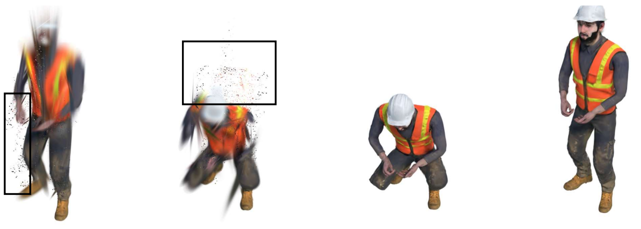
-Supplementary Material-
4D Gaussian Splatting for Real-Time Dynamic Scene Rendering

Guanjun Wu[1*], Taoran Yi[2*], Jiemin Fang[3†], Lingxi Xie[3], Xiaopeng Zhang[3],
Wei Wei[1], Wenyu Liu[2], Qi Tian[3], Xinggang Wang[2†‡]
[1]School of CS, Huazhong University of Science and Technology
[2]School of EIC, Huazhong University of Science and Technology [3]Huawei Inc.
{guajuwu, taoranyi, weiw, liuwy, xgwang}@hust.edu.cn
{jaminfong, 198808xc, zxphistory}@gmail.com tian.qi1@huawei.com



(a) Ours w/o dx (b) Ours

Figure 1. Visualization of ablation study about ϕ_x.

A. Appendix

A.1. Introduction

In the supplementary material, we mainly introduce our hyperparameter settings of experiments in Sec. A.2. Then more ablation studies are conducted in Sec. A.3. Finally, we delve into the limitations of our proposed 4D-GS in Sec. A.4.

A.2. Hyperparameter Settings

Our hyperparameters mainly follow the settings of 3D-GS [8]. The basic resolution of our multi-resolution HexPlane module $R(i,j)$ is set to 64, which is upsampled by 2 and 4. The learning rate is set as 1.6×10^{-3}, decayed to $1.6e-4$ at the end of training. The Gaussian deformation decoder is a tiny MLP with a learning rate of 1.6×10^{-3} which decreases to 1.6×10^{-3}. The batch size in training is set to 1. The opacity reset operation in [8] is not used as it does not bring evident benefit in most of our tested scenes. Besides, we find that expanding the batch size will indeed contribute to rendering quality but the training cost increases accordingly.

Different datasets are constructed under different capturing settings. D-NeRF [14] is a synthesis dataset in which each timestamp has only one single captured image following the monocular setting. This dataset has no background which is easy to train, and can reveal the upper bound of our proposed framework. We change the pruning interval to 8000 and only set a single upsampling rate of the multi-resolution HexPlane Module $R(i,j)$ as 2 because the structure information is relatively simple in this dataset. The training iteration is set to 20000 and we stop 3D Gaussians from growing at the iteration of 15000.

The DyNeRF dataset [9] includes 15 – 20 fixed camera setups, so it's easy to get the sfm [15] point in the first frame, we utilize the dense point-cloud reconstruction and downsample it lower than 100k to avoid out of memory error. Thanks to the efficient design of our 4D Gaussian splatting framework and the tiny movement of all the scenes, only 14000 iterations are needed and we can get the high rendering quality images.

HyperNeRF's dataset is captured with less than 2 cameras in feed-forward settings. We change the upsampling resolution up to $[2,4]$ and the hidden dim of the decoder into 256. Similar to other works [3, 13], we found that Gaussian deformation fields always fall into the local minima that link the correlation of motion between cameras and objects even with static 3D Gaussian initialization. And we're going to reserve the splitting of the relationship in the future works.

A.3. More Ablation Studies

Position Deformation. We find that removing the output of the position deformation head can also model the object motion. It is mainly because leaving some 3D Gaussians in the dynamic part, keeping them small in shape, and then scaling them up at a certain timestamp can also model the dynamic part. However, this approach can only model

---

[*] Equal contributions.
[†] Project Lead.
[‡] Corresponding author.

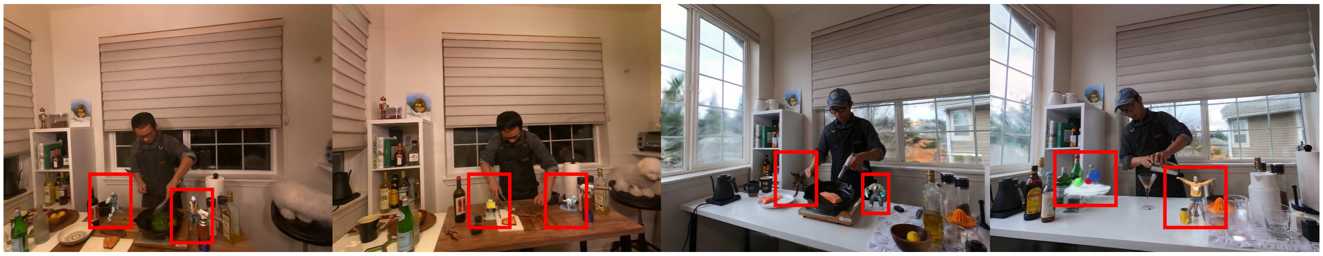

(a) Cook Spinach (b) Cut Beef (c) Flame Salmon (d) Coffee Martini

Figure 2. More visualization of composition in 4D Gaussians. (a) Composition with Punch and Standup. (b) Composition with Lego and Trex. (c) Composition with Hellwarrior and Mutant. (d) Composition with Bouncingballs and Jumpingjacks.

Table 1. Perscene results of HyperNeRF's vrig datasets [13] by different models.

Method	3D Printer		Chicken		Broom		Banana	
	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM	PSNR	MS-SSIM
Nerfies [12]	20.6	0.83	26.7	0.94	19.2	0.56	22.4	0.87
HyperNeRF [13]	20.0	0.59	26.9	0.94	19.3	0.59	23.3	0.90
TiNeuVox-B [3]	22.8	0.84	28.3	0.95	21.5	0.69	24.4	0.87
FFDNeRF [6]	22.8	0.84	28.0	0.94	21.9	0.71	24.3	0.86
3D-GS [8]	18.3	0.60	19.7	0.70	20.6	0.63	20.4	0.80
Ours	22.1	0.81	28.7	0.93	22.0	0.70	28.0	0.94

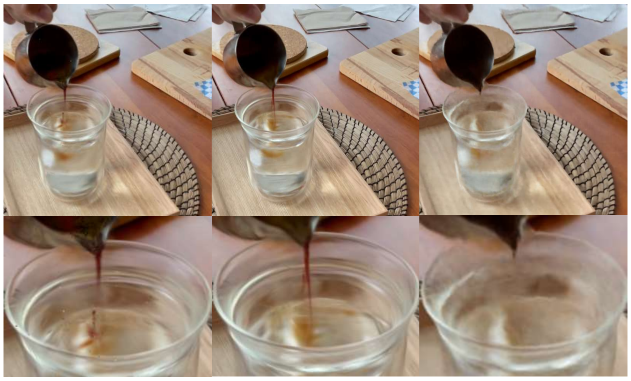

(a) Ours (b) Ours w φ_{shs}, φ_o (c) TiNeuVox

Figure 3. Visualization of ablation study in $\phi_{\mathcal{C}}$ and ϕ_{α} comparing with TiNeuVox [3].

coarse object motion and lost potential for tracking. The visualization is shown in Fig. 1.

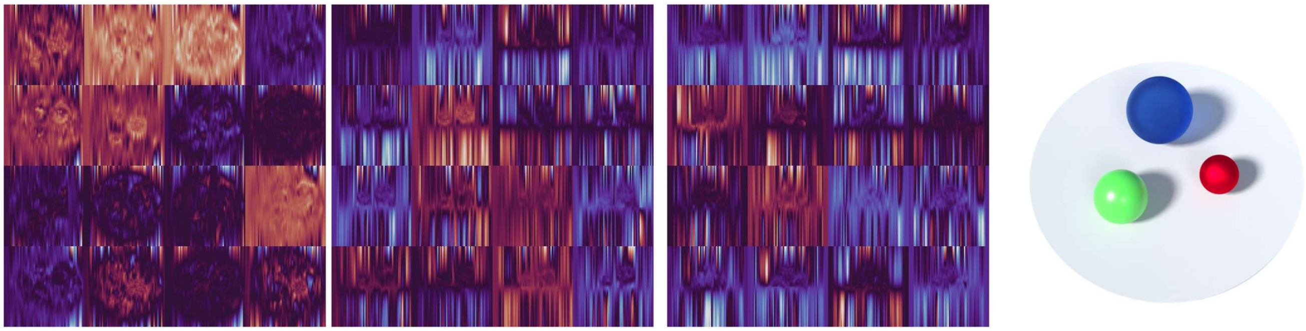
Editing with 4D Gaussians. We provide more visualization in editing with 4D Gaussians with Fig. 2. This work only proposes a naive approach to transformation. It is worth noting that when applying the rotation of the scenes, 3D Gaussian's rotation quaternion q and scaling coefficient s need to be considered. Meanwhile, some interpolation methods should be applied to enlarge or reduce 4D Gaussians.

Color and Opacity's Deformation. When encountered with fluid or non-rigid motion, we adopt another two output MLP decoder $\phi_{\mathcal{C}}$, ϕ_{α} to compute the deformation of 3D Gaussian's color and opacity $\Delta\mathcal{C} = \phi_{\mathcal{C}}(f_d)$, $\Delta\alpha = \phi_{\alpha}(f_d)$. Tab. 4 and Fig. 3 show the results in comparison with TiNeuVox [3]. However, it is worth noting that modeling Gaussian color and opacity change may cause irrational shape changes when rendering novel views. *i.e.* the Gaussians on the surface should move with other Gaussians but stay in the place and the color is changed, making the tracking difficult to achieve.

Spatial-temporal Structure Encoder. We have explored why 4D-GS can achieve such a fast convergence speed and rendering quality. As shown in Fig. 4, we visualize the full features of R_1 in bouncingballs. It's explicit that in the $R_1(x, y)$ plane, the spatial structure of the scenes is encoded. Similarly, $R_1(x, z)$ and $R_1(y, z)$ also show different view structure features. Meanwhile, temporal voxel grids $R_1(x, t), R_1(y, t)$ and $R_1(z, t)$ also show the integrated motion of the scenes, where large motions always stand for explicit features. So, it seems that the proposed HexPlane module encodes the features of spatial and temporal information.

A.4. More Discussions

Monocular Dynamic Scene Reconstruction. In monocular settings, input data are sparse from both cameration


(a) $R_1(x, y)$ (b) $R_1(x, z)$ (c) $R_1(y, z)$ (d) Training View 1

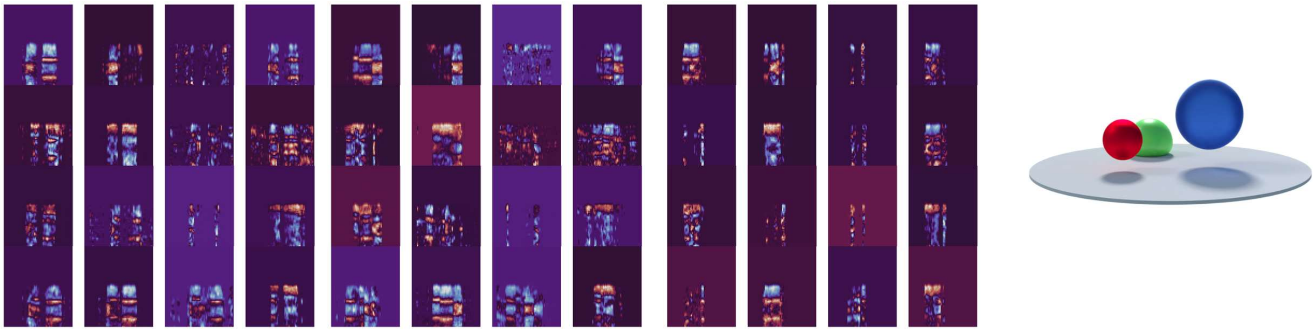

(e) $R_1(x, t)$ (f) $R_1(y, t)$ (g) $R_1(z, t)$ (h) Training View 2

Figure 4. More visualization of the HexPlane voxel grids $R(i, j)$ in bouncing balls. (a)-(c), (e)-(f) stand for visualization of $R_1(i, j)$, where grids resolution equals to 64×64.

Table 2. Per-scene results of DyNeRF's [9] datasets.

Method	Cut Beef		Cook Spinach		Sear Steak	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
NeRFPlayer [16]	31.83	0.928	32.06	0.930	32.31	0.940
HexPlane [2]	32.71	0.985	31.86	0.983	32.09	0.986
KPlanes [4]	31.82	0.966	32.60	0.966	32.52	0.974
MixVoxels [17]	31.30	0.965	31.65	0.965	31.43	0.971
Ours	32.90	0.957	32.46	0.949	32.49	0.957
Method	Flame Steak		Flame Salmon		Coffee Martini	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
NeRFPlayer [16]	27.36	0.867	26.14	0.849	32.05	0.938
HexPlane [2]	31.92	0.988	29.26	0.980	-	-
KPlanes [4]	32.39	0.970	30.44	0.953	29.99	0.953
MixVoxels [17]	31.21	0.970	29.92	0.945	29.36	0.946
Ours	32.51	0.954	29.20	0.917	27.34	0.905

pose and timestamp dimensions. This may cause the local minima of overfitting with training images in some complicated scenes. As shown in Fig. 5, though 4D-GS can render relatively high quality in the training set, the strong overfitting effects of the proposed model cause the failure of rendering novel views. To solve the problem, more priors such as depth supervision or optical flow may be needed.

Large Motion Modeling with Multi-Camera Settings. In the DyNeRF [9]'s dataset, all the motion parts of the scene are not very large and the multi-view camera setup also provides a dense sampling of the scene. That is the

Table 3. Per-scene results of synthesis datasets.

Method	Bouncing Balls			Hellwarrior			Hook			Jumpingjacks		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
3D-GS [8]	23.20	0.9591	0.0600	24.53	0.9336	0.0580	21.71	0.8876	0.1034	23.20	0.9591	0.0600
K-Planes[4]	40.05	0.9934	0.0322	24.58	0.9520	0.0824	28.12	0.9489	0.0662	31.11	0.9708	0.0468
HexPlane[2]	39.86	0.9915	0.0323	24.55	0.9443	0.0732	28.63	0.9572	0.0505	31.31	0.9729	0.0398
TiNeuVox[3]	40.23	0.9926	0.0416	27.10	0.9638	0.0768	28.63	0.9433	0.0636	33.49	0.9771	0.0408
Ours	40.62	0.9942	0.0155	28.71	0.9733	0.0369	32.73	0.9760	0.0272	35.42	0.9857	0.0128

Method	Lego			Mutant			Standup			Trex		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
3D-GS [8]	23.06	0.9290	0.0642	20.64	0.9297	0.0828	21.91	0.9301	0.0785	21.93	0.9539	0.0487
K-Planes [4]	25.49	0.9483	0.0331	32.50	0.9713	0.0362	33.10	0.9793	0.0310	30.43	0.9737	0.0343
HexPlane [2]	25.10	0.9388	0.0437	33.67	0.9802	0.0261	34.40	0.9839	0.0204	30.67	0.9749	0.0273
TiNeuVox [3]	24.65	0.9063	0.0648	30.87	0.9607	0.0474	34.61	0.9797	0.0326	31.25	0.9666	0.0478
Ours	25.03	0.9376	0.0382	37.59	0.9880	0.0167	38.11	0.9898	0.0074	34.23	0.9850	0.0131

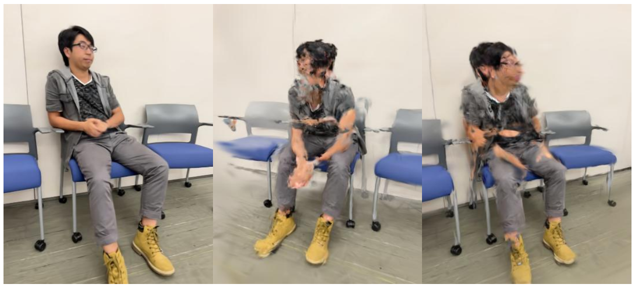

(a) Training View (b) Novel View 1 (c) Novel View 2

Figure 5. Novel view rendering results in the iPhone datasets [5].

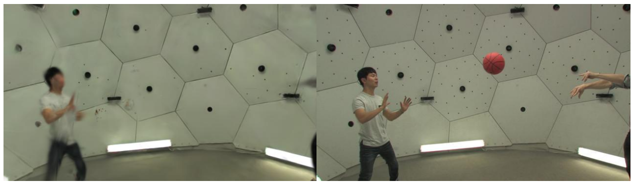

(a) Ours (b) Ground Truth

Figure 6. Rendering results on sports dataset [7] also used in Dynamic3DGS [11].

Table 4. Ablation Study on ϕ_C and ϕ_α, comparing with TiNeuVox [3] in Americano of HyperNeRF [13]'s dataset.

Method	Americano	
	PSNR	MS-SSIM
TiNeuVox-B [3]	28.4	0.96
Ours w/ ϕ_C,ϕ_α	31.53	0.97
Ours	30.90	0.96

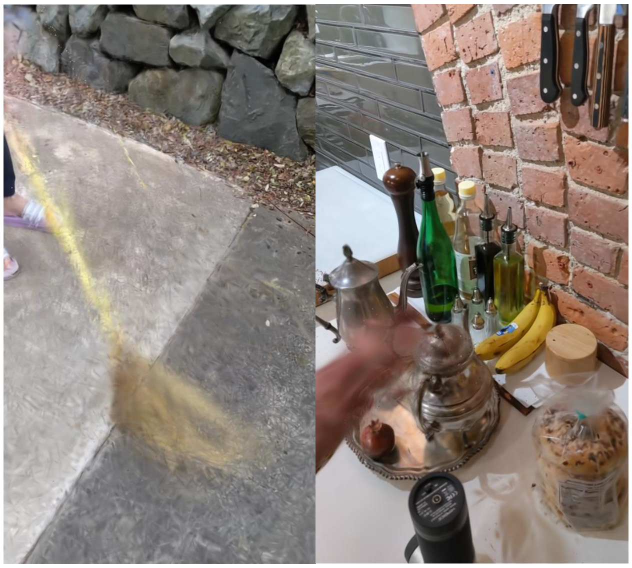

(a) Broom (b) Teapot

Figure 7. 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.

reason why 4D-GS can perform a relatively high rendering quality. However, in large motion such as sports datasets [7] used in [11], 4DGS cannot fit well within short times as shown in Fig. 6. Online training [1, 11] or using information from other views like [10, 18] could be a better approach to solve the problem with multi-camera input.

Large Motion Modeling with Monocular Settings. 4D-GS uses a deformation field network to model the motion of 3D Gaussians, which may fail in modeling large motions or dramatic scene changes. This phenomenon is also observed in previous NeRF-based methods [3, 9, 13, 14], producing blurring results. Fig. 7 shows some failed samples. Exploring more useful priors could be a promising future direction.

References

[1] Jad Abou-Chakra, Feras Dayoub, and Niko Sünderhauf. Particlenerf: Particle based encoding for online neural radiance fields in dynamic scenes. *arXiv preprint arXiv:2211.04041*, 2022. 4

[2] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 3, 4

[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. 1, 2, 4

[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. 3, 4

[5] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. 4

[6] Xiang Guo, Jiadai Sun, Yuchao Dai, Guanying Chen, Xiaoqing Ye, Xiao Tan, Errui Ding, Yumeng Zhang, and Jingdong Wang. Forward flow for novel view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16022–16033, 2023. 2

[7] Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social motion capture. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3334–3342, 2015. 4

[8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4):1–14, 2023. 1, 2, 4

[9] 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, 3, 4

[10] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. High-fidelity and real-time novel view synthesis for dynamic scenes. In *SIGGRAPH Asia Conference Proceedings*, 2023. 4

[11] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 4

[12] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[13] 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. 1, 2, 4

[14] 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. 1, 4

[15] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1

[16] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 3

[17] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19706–19716, 2023. 3

[18] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 4