# DiVa-360: The Dynamic Visual Dataset for Immersive Neural Fields

Cheng-You Lu[1*] Peisen Zhou[1*] Angela Xing[1*] Chandradeep Pokhariya[2] Arnab Dey[3]
Ishaan Nikhil Shah[2] Rugved Mavidipalli[1] Dylan Hu[1] Andrew I. Comport[3]
Kefan Chen[1] Srinath Sridhar[1]

[1]Brown University, [2]IIIT Hyderabad, [3]I3S-CNRS/Université Côte d'Azur
https://ivl.cs.brown.edu/research/diva

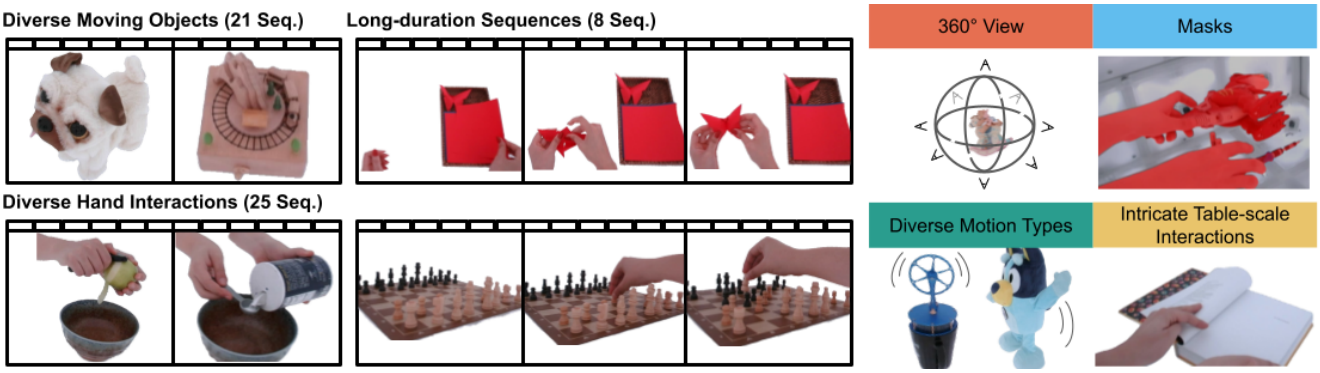


Figure 1. DiVa-360 is a **real-world** 360° multi-view visual dataset of dynamic tabletop scenes captured using a customized low-cost capture system consisting of 53 RGB cameras. DiVa-360 provides (1) 360° coverage of dynamic scenes, (2) foreground-background segmentation masks, synchronized audio, and detailed text descriptions, and (3) diverse scenes with intricate motions.

## Abstract

*Advances in neural fields are enabling high-fidelity capture of the shape and appearance of dynamic 3D scenes. However, their capabilities lag behind those offered by conventional representations such as 2D videos because of algorithmic challenges and the lack of large-scale multi-view real-world datasets. We address the dataset limitation* ***in our CVPR 2024 paper****, DiVa-360, a real-world 360° dynamic visual dataset that contains synchronized high-resolution and long-duration multi-view video sequences of table-scale scenes captured using a customized low-cost system with 53 cameras. It contains 21 object-centric sequences categorized by different motion types, 25 intricate hand-object interaction sequences, and 8 long-duration sequences for a total of 17.4 M image frames. In addition, we provide foreground-background segmentation masks, synchronized audio, and text descriptions. We benchmark the state-of-the-art dynamic neural field methods on DiVa-360 and provide insights about existing methods and future challenges on long-duration neural field capture.*

## 1. Introduction

Neural fields [64], or neural implicit representations, have recently emerged as useful representations in computer vision, graphics, and robotics [56, 64] for capturing properties such as radiance [4, 5, 26, 40, 41], shape [30, 38, 42, 43, 59, 69], and dynamic motion [8, 16, 29, 34, 36, 45, 58, 60, 62]. Their high fidelity, continuous representation, and implicit compression [14] properties make them attractive as immersive digital representations of our dynamic world.

However, despite their popularity, neural fields remain less capable than conventional representations for representing dynamic scenes. Though we can easily watch hours-long 2D videos, this is not yet achievable efficiently with 3D neural fields due to long training times [3, 9, 15, 26, 28, 29, 36, 54, 58, 60]. We believe that **large-scale, real-world** datasets of dynamic scenes with associated benchmarks are essential for continued progress in this problem. While some real-world dynamic datasets exist [8, 13, 29, 30, 35, 45, 55, 65, 67, 70], they are limited to room-scale scenes or specific categories like hu-

*Equal Contribution

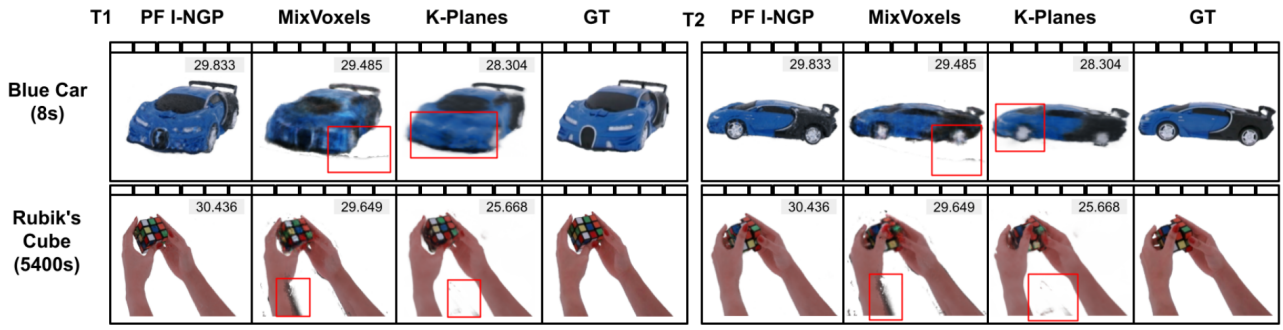


Figure 2. Here, we showcase reconstruction results across time steps from PF I-NGP [41], MixVoxels [58], and K-Planes [15] trained on our dataset. Though PF I-NGP does not directly utilize temporal information, it performs better than MixVoxels and K-Planes. We demonstrate more visualization results in the supplementary Section 2.

mans [19, 20, 22, 33, 46, 71], or are captured with monocular or forward-facing cameras that do not always provide sufficient multi-view cues for immersive reconstruction [17, 30, 35, 45]. Furthermore, most of the sequences in these datasets [13, 29, 30, 35, 45, 67, 70] are short, often less than 15 seconds, limiting their use for building methods that capture long-duration scenes.

To address these limitations, we present **DiVa-360**, a real-world dynamic visual dataset that contains synchronized high-resolution long-duration table-scale sequences captured by a 360° multi-view camera system (see Figure 1). Our dataset includes high-resolution (1280×720), high-framerate (120 FPS), and up to 3 mins long videos captured from 53 RGB cameras spanning 360°. We provide 46 dynamic sequences, including 21 object-centric sequences, 25 hand-object interaction sequences with human routine activities, and 8 long duration dynamic sequences. In total, DiVa-360 dynamic dataset contains **17.4 M** image frames and foreground-background segmentation masks of 53 dynamic scenes over **2738** seconds.

Capturing such large-scale data requires advances in capture systems and benchmarking metrics. We have built a new low-cost capture system called **BRICS (Brown Interaction Capture System)** which is designed to capture synchronized, high-framerate, and high-fidelity data. In addition, we propose standardized metrics for reconstruction quality and runtime, and compare baseline methods on these metrics [15, 41, 58]. Surprisingly, we observe that methods that model each frame in a dynamic sequence without directly using temporal information [41] outperform state-of-the-art dynamic methods [15, 58] in terms of reconstruction quality and even training speed (see Figure 2). To summarize, we make the following contributions:

- **BRICS**: A low-cost capture system specifically designed for 360° capture of table-scale dynamic scenes with 53 synchronized RGB cameras.
- **DiVa-360 Dataset**: The largest dataset (17.4 M frames) for dynamic neural fields with 21 object-centric sequences categorized by different motion types, 25 hand-object interaction sequences including routine human activities, and 8 long-duration dynamic sequences.
- **Benchmark & Analysis**: We benchmark the dataset with state-of-the-art methods and enable a better understanding of the current state of dynamic neural fields.

We believe our work can help the community take a leap from the current focus on short dynamic videos toward a more holistic understanding of longer dynamic scenes.

## 2. Related Work

**Neural Fields**: Neural fields, or coordinate-based implicit neural networks, have generated considerable interest in computer vision [64] because of their ability to represent geometry [11, 37, 43] and appearance [34, 40, 53]. Neural radiance fields (NeRF) [40] and its variants [4, 31, 42, 58, 69] uses a multilayer perceptron (MLP) to model density and color for photorealistic novel view synthesis and 3D reconstruction. Since the training cost of NeRFs is high, several methods have tried to address this limitation [10, 26, 41, 49]. Naturally, some approaches have also turned their focus towards dynamic neural fields [9, 15, 16, 29, 30, 34, 36, 44, 45, 47, 54, 58, 60, 66]. However, due to the lack of long-duration datasets, these methods have been limited to brief sequences. Our work enables further research in long-duration dynamic neural field research with a richer dataset containing long sequences.

**Multi-Camera Capture Systems**: Capturing multi-view data with high resolution and framerate requires specialized hardware and software systems. The earliest multi-camera capture systems were extensions of stereo cameras to 5–6 cameras [25], which were later extended to capture a hemispherical volume [24] with up to 50 cameras for 3D and 4D reconstruction using non-machine learning techniques [57]. The focus of most existing multi-camera capture systems has been on room-scale scenes for human or environment capture [23, 72]. While some table-scale datasets exist, notably for hand interaction capture [7, 73], they have only a limited number of cameras. In contrast, our BRICS sys-

tem is specially designed for dense 53-view visual capture of table-scale scenes, and our sequences showcase intricate interactions in high fidelity.

**Datasets for Dynamic Neural Fields**: While plenty of datasets exist for NeRF methods [1, 5, 12, 21, 27, 39, 40, 48, 61, 68] their focus has been on static scenes. For dynamic scenes, numerous datasets such as DyNeRF [29], NDSD [70], ILFV [8], NeRF-DS [67], and Deep3DMV [32] exist, but they are limited to only a short duration (~15s), or have only forward-facing cameras. BlockNeRF [55] lacks focus on objects and provides limited views. Eyeful Tower [65] provides dynamic data up to 2000 s long, but the framerate is less than 4 FPS. Monocular videos of human faces [44, 45], human activities [30], or outdoor scene [35] have been used for neural field reconstructions, but a single camera restricts visibility resulting in low effective multi-view factors (EMF)[17]. While Objaverse [13] and SAPIEN [63] provide articulated objects, they are not sourced from the real world. Our dataset stands out by offering a 360° view of real-world (non-synthetic) long dynamic sequences of objects and hand-object interaction captured by 53 synchronized cameras (see supplementary Table 1 and 2). Furthermore, each sequence is accompanied by foreground-background segmentation masks. Hence, we do not need to worry about the domain gap, the influence from the background, and the insufficient multiview cues.

## 3. Brown Interaction Capture System (BRICS)

To capture long-duration sequences of table scale objects and interactions, we designed and built our own hardware and software called the **Brown Interaction Capture System (BRICS)** which is shown in supplementary Figure 1.

**BRICS Hardware**: Our system uses an aluminum frame with fitted panels across each side that contain RGB cameras, microphones, and LED light strips. For 360° capture, we installed a transparent shelf to place objects. A custom communication setup that compresses and transmits data to a control workstation, comprehensive capture capabilities, and efficient data management all allow for 360° view capture with low latency.

**BRICS Software**: We design specialized software for managing data and adopt network-based synchronization [2] with an accuracy of 2-3 ms. For camera calibration, we capture a single calibration frame with ArUco markers affixed to the wall. We generate camera poses for the 53 cameras using COLMAP [50, 51], and refine them using I-NGP's [41] photometric loss for improved reconstruction quality. Finally, we built software for efficiently transferring terabytes of data from the control workstation to cloud storage. Our goal is to make this dataset useful for learning long-duration dynamic neural fields of appearance - existing methods [3, 9, 15, 28, 29, 36, 54, 58, 60] have been limited to only short durations (~10s). We fully benchmark all sequences in our dynamic dataset. In total, DiVa-360 dynamic dataset contains **17.4 M** image frames and foreground-background segmentation masks of 53 dynamic scenes over **2738** seconds. To our knowledge, this is the largest-scale dynamic dataset with a focus on table-scale interactions.

## 4. DiVa-360 Dataset

**Dynamic Objects**: We captured 21 dynamic sequences with everyday objects and toys that move. To be representative of real-world motions, we chose objects with different types of motion: (1) Slow motion: objects that perform slow, continuous motions, (2) Fast motion: objects that move or transform drastically (3) Detailed motion: objects that perform precise small motions (4) Repetitive motion: objects that repeat the same motion pattern (5) Random motion: objects that perform indeterministic motions (see supplementary Table 5).

**Interactions**: In addition to dynamic objects, we also include 25 hand-object interaction scenes representing intricate real-world activities. The interactions included are hand activities commonly observed in everyday life. We hope these hand-centric interactions encourage future modeling of complex hand dynamics.

**Long-Duration Sequences**: Although dynamic objects and interaction datasets have covered several long-duration videos, we further provide a long-duration dynamic dataset with 8 sequences of at least 120 seconds. The existing methods have shown fast training speeds for 10s long sequences, but more efficient methods that can operate on longer sequences are needed. Hence, this dataset is aimed at enabling future research in long-duration dynamic neural fields.

**Foreground-Background Segmentation**: Manually segmenting every frame of the sequences is infeasible due to the quantity and view inconsistency. Therefore, we developed a segmentation method using I-NGP [41]. For each frame, we fit an I-NGP model and progressively reduce the bounding box to ensure the model does not render the walls of BRICS. The rendering is applied to the raw data as a segmentation mask, and connected components smaller than a threshold are removed to refine the result. Since the segmentation is generated from I-NGP, the masks are multi-view consistent.

## 5. Benchmarks & Experiments

In this section, we show how DiVa-360 can be used to benchmark dynamic neural field methods using standardized metrics.

### 5.1. Benchmark Comparisons

Our goal is to compare state-of-the-art methods for dynamic neural field reconstruction on our dataset. Specifically, we choose three methods: (1) Per-Frame I-NGP (PF I-NGP) [41], a NeRF model which we train on individual

| Baseline | PSNR↑ | SSIM↑ | LPIPS↓ | JOD↑ | Train (s/f)↓ | Render (s/f)↓ |
|---|---|---|---|---|---|---|
| PF I-NGP [41] | **28.31** ± 3.27 | **0.94** ± 0.03 | **0.08** ± 0.04 | **7.61** ± 0.88 | 48.70 ± 4.40 | **0.94** ± 0.25 |
| MixVoxels [58] | 27.68 ± 2.51 | **0.94** ± 0.03 | 0.09 ± 0.04 | 7.56 ± 0.94 | 57.55 ± 6.96 | 1.48 ± 0.49 |
| K-Planes [15] | 26.39 ± 3.13 | 0.92 ± 0.03 | 0.19 ± 0.07 | 7.18 ± 1.08 | **47.59** ± 5.13 | 3.03 ± 0.20 |

Table 1. We compare the rendering quality and train/render time of PF I-NGP, MixVoxels, and K-Planes for dynamic scenes. Surprisingly, PF I-NGP achieves higher rendering quality and equal or even faster training speed than MixVoxels and K-Planes without directly using temporal information from the adjacent frames.

frames, (2) MixVoxels [58], a state-of-the-art dynamic neural radiance field that uses variation fields to decompose scenes into static and dynamic voxels, and (3) K-Planes [15] which encourages natural decomposition through planar factorization with L1 regularization for space-time decomposition.

**Pre-processing**: We downsample all our sequences to 30 FPS and then segment all frames following Section 4. We split all of our sequences into 5-second chunks (150 frames with 30 FPS, except for PF I-NGP, which has chunk size 1) and then train the above methods per chunk. We select 35 out of 53 cameras for training, excluding cameras from the bottom row of the side panels due to reflections caused by the glass panel in BRICS. We randomly select one camera from each side for a total of 6 cameras for testing. Additionally, we undistort the images with OpenCV [6] and then crop them to $1160 \times 550$.

**Results**: We quantitatively compare the three methods in Table 1. Surprisingly, although PF I-NGP is trained on each frame individually without directly utilizing temporal information, its reconstruction quality is better than MixVoxels and K-Planes in terms of PSNR, SSIM, and LPIPS. However, PF I-NGP suffers from temporal inconsistency (see supplementary Figure 4 and 25). Furthermore, PF I-NGP requires over six times more storage space per time step than MixVoxels or K-Planes (see supplementary Figure 2 and 3). Although MixVoxels is designed for dynamic scenes, its training and inference times are higher than PF I-NGP (with a higher variance). K-Planes has training times similar to PF I-NGP but has significantly longer inference times. We also notice that MixVoxels struggles to capture the dynamic components of the scenes, leading to blurry and noisy reconstruction (see Figure 2). We hypothesize that this is caused by insufficient capacities of the dynamic voxels when there are a lot of dynamic samples. In contrast, K-Planes struggles to capture the static components, such as the background of the scenes, especially in the parts where there is little or no motion. This could be the result of overfitting and contamination from the dynamic planes due to incorrect space-time decomposition.

### 5.2. Experimental Analysis

In theory, temporal information can improve the performance of learning-based methods [18, 52], but benchmark results in Section 5.1 demonstrate that PF I-NGP outperforms MixVoxels and K-Planes. To investigate model sensitivity to temporal information, we split the 30-second long sequences into 2, 3, 6, and 12 chunks and train one dynamic NeRF model per chunk. We find that Mixvoxels' performs roughly the same across different numbers of chunks whereas K-Planes performs better with more temporal information (see supplementary Section 3).

Intuitively, neural fields trained with higher-resolution images should result in better reconstruction quality. To test this, we compare model performances across different resolutions ($1160 \times 550$, $674 \times 320$, $464 \times 220$). Surprisingly, we found that the performance of PF I-NGP remains almost the same whereas MixVoxels and K-Planes perform better at lower resolutions and suffer from blurry details across all resolutions (see supplementary Section 3).

## 6. Conclusion

We have introduced DiVa-360, a real-world 360° dynamic visual dataset that contains synchronized long-duration sequences of table-scale moving objects and interactive scenes. We propose a new BRICS capture system for synchronized long-duration data capture, which also acts as a rich multimodal data capturing system (see supplementary Section 1). DiVa-360 consists of a dynamic dataset of high-resolution, high-framerate, long (5s to 3 mins), and synchronized videos captured simultaneously from 53 RGB cameras within the capture space. In total, DiVa-360 contains 17.4 M images. We benchmark the existing state-of-the-art dynamic neural fields with DiVa-360 dynamic dataset and demonstrate that there is still room for improvement in terms of training and rendering speed, hardware requirement, and imbalance capacity.

## 7. Acknowledgements

This work is supported by NSF grants CAREER #2143576 and CNS #2038897, ONR grant N00014-22-1-259, ONR DURIP grant N00014-23-1-2804, a gift from Meta Reality Labs, an AWS Cloud Credits award, and NSF CloudBank. Arnab Dey is supported by H2020 COFUND program BoostUrCareer under Marie SklodowskaCurie grant agreement #847581. We thank George Konidaris, Stefanie Tellex, Rohith Agaram, and all Brown IVL members.

## References


[1] Henrik Aanæs, Rasmus Jensen, George Vogiatzis, Engin Tola, and Anders Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120, 2016. 3

[2] Sameer Ansari, Neal Wadhwa, Rahul Garg, and Jiawen Chen. Wireless software synchronization of multiple distributed cameras. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–9. IEEE, 2019. 3

[3] 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. *arXiv preprint arXiv:2301.02238*, 2023. 1, 3

[4] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 1, 2

[5] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 1, 3

[6] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 4

[7] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8709–8719, 2019. 2

[8] Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. In *ACM Transactions on Graphics (Proc. SIGGRAPH)*. ACM, 2020. 1, 3

[9] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023. 1, 2, 3

[10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[12] Luca De Luigi, Damiano Bolognini, Federico Domeniconi, Daniele De Gregorio, Matteo Poggi, and Luigi Di Stefano. Scannerf: a scalable benchmark for neural radiance fields. In *Winter Conference on Applications of Computer Vision*, 2023. WACV. 3

[13] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022. 1, 2, 3

[14] Emilien Dupont, Adam Goliński, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. Coin: Compression with implicit neural representations. *arXiv preprint arXiv:2103.03123*, 2021. 1

[15] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023. 1, 2, 3, 4

[16] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 1, 2

[17] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022. 2, 3

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 4

[19] Mustafa Işık, Martin Rünz, Markos Georgopoulos, Taras Khakhulin, Jonathan Starck, Lourdes Agapito, and Matthias Nießner. Humanrf: High-fidelity neural radiance fields for humans in motion. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 2

[20] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339, 2013. 2

[21] Yoonwoo Jeong, Seungjoo Shin, Junha Lee, Chris Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Perfception: Perception using radiance fields. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 3

[22] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 2

[23] 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. 2

[24] Takeo Kanade and PJ Narayanan. Virtualized reality: perspectives on 4d digitization of dynamic events. *IEEE Computer Graphics and Applications*, 27(3):32–40, 2007. 2

[25] Takeo Kanade, Hiroshi Kano, Shigeru Kimura, Atsushi Yoshida, and Kazuo Oda. Development of a video-rate stereo machine. In *Proceedings 1995 IEEE/RSJ International Conference on Intelligent Robots and Systems. Human Robot Interaction and Cooperative Robots*, pages 95–100. IEEE, 1995. 2

[26] 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, 2

[27] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 3

[28] Lingzhi Li, Zhen Shen, Li Shen, Ping Tan, et al. Streaming radiance fields for 3d video synthesis. In *Advances in Neural Information Processing Systems*. 1, 3

[29] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, and Zhaoyang Lv. Neural 3d video synthesis. *CoRR*, abs/2103.02597, 2021. 1, 2, 3

[30] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3

[31] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[32] Kai-En Lin, Lei Xiao, Feng Liu, Guowei Yang, and Ravi Ramamoorthi. Deep 3d mask volume for view synthesis of dynamic scenes. In *ICCV*, 2021. 3

[33] Jia-Wei Liu, Yan-Pei Cao, Tianyuan Yang, Eric Zhongcong Xu, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Hosnerf: Dynamic human-object-scene neural radiance fields from a single video. *arXiv preprint arXiv:2304.12281*, 2023. 2

[34] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, 2019. 1, 2

[35] Chongshan Lu, Fukun Yin, Xin Chen, Wen Liu, Tao Chen, Gang Yu, and Jiayuan Fan. A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7557–7567, 2023. 1, 2, 3

[36] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 1, 2, 3

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[38] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1

[39] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019. 3

[40] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3

[41] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022. 1, 2, 3, 4

[42] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 1, 2

[43] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[44] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2, 3

[45] 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. *ACM Trans. Graph.*, 40(6), 2021. 1, 2, 3

[46] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 2

[47] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *arXiv preprint arXiv:2011.13961*, 2020. 2

[48] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *International Conference on Computer Vision*, 2021. 3

[49] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[51] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3

[52] Javier Selva, Anders S Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B Moeslund, and Albert Clapés. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 4

[53] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 2

[54] 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. 1, 2, 3

[55] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. *arXiv*, 2022. 1, 3

[56] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Wang Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. Advances in neural rendering. In *Computer Graphics Forum*, 2022. 1

[57] Christian Theobalt, Marcus A Magnor, Pascal Schüler, and Hans-Peter Seidel. Combining 2d feature tracking and volume reconstruction for online video-based human motion capture. *International Journal of Image and Graphics*, 4(04):563–583, 2004. 2

[58] 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. 1, 2, 3, 4

[59] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction, 2022. 1

[60] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Wang Xinggang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 1, 2, 3

[61] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, Dahua Lin, and Ziwei Liu. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[62] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9421–9431, 2021. 1

[63] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11097–11107, 2020. 3

[64] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 2022. 1, 2

[65] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia Conference Proceedings*, 2023. 1, 3

[66] Zhen Xu, Sida Peng, Haotong Lin, Guangzhao He, Jiaming Sun, Yujun Shen, Hujun Bao, and Xiaowei Zhou. 4k4d: Real-time 4d view synthesis at 4k resolution. 2023. 2

[67] Zhiwen Yan, Chen Li, and Gim Hee Lee. Nerf-ds: Neural radiance fields for dynamic specular objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8285–8295, 2023. 1, 2, 3

[68] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. *Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[69] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2

[70] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020. 1, 2, 3

[71] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR2021)*, 2021. 2

[72] Zhixuan Yu, Jae Shin Yoon, In Kyu Lee, Prashanth Venkatesh, Jaesik Park, Jihun Yu, and Hyun Soo Park. Humbi: A large multiview dataset of human body expressions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2990–3000, 2020. 2

[73] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 813–822, 2019. 2