DASH: 4D Hash Encoding with Self-Supervised Decomposition for Real-Time Dynamic Scene Rendering

Supplementary Material

In the supplementary material, we provide additional implementation details in Appendix A. Then more experimental results are conducted in Appendix B. Afterward, we introduce further ablation studies in Appendix C. Finally, we discuss the failure cases of our proposed DASH in Appendix D.

A. Implementation Details

A.1. Hyperparameter Settings

Our hyperparameters mainly follow the settings of 3DGS [6]. For the 3D hash encoder in decomposition, we set the hash table size to 2^{19}, the number of levels to 16, and the feature dimension per level to 2. We set the 4D hash encoding parameters to $L = 32$ levels, $T_l = 2^{19}$ for the hash table size, and $F = 2$ for the feature dimension per level. In the decomposition stage, the loss weights are empirically set to $\lambda_s = 0.1$ and $\lambda_c = 0.2$. The threshold τ is defined as $||\Delta p||$ at the top k% percentile, where k is determined based on scene characteristics. Specifically, for Neural 3D Video [8] dataset, k% is set between 5% to 10%, while for Technicolor Light Field [11] dataset, it ranges from 15% to 20%. In the deformation field training stage, we apply $\lambda_r = 0.5$ and $\lambda_c = 0.2$. The learning rate schedule primarily follows Grid4D [5], with the MLP decoder's learning rate adjusted based on scene scale. Additionally, the learning rate for grid hash encoders is set 10–50 times higher than that of the MLP decoder. We use Adam [7] optimizer with $\beta = (0.9, 0.999)$ and set the background to black.

A.2. Evaluation Details

For LPIPS computation, we use the AlexNet LPIPS variant for all of our comparisons in the main paper (as do all of the baseline methods).

To ensure fair SSIM comparisons across datasets, we employ the scikit-image implementation for Neural 3D Video [8] and Technicolor Light Field [11] datasets following K-Planes [4] and E-D3DGS [2].

B. Additional Results

B.1. Per-scene Results on Technicolor Light Field

We provide the per-scene results for the experiments on the real-world Technicolor Light Field [11] dataset. Tab. 1 and Fig. 1 illustrate the comparisons. It can be observed that DASH exhibits superior rendering quality compared to the previous methods, demonstrating the effectiveness and generalization of our method under various scenes.

B.2. Generality of Decomposition

We provide additional qualitative results demonstrating the integration of our method into Grid4D (denoted as Grid4D+dec). As shown in Fig. 2, Grid4D+dec significantly enhances detail rendering compared to Grid4D. We attribute this improvement mainly to our decomposition method. It effectively separates static and dynamic components, allowing the network to better focus on dynamic regions.

C. Additional Ablations

C.1. Ablation on Hash Table Size

We conduct ablation studies on hash table sizes ranging from 2^{16} to 2^{19}. As shown in Tab. 2, reducing the hash table size leads to performance degradation while decreasing model size. We attribute this trade-off to increased hash collision rates at smaller table sizes, which compromise feature query accuracy and ultimately degrade reconstruction quality. Notably, even with these smaller models, compelling results can still be achieved when memory constraints are paramount.

C.2. Ablation on Hash Resolution Levels

We conduct ablation studies on the number of hash resolution levels, evaluating configurations from 8 to 32. As shown in Tab. 3, reducing the number of levels results in performance degradation while decreasing model size. We attribute this trade-off to insufficient high-frequency feature extraction at coarser resolutions, which limits the model's ability to capture fine details and ultimately compromises reconstruction quality. Importantly, even with these smaller model configurations, satisfactory performance can still be attained when storage efficiency is prioritized over rendering fidelity.

C.3. Ablation on Encoders

We conducted ablation studies on the encoders, as shown in Tab. 4. Specifically, we use different encoding methods on decomposed dynamic components. Ablation results shows our method outperforms others. This demonstrates that 4D hash provides better feature encoding for dynamic parts by mitigating feature overlap.

D. Failure Cases

Large Motion Modeling with Monocular Settings. In monocular settings, the input is sparse in both camera pose

Method	Birthday			Fabien			Painter		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DyNeRF [8]	29.20	0.952	0.067	32.76	0.965	0.242	35.95	0.972	0.146
HyperReel [1]	29.99	-	0.053	34.70	-	0.186	35.91	-	0.117
E-D3DGS [2]	32.37	0.964	0.066	34.78	0.957	0.145	36.18	0.968	0.097
Grid4D [5]	32.02	0.967	0.058	33.94	0.948	0.181	35.64	0.963	0.120
Grid4D+dec	31.59	0.965	0.046	34.87	0.957	0.151	36.60	0.972	0.083
Ours	32.97	0.968	0.039	35.52	0.960	0.135	36.87	0.973	0.081

Method	Theater			Train			Mean		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DyNerf [8]	29.53	0.939	0.188	31.58	0.962	0.067	31.80	0.958	0.142
HyperReel [1]	33.32	-	0.115	29.74	-	0.072	32.73	-	0.109
E-D3DGS [2]	31.10	0.937	0.145	31.36	0.951	0.074	33.16	0.955	0.105
Grid4D [5]	31.12	0.936	0.174	30.21	0.929	0.124	32.59	0.949	0.128
Grid4D+dec	31.19	0.945	0.140	30.97	0.947	0.083	33.04	0.955	0.114
Ours	32.51	0.948	0.135	31.81	0.952	0.075	33.94	0.960	0.097

Table 1. Additional quantitative comparisons on Technicolor Light Field [11] dataset. The best, the second best, and the third best are colored in table cells. Results of DyNeRF [8] and HyperReel [1] are from their original paper, while we calculate metrics of all Gaussian-based methods by running their official codes.

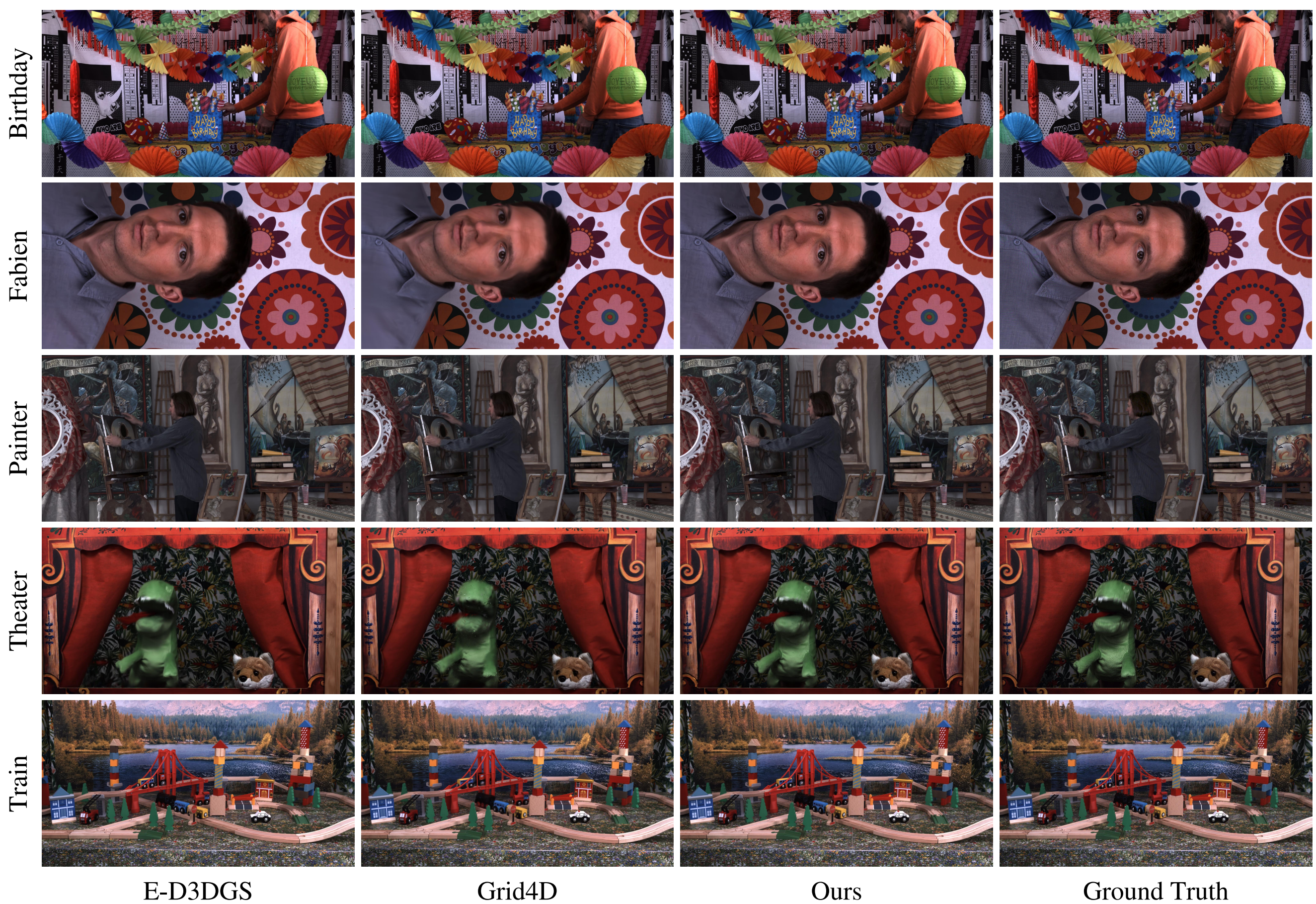

Figure 1. Additional qualitative comparisons on Technicolor Light Field [11] dataset.

Figure 2. Additional qualitative results on Grid4D+dec.

Table Size	PSNR	SSIM	LPIPS	Model Size (MB)
2^{16}	32.97	0.974	0.043	**16**
2^{17}	33.05	0.974	0.043	31
2^{18}	33.08	0.975	0.041	60
2^{19}	**33.16**	**0.980**	**0.040**	115

Table 2. Additional quantitative ablation results on hash table size in the cook spinach scene from Neural 3D Video [8] dataset.

Levels	PSNR	SSIM	LPIPS	Model Size (MB)
8	32.92	0.974	0.045	**28**
16	33.08	0.975	0.042	57
24	33.14	0.975	0.041	86
32	**33.16**	**0.980**	**0.040**	115

Table 3. Additional quantitative ablation results on the number of hash resolution levels in the cook spinach scene from Neural 3D Video [8] dataset.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Plane-based [12] + dec	31.65	0.964	0.056
Grid-based [5] + dec	31.74	0.967	**0.049**
Ours	**32.22**	**0.969**	0.050

Table 4. Quantitative comparisons on Neural 3D Video dataset.

and timestamp dimensions. This may cause the local minima of overfitting with training images in some complicated scenes. Our method may fail in modeling large motions or dramatic scene changes. This phenomenon is also observed in previous NeRF-based methods [3, 8–10] and Gaussian-based methods [5, 12, 13], producing blurring results. Fig. 3 shows some failed samples.

(a) Broom (b) Teapot

Figure 3. Failure cases of modeling large motions and dramatic scene changes. (a) The sudden motion of the broom makes optimization harder. (b) Teapots have large motion and a hand is entering/leaving the scene.

References

[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16610–16620, 2023. 2

[2] Jeongmin Bae, Seoha Kim, Youngsik Yun, Hahyun Lee, Gun Bang, and Youngjung Uh. Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In *European Conference on Computer Vision*, pages 321–335. Springer, 2024. 1, 2

[3] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 3

[4] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-Planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 1

[5] Xu Jiawei, Fan Zexin, Yang Jian, and Xie Jin. Grid4D: 4D decomposed hash encoding for high-fidelity dynamic scene rendering. *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1, 2, 3

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[8] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard 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. 1, 2, 3

[9] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[10] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc 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. 3

[11] Neus Sabater, Guillaume Boisson, Benoit Vandame, Paul Kerbiriou, Frederic Babon, Matthieu Hog, Remy Gendrot, Tristan Langlois, Olivier Bureller, Arno Schubert, et al. Dataset and pipeline for multi-view light-field video. In *Proceedings of the IEEE conference on computer vision and pattern recognition Workshops*, pages 30–40, 2017. 1, 2

[12] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. 3

[13] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3D Gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 3