Appendix: Masked Space-Time Hash Encoding for Efficient Dynamic Scene Reconstruction

Anonymous Author(s)
Affiliation
Address
email

1 Implementation Details

For the hash encoding, we adopt multi-resolution hash tables with exponential progressive growing sizes setting of Instant-NGP [5]. Specifically, we instantiate the 4D hash table of $L = 16$ levels with grid resolution from a space-time resolution of $16 \times 16 \times 16 \times 15$ to the maximum resolution of $2048 \times 2048 \times 2048 \times \mathcal{T}$, where $\mathcal{T}$ is the frame number. For the 3D hash table, the spatial resolution is the same as the 4D hash table except that it has no temporal dimension. The hash table size is set to 2^{19} for both the 3D and 4D hash tables. Interpolations are applied to the 3D and 4D hash tables, utilizing tri-linear and tetra-linear functions, respectively. The feature dimension for each hash item is $F = 2$, which is the same as Instant-NGP.

For the mask and uncertainty encoding, we use the voxel-grid representation with 64^3 space resolution with tri-linear interpolation. We empirically found that a relatively small spatial resolution is enough to achieve high-quality results.

We implement the proposed method based on PyTorch and develop a customized CUDA extension to facilitate the incorporation of the rectangular hash grid, as required by MSTH. All experiments are performed on an NVIDIA RTX 3090 with 24GB RAM.

2 Campus Dataset

The Campus dataset is collected on a campus environment characterized by complex and realistic scenes, including intricate dynamics such as pedestrians walking around, fountains with splashing water droplets, and swiftly moving vehicles on the roadways. Fig. 2 showcases some of the captured dynamic scenes. These scenes exhibit more challenging time-variant patterns with large movement areas. The capturing process is similar to DyNeRF [4]. We build a multi-view capture system including 24 GoPro Black Hero 9 cameras. All videos are captured with a resolution of 3840×2160 and a frame rate of 30 FPS. The videos of different views are synchronized by aligning the time codes. The intrinsic and extrinsic parameters are estimated with COLMAP [8].

3 Additional Ablations

3.1 Hash Table Size

We conduct an evaluation to assess the impact of varying sizes of 4D hash tables. Fig. 1 showcases the novel view rendering performances across different hash table sizes. As the number of hash table items increases, we note a corresponding decrease in LPIPS scores, which eventually saturate at a range of 2^{18} to 2^{19}. Our findings suggest that LPIPS is more consistent with human perception, particularly in dynamic scenes. Furthermore, we observe that PSNR may not be a reliable indicator of reconstruction quality and is not sensitive to the artifacts caused by hash collisions.

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

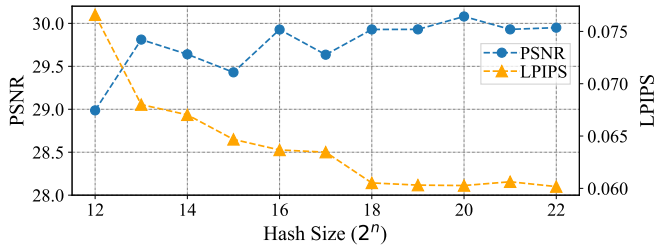

Figure 1: Ablation on the sizes of 4D hash tables on the *flame-salmon* scene. We show the PSNR and LPIPS curves.

	PSNR ↑	LPIPS ↓
Mask	29.64	0.087
Pearson	29.87	0.107
Hard	28.74	0.124
MI	**29.93**	**0.063**

Table 1: Ablation study on different approaches to maximize correlation. Maximizing mutual information outperforms other choices.

Figure 2: We visualize four dynamic scenes of the Campus dataset, which is collected by a synchronized GoPro camera array. Video demonstrations are provided in the supplemental video.

3.2 Correlation

To maximize the correlation between the estimated uncertainty u and the mask m, we conduct an evaluation of different correlation methods: (1) the Pearson correlation coefficient, which is commonly used for measuring the linear correlation (We minimize the Pearson to force the negative correlation between u and m). (2) A hard-coded linear relation, i.e., $m = a \times u + b$ with a learnable a and b. (3) MSTH with mutual information. Tab. 1 demonstrates the results. The Pearson coefficient and the hard-coded relation both achieve degenerated performances since they model the relation of m and u linearly, which makes the mask m reside near a consecutive small interval.

3.3 Ray Sampling

In our evaluation, we observed that the ray sampling method has minimal impact on scenes in the Plenoptic Video dataset and the Google Immersive dataset. This finding can be attributed to the relatively simple and slow movements present in these datasets, which are easier to capture with greater accuracy. However, in the campus dataset, we found that the proposed space-time weighted ray sampling is beneficial to generate sharp boundaries for the moving object. Fig. 3 visually demonstrates the effectiveness of our proposed ray sampling strategy. It clearly illustrates that the proposed sampling strategy leads to clear boundaries and improved reconstruction details.

4 Additional Results

Per-scene results. We report the per-scene metrics for Plenoptic Video dataset in Tab. 3, Google Immersive dataset in Tab. 4, and D-NeRF dataset in Tab. 5. We also visualize the qualitative results on Fig. 4, Fig. 6, Fig. 5 and 7. For video comparisons and results, we provide them in the supplemental video. We highly recommend watching it for a better visual comparison.

Details of SSIM. For SSIM evaluation, prior methods adopt two different settings in data range, which results in different SSIM values. Most methods [4, 3, 7] adopt the data range values 2.0, which is the default argument when calling the SSIM function in the scikit-image library. In our paper, we report SSIM in this setting to be consistent with prior methods. HyperReel employs the variant of SSIM that utilizes a data range of 1.0, resulting in lower SSIM scores than that evaluated with a data range of 2.0. For a fair comparison with HyperReel, we also measure the

Table 2: SSIM comparison with HyperReel[1].

Method	Plenoptic Video	Google Immersive
HyperReel [1]	92.7	87.4
Ours	**94.3**	**91.8**

w/o weighted w/ weighted w/o weighted w/ weighted

Figure 3: Qualitative comparison for ablation on ray sampling strategy. The images are chosen from *Talk* in our proposed Campus dataset. We set $\tau_1 = 21$ and $\tau_2 = 5$, which yields an expected sampling rate that was marginally higher for the dynamic regions and frames than for the static counterparts.

Table 3: Quantitative results on six scenes of Plenoptic Video dataset [4]. †denotes the HexPlane [3] setting which removes the coffee-martini scene.

	Coffee	Cook Spinach	Cut Beef	Flame Salmon	Flame Steak	Sear Steak	Mean	Mean†
PSNR	28.72	33.62	33.75	29.93	34.13	34.07	32.37	33.10
SSIM	95.08	97.57	97.67	95.97	98.03	98.07	97.06	97.46
LPIPS	0.077	0.056	0.053	0.063	0.042	0.043	0.056	0.051

Table 4: Quantitative results on seven scenes of Google Immersive dataset [2].

	Welder	Flames	Trunk	Horse	Exhibit	Face	Alexa	Mean
PSNR	26.76	30.55	28.27	28.95	29.79	31.55	31.15	29.57
SSIM	88.71	91.69	95.12	96.27	95.42	98.11	97.49	94.68
LPIPS	0.133	0.092	0.092	0.099	0.088	0.076	0.069	0.093

Table 5: Quantitative results on eight scenes of D-NeRF dataset [6].

	Lego	Bouncing Balls	Hell Warrior	Hook	Jumping Jacks	Mutant	Standup	Trex	Mean
PSNR	26.40	39.27	25.05	28.04	31.32	34.51	33.85	32.33	31.35
SSIM	95.13	99.54	96.20	96.41	98.35	99.03	98.85	98.17	97.71
LPIPS	0.060	0.007	0.051	0.034	0.023	0.009	0.012	0.027	0.028

SSIM in this setting. The comparison is shown in Tab. 2, where our method surpasses +1.6 SSIM in the Plenoptic Video dataset and +4.4 SSIM in the Google Immersive dataset.

References

[1] B. Attal, J.-B. Huang, C. Richardt, M. Zollhoefer, J. Kopf, M. O'Toole, and C. Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. *arXiv preprint arXiv:2301.02238*, 2023.

[2] M. Broxton, J. Flynn, R. Overbeck, D. Erickson, P. Hedman, M. Duvall, J. Dourgarian, J. Busch, M. Whalen, and P. Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics (TOG)*, 39(4):86–1, 2020.

[3] A. Cao and J. Johnson. Hexplane: a fast representation for dynamic scenes. *arXiv preprint arXiv:2301.09632*, 2023.

[4] T. Li, M. Slavcheva, M. Zollhoefer, S. Green, C. Lassner, C. Kim, T. Schmidt, S. Lovegrove, M. Goesele, R. Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022.

[5] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.

[6] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.

[7] Sara Fridovich-Keil and Giacomo Meanti, F. R. Warburg, B. Recht, and A. Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023.

[8] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

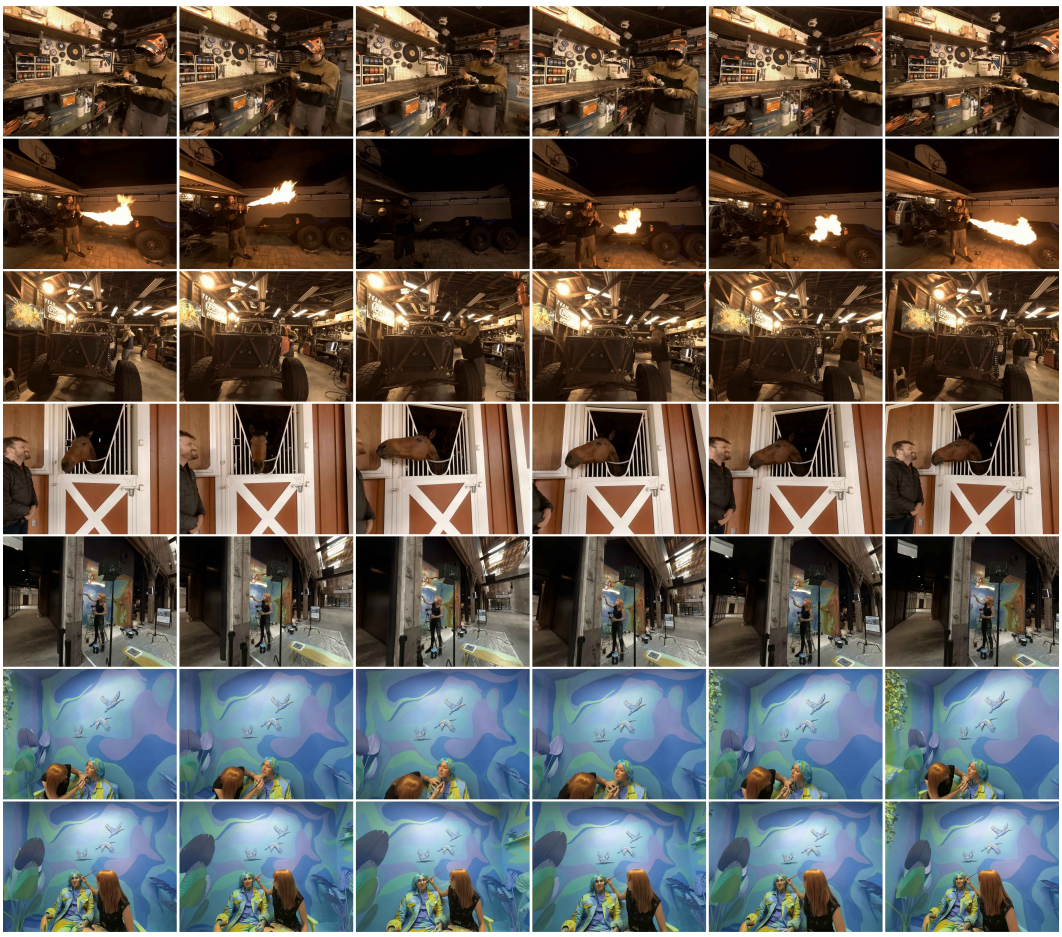

Figure 4: Qualitative results on Google Immersive dataset. Scenes from top to bottom are: *Welder*, *Flames*, *Truck*, *Horse*, *Alexa Meade Exihibit*, *Face Paint 1*, *Face Paint 2*.

Figure 5: Qualitative results on Campus dataset. Scenes from top to bottom are: *Playing*, *Fountain*, *Pedestrian*, *Talk*.

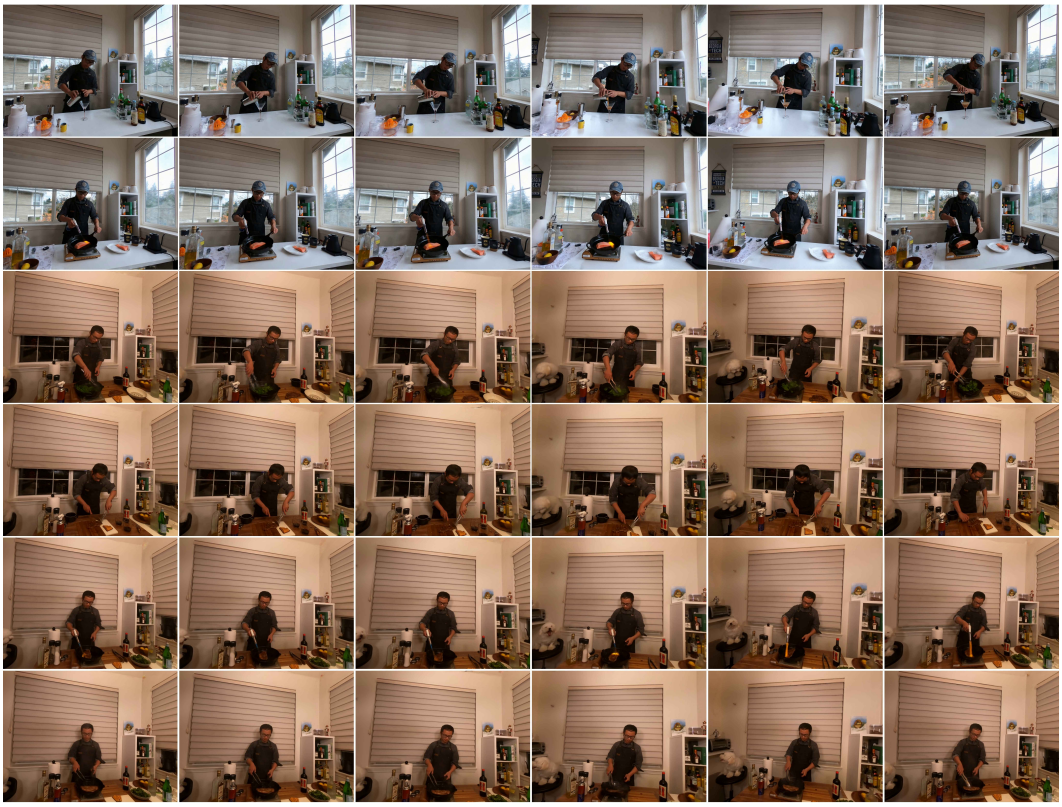

Figure 6: Qualitative results on Plenoptic Video datasets. scenes from top to bottom are: *coffee-martini*, *flame-salmon*, *cook-spinach*, *cut-roasted-beef*, *flame-steak*, *sear-steak*

Figure 7: Qualitative results on D-NeRF dataset. Scenes from top to bottom are: *BouncingBalls*, *Hook*, *HellWarrior*, *JumpingJack*, *Mutant*, *StandUp*, *Trex*.