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 m64722
October 2023, Hannover

Title: **[INVR] Report on EE2.1: Inter Mode Dynamic NeRF Investigation**

Source: **Jun-Hyeong Park, Jaeyeol Choi, Jong-Beom Jeong, Eun-Seok Ryu (SKKU)**

1 Introduction

In this contribution, we present our research findings on the 'EE2.1 Rendering Quality of Dynamic NeRF.' Specifically, we focused on the Neural Residual Radiance Fields (ReRF) method among the proposed 3D INVR methods, notable for its use of the Inter mode strategy for dynamic NeRF. Our results reveal that ReRF effectively restores dynamic scenes with temporal-multiview consistency. However, when trained on natural content under same test conditions with CG content, a pervasive 'foggy' artifact emerged. Therefore, we recommend further investigation for EE2.1, including comprehensive hyperparameter and training strategy modifications. Additionally, to achieve a more compact representation, we suggest conducting further experiments with a frame-based compression module applicable to ReRF for EE2.2.

2 Neural Residual Radiance Fields (ReRF)

ReRF is a state-of-the-art NeRF variant presented at CVPR2023 that uses the inter mode to compactly represent dynamic scenes. The approach of ReRF explicitly models the residual and motion information between adjacent timestamps in the feature space using a voxel-grid representation. This methodology allows for a highly compact representation of dynamic scenes while achieving high-quality temporal-multiview consistency.

2.1 ReRF Training Strategies

ReRF's primary training strategy for the initial frame relies on an explicit voxel grid representation like DVGO [3] approach. This method makes use of explicit density and color feature grids, represented as $V\sigma$ and Vc. A compact MLP(ϕ) approximates these feature grids into radiance fields, serves as a feature decoder. This MLP operates globally across all frames. Following this, for the representation of subsequent frames, the process advances through the Motion Grid Estimation and Residual Grid Optimization stages.

[Motion Grid Estimation] To estimate motion between frames, an initial dense motion field (D_t) is calculated. This field gives a comprehensive mapping from the current frame back to the previous one, serving as a reference space. Within this grid, a voxel (p_t) can reference another voxel, (p_{t-1}), from the preceding frame. For a streamlined motion representation, we employ motion pooling techniques. Specifically, the dense motion field (D_t) is segmented into 8x8x8 cubes, and the motion inside each cube is averaged out. This approach results in a compact motion grid (M_t) that's a staggering 512 times more concise than the initial dense grid, enhancing its compressibility.

[Residual Grid Optimization] After motion estimation, there's a need to adjust for any discrepancies or nuances that aren't captured by the motion grid, and this is where the residual grid comes into play. Specifically, we warp the previous feature grid, (f_{t-1}), to generate the current base grid ($\hat{f}_t$) with the aid of the compact motion grid (M_t). During the optimization of the residual grid, only the residuals (r_t) are updated by backpropagating the gradients, keeping ($\hat{f}_t$) and the small MLP (ϕ) static. An L1 loss is applied to (r_t) to enhance its sparsity, reinforcing the grid's compactness, and making it more suitable for compression. Through this process, we obtain the residual information in a form that is sparse and optimized for compression.

2.2 ReRF Compression Module

Adopting a traditional keyframe-based strategy, the ReRF-based codec divides the feature grid sequences into continuous groups of feature grids (GOFs). Each GOF starts with an I-feature grid (keyframe) which contains the most crucial details for the following P-feature grid sequence. Initially, these grids are reshaped into matrices, facilitating data representation. Principal component analysis (PCA) is employed on (r_t) grids to emphasize significant data directions. After the PCA process, both the (f_1) grid and the subsequent (r_t) grids undergo a 3D discrete cosine transform (DCT) by first segmenting them into 8x8x8 voxel cubes for more efficient data handling. These transformed values are then quantized, creating a simplified version of the data. The final step involves encoding this quantized data with techniques like Huffman coding, optimizing it for transmission. Through these consecutive steps, ReRF effectively compresses data, ensuring its compactness for long-duration dynamic scenes.

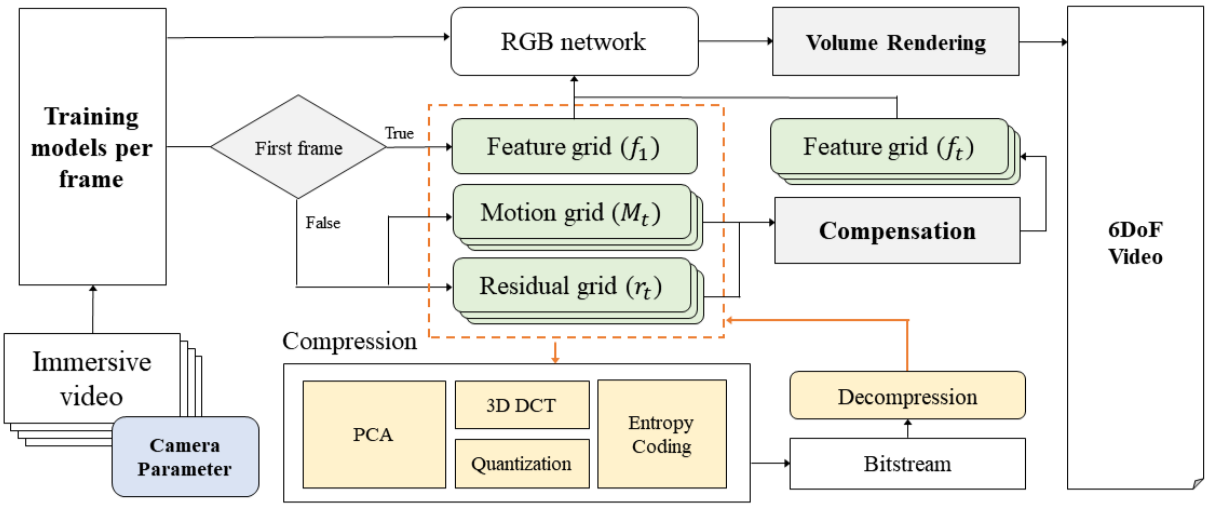

Figure 1. ReRF pipeline

In summary, ReRF employs a two-stage training strategy for dynamic scenes. Starting with an explicit voxel grid representation for the first frame, it then utilizes motion grid estimation and residual grid optimization to efficiently represent subsequent frames. This approach effectively captures inter-frame feature similarities and discrepancies, resulting in a concise and compressible model. Compression pipeline comprises keyframe-based grouping, reshaping, PCA application on (r_t) grids, 3D DCT transformation, quantization, and entropy coding. By deviating from using canonical space, which is the typical method in dynamic NeRF, this approach can effectively handle long-duration sequences. Figure 1 illustrates the entire process. For an in-depth understanding and further explanations, we recommend consulting the referenced paper [2].

3 Experimental Conditions

According to the EE documentation [1], CG test materials “Mirror” and “Garage” [4] are selected. For natural content, we chose “Coffee Martini” and “SKKU VRroom1D” [5]. We follow the training configuration described in the EE documentation, including view split for training-testing and frame collection, to conduct our experiments.

ReRF's training incorporates progressive scaling, allowing us to configure checkpoint for resizing the voxel grid. Depending on the chosen number of checkpoints, the initial voxel size is reduced, and at each checkpoint, the resolution of the voxel grid is doubled along each axis. This approach ensures that as the training progresses, ReRF operates on increasingly dense voxel grids, facilitating fine learning.

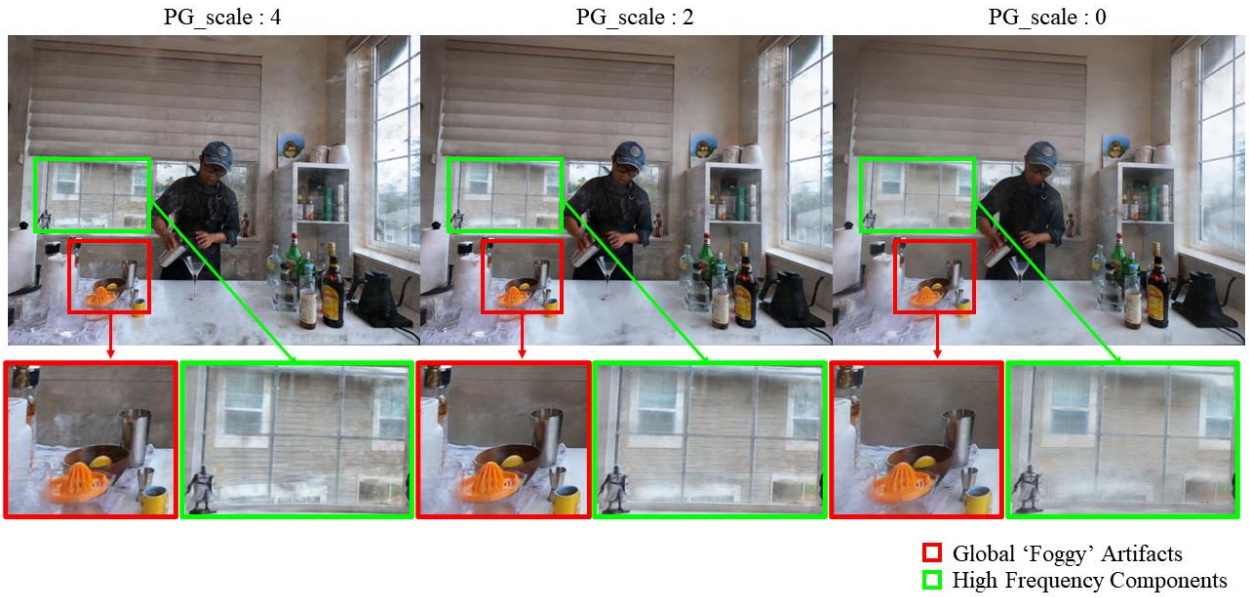

Figure 2. Enlarged noticable sections of "Coffee Martini" with different PG scaling

This training methodology was introduced to reduce training time and enhance the detailed representation of high-frequency components. However, we observed that when reconstructing natural content, initiating training with dense voxels from the beginning helps prevent the occurrence of global artifacts. Figure 2 presents the rendering results with adjusted progressive scaling for natural contents. As a result, in this experiment, particularly for natural content, we decreased the number of checkpoints in the progressive scaling process and increased the voxel grid size compared to CG content. For detailed experiment configurations, please refer to Table 1.

Table 1. Training configuration

	Mirror	Garage	Coffee Martini	SKKU VRroom
Data Type	CG	CG	NC	NC
Test Views	v6, v8	v61	v0	v15
Training Views	All of the rest (13views)	6 x 6 views with 1 interval camera	All of the rest (16views)	All of the rest (29 views)
Training Frames	32	90	32	32
Iteration (per frame)	30K	30K	30K	30K
Voxel grid size	384*384*216	384*384*216	384*384*256	384*384*256
PG scaling	4	4	2	2

4 Experiment Results

We evaluated the rendering quality of ReRF in its uncompressed version, conducting experiments with both CG contents and natural contents. For each sequence, we rendered results from test views that were not included in the training set and compared them to the ground truth. We provide image quality measurement metrics, including PSNR, SSIM, and LPIPS, for the evaluation.

We observe that ReRF can achieve high reconstruction quality on CG contents. As mentioned, model adjustments were applied to enhance the reconstruction quality for natural contents. The experimental results indicated well-preserved temporal-multiview consistency on natural contents. However, global artifacts persisted, and the reconstruction quality in high-frequency components exhibited relatively unsatisfactory results on natural contents. Figures 3-6 display subjective results for each sequence, and Table 2 presents the evaluation metrics on test view.

(a) 1st frame (b) 32nd frame

Figure 3. Rendering results of “Mirror”

(a) 1st frame (b) 90th frame

Figure 4. Rendering results of “Garage”

(a) 1st frame (b) 32nd frame

Figure 5. Rendering results of “Coffee Martini”

(a) 1st frame (b) 32nd frame

Figure 6. Rendering results of "SKKU VRroom1D"

Table 2. Rendering Quality Assessment on Test Views

Method Employed: ReRF (Uncompressed)				
Sequence Name	Mirror (CG)	Garage (CG)	Coffee Martini (NC)	SKKU VRroom (NC)
Test View	v6	v61	v0	v15
PSNR(dB)	30.19	30.35	24.02	23.38
SSIM	0.8821	0.8701	0.8332	0.8287
LPIPS	0.1328	0.1236	0.1618	0.1839

5 Conclusion

In conclusion, ReRF demonstrates the potential for reconstructing dynamic scenes with temporal-multiview consistency using the inter mode strategy. However, further experiments are required to achieve higher rendering quality for uncompressed model of ReRF especially in natural content. Additionally, experiments with a frame-based compression module applicable to ReRF is also required. We recommend for continued exploration in EE2 with various inter mode-based dynamic NeRF variants, including ReRF.

6 References

[1] G. Bang, Y. Liao, P. Hellier, M. Teratani, Exploration experiments on implicit neural visual representation, ISO/IEC JTC 1/SC 29/WG 04 N 0389, Geneva, July 2023.

[2] L. Wang, Q. Hu, Q. He, Z. Wang, J. Yu, T. Tuytelaars, L. Xu, M. Wu, Neural Residual Radiance Fields for Streamably Free-Viewpoint Videos, CVPR 2023.

[3] Sun, Cheng, M. Sun, H. T. Chen, Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction, CVPR 2022.

[4] Gun Bang, Jinho Lee, Jungwon Kang, [INVR]EE2.1 report with New INVR Video content, ISO/IEC JTC 1/SC 29/WG 4, m64394, Geneva, July 2023.
[5] Jaeyeol Choi, Yeongil Ryu, Yihyun Choi, [INVR]EE2.1-Related: Report with New Natural INVR Video Contents: SKKU_VRroom, ISO/IEC JTC 1/SC 29/WG 4, m64721, Hannover, October 2023.