and (iii) results in outcomes that are highly dependent on the performance of the basic view selection algorithm of MIV.
- None of the above conditions can be considered fair experiments compared to INVR's NeRF and 3DGS methods, which use all views. This is because the number of training views does not directly affect the model size. The model size is influenced by the number of parameters within the neural network, and for 3DGS, the number of gaussians and spherical harmonics degrees are important.
- Experimental results showed that having more atlases (=containing all training views in atlases) is beneficial in RD performances. And for a fair comparison with NeRF and 3DGS methods which use all training views, increasing number of atlases to have all training views is recommended.

(a) (b) (c) (d)

Figure 5. Enlarged noticeable sections of texture and synthesized geometry in *v01* (RP0, Full), (a) enlarged texture, (b) enlarged geometry showing block artifacts, (c) enlarged texture, (d) enlarged geometry showing damaged object.

Discussion 2. IVDE configuration

- As shown in Figure 3, RP0 does not always show the best performance, and evaluation value differences are low, compared to the MIV. This implies that the compression rate of the texture has little impact on the quality of the test view. Synthesis artifacts always occur due to the inaccuracy of depth when synthesizing geometry with IVDE, regardless of the texture compression rate.
- Figure 5 shows the depth synthesized with IVDE and the original texture in the Full condition. The depth generated by IVDE exhibits stair-step artifacts and some information loss, leading to a decrease in the quality of the synthesized test view. This occurs regardless of the texture compression rate and may contribute to the flat appearance of the RD-curve.
- Tuning of IVDE configuration for each sequence will improve the quality of synthesized test view, especially in high bitrate (RP1). Then, drawing a non-flat RD-curve might be possible, but this needs further investigation.

Discussion 3. QP tuning

- The current INVR CTC recommends 5-50 Mbps of bitrates. In the draft CfP, four bitrate points for each sequence need to be defined, considering characteristics of test sequences.

5 Conclusion

MIV DSDE anchor generation was conducted for *VRroom1D*. Increasing number of atlases to include all training views is recommended to have a non-flat RD-curve and a fair comparison with NeRF and 3DGS methods. Further, for a non-flat RD-curve, optimizing IVDE configuration for each sequence is recommended as a part of EE3. Appropriate bitrate points are also needed for each sequence.

6 References

[1] J. Choi, Y. Ryu, Y. Choi, J. -B. Jeong, J. -H. Park, I. Yang, E. -S. Ryu, "[INVR]EE2.1-Related: Report with New Natural INVR Video Contents: SKKU_VRroom", ISO/IEC JTC1/SC29/WG4 input document m64721, October 2023, Hannover.

[2] A. Dziembowski, B. Kroon, J. Jung, "Common test conditions for MPEG immersive video", ISO/IEC JTC1/SC29/WG4 output document N0406, October 2023, Hannover.

[3] Y. Liao, G. Bang, O. L. Meur, M. Teratani, "Common test condition on implicit neural visual representation", ISO/IEC JTC1/SC29/WG4 output document N0518, April 2024, Rennes.