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

Anonymous Author
Institution
author@i1.org

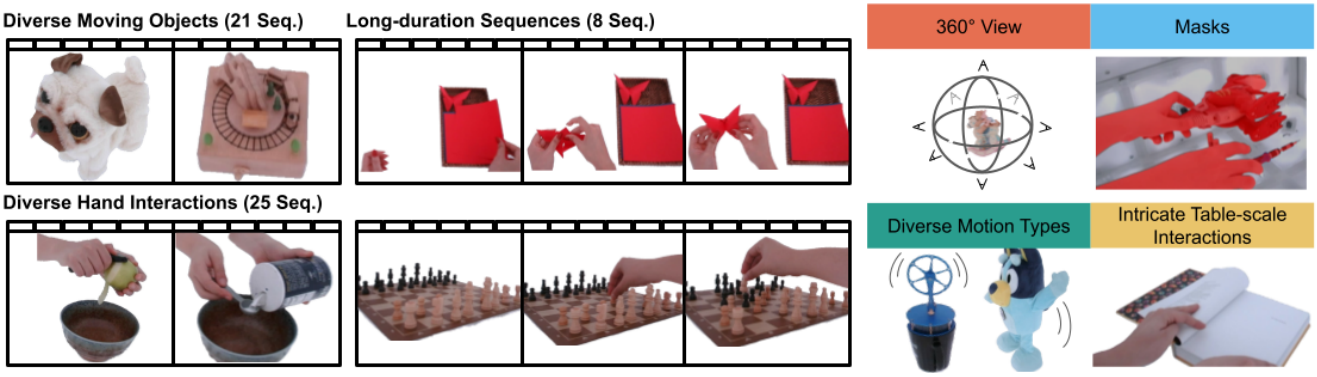

Figure 1. DiVa-360 is a **real-world** 360° multi-view visual dataset of dynamic tabletop scenes captured using a customized low-cost capture system consisting of 53 RGB cameras. The DiVa-360 dataset consists of 21 diverse moving object sequences, and 33 intricate hand-object interaction sequences, including 8 long-duration sequences (2-3 mins). DiVa-360 provides (1) 360° coverage of dynamic scenes, (1) foreground-background segmentation masks, synchronized audio, and detailed text descriptions, and (3) diverse table-scale scenes with intricate motions (*e.g.*, repairing objects). The goal of DiVa-360 is to facilitate research in dynamic long-duration neural fields.

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 with 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 [69], or neural implicit representations, have recently emerged as useful representations in computer vision, graphics, and robotics [60, 69] for capturing properties such as radiance [4, 5, 27, 43, 44], shape [32, 41, 45, 46, 64, 74], and dynamic motion [9, 17, 31, 36, 38, 48, 63, 65, 67]. Their high fidelity, continuous representation, and implicit compression [15] 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. Consider this: we can easily watch hours-long 2D videos, a task that cannot yet be achieved efficiently with 3D neural fields due to long training times [3, 10, 16, 27, 30, 31, 38, 58, 63, 65]. 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 [9, 14, 31, 32, 37, 48, 59, 70, 72, 75], they are limited to room-scale scenes or specific categories like humans [20, 21, 23, 35, 49, 76], or are captured with monocular or forward-facing cameras that do not always provide sufficient multi-view cues for immersive reconstruction [18, 32, 37, 48]. Furthermore, most of the sequences in these datasets [14, 31, 32, 37, 48, 72, 75] 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-

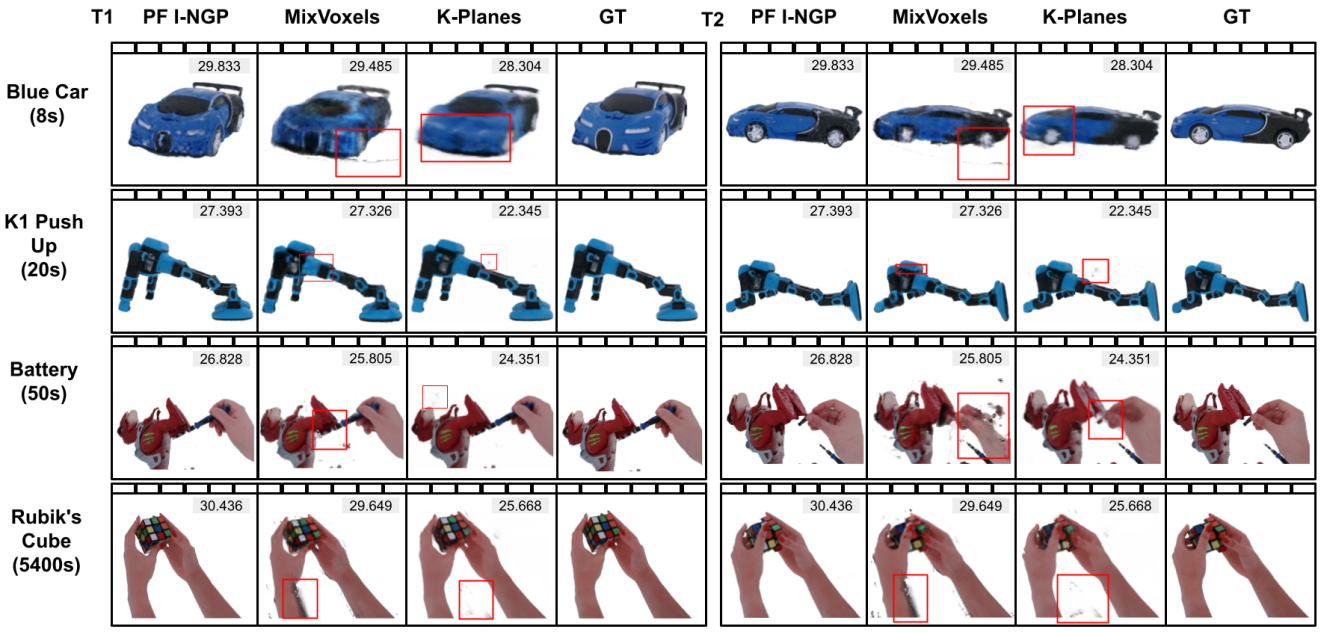

Figure 2. DiVa-360 provides multi-view dynamic sequences for dynamic neural field methods. The dataset contains a variety of object and motion types. Here, we showcase reconstruction results across time steps from PF I-NGP [44], MixVoxels [63], and K-Planes [16] trained on our dataset. Surprisingly, the rendering results of PF I-NGP, a method that does not directly utilize temporal information from adjacent frames, are better than those of MixVoxels and K-Planes. MixVoxels struggles with complex motion data, such as hands, while K-Planes suffers from floaters in the background. We demonstrate more visualization results in the supplementary Section 3.

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 simultaneously from 53 RGB cameras spanning 360° volume within the capture space. We provide 46 dynamic sequences, including 21 object-centric sequences categorized by different motion types and 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 of 53 dynamic scenes over **2738** seconds. We also provide foreground-background segmentation masks, accompanying audio data from microphones, and detailed text descriptions of the activity observed.

Capturing such large-scale data requires advances in capture systems, as well as benchmarking metrics. We have built a new low-cost capture system called **TRICS (Temporal 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 [16, 44, 63]. We perform a systematic analysis of current methods and characterize their performance on different sequence durations, image resolutions, motion types, and viewpoints. Surprisingly, we observe that methods that model each frame in a dynamic sequence without directly using temporal information [44] outperform state-of-the-art dynamic methods [16, 63] in terms of reconstruction quality and even training speed (see Figure 2). In addition, existing methods [16, 63] are biased toward moving objects' shapes, thereby losing high-frequency information and fine-grained details. Finally, existing state-of-the-art methods [16, 63] prefer different amounts of temporal information, with one [16] starting to outperform another [63] after acquiring more temporal information. To summarize, we make the following contributions:

- **TRICS**: 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.

Dataset	Real	Mask	360° view	Multiview	Object	Scene
HyperNeRF [48]	✓	✗	✗	✗	✗	✓
OMMO [37]	✓	✗	✗	✗	✗	✓
Block-NeRF [59]	✓	✗	✗	✓	✗	✓
Eyeful Tower [70]	✓	✗	✗	✓	✗	✓
DyNeRF [31]	✓	✗	✗	✓	✗	✓
ILFV [9]	✓	✗	✗	✓	✗	✓
Deep3DMV [34]	✓	✗	✗	✓	✗	✓
NeRF-DS [72]	✓	✓	✗	✓	✗	✓
NDSD [75]	✓	✓	✗	✓	✗	✓
D-NeRF [50]	✗	—	✓	✗	✓	✗
Objaverse [14]	✗	—	✓	✓	✓	✓
DiVa-360	✓	✓	✓	✓	✓	✓

Table 1. We compare featured properties of our DiVa-360 with other object-centric and scene-centric datasets. DiVa-360 is a unique dataset that contains real-world 360° multiview object-centric and scene-centric data with foreground-background masks.

Dataset	#Camera	FPS	#Scenes	#Frames	Average length (s)
Objaverse [14]	—	—	3k	—	—
NeRF-DS [72]	2	—	8	10.2k	—
Block-NeRF [59]	12	10	1	12k	100*
HyperNeRF [48]	1	15	17	13.8k	27
OMMO [37]	1	—	33	14.7k	—
Eyeful Tower [70]	22	< 4	11	28.6k	**2k**†
DyNeRF [31]	18	30	6	37.8k	10
ILFV [9]	46	30	15	270.4k	13
Deep3DMV [34]	10	**120**	**96**	3.8M	33
DiVa-360	**53**	**120**	54	**17.4M**	51

Table 2. Specifications of our DiVa-360 dataset and other dynamic datasets. ∗ indicates that despite Block-NeRF consisting of a 100s-long video, it is made up of numerous transient street scenes, each with restricted view coverage. † indicates that although the average length of Eyeful Tower is 2000 s, the FPS is less than 4. Our DiVa-360 dataset is the largest visual dataset for dynamic neural fields captured at 120 FPS with an average video length of 51 s.

2. Related Work

Neural Fields: Neural fields, or coordinate-based implicit neural networks, have generated considerable interest in computer vision [69] because of their ability to represent geometry [12, 40, 46] and appearance [36, 43, 57]. Neural radiance fields (NeRF) [43] and its variants [4, 33, 45, 63, 74] uses a multilayer perceptron (MLP) to model density and color, leading to photorealistic novel view synthesis and 3D reconstruction. Since the training cost of NeRFs is high, several methods have tried to address this limitation [11, 27, 44, 53]. Naturally, some approaches have also turned their focus towards dynamic neural fields [10, 16, 17, 31, 32, 36, 38, 47, 48, 50, 58, 63, 65, 71]. However, these methods have thus far been limited to only brief sequences, partly as a result of the unavailability of long-duration datasets. Our work enables further research in long-duration dynamic neural field research with a more comprehensive and richer dataset with 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 [26], which were later extended to capture a hemispherical volume [25] with up to 50 cameras for 3D and 4D reconstruction using non-machine learning techniques [61]. The focus of most existing multi-camera capture systems has been on room-scale scenes for human or environment capture [24, 24, 77]. While some table-scale datasets exist, notably for hand interaction capture [8, 79], they have only a limited number of cameras. In contrast, our TRICS system is specially designed for dense 53-view visual capture of table-scale scenes, and our sequences showcase intricate interactions (*e.g.*, small tool use) in high fidelity.

Datasets for Dynamic Neural Fields: While plenty of datasets exist for NeRF methods [1, 5, 13, 22, 29, 42, 43, 51, 66, 73] their focus has been on static scenes. For dynamic scenes, numerous datasets such as DyNeRF [31], NDSD [75], ILFV [9], NeRF-DS [72], and Deep3DMV [34] exist, but they are limited to only a short duration (~15s), or have only forward-facing cameras preventing them from enabling 360° capture. BlockNeRF [59] provides street view videos incorporating dynamic elements but lacks a focus on objects, and does not provide many views. This creates fleeting scenes that do not encompass full 360° camera coverage.

Eyeful Tower [70] provides dynamic data up to 2000 s long, but the framerate is less than 4 FPS. Monocular videos of human faces [47, 48], human activities [32], or outdoor scene [37] have been used for neural field reconstructions, but a single camera restricts visibility resulting in low effective multi-view factors (EMF)[18]. While datasets like Objaverse [14] and SAPIEN [68] provide articulated objects, they are not sourced from the real world. Our dataset stands out by offering a 360° view of real-world long dynamic sequences with objects and hand-object interaction captured by 53 synchronized cameras (see Table 1 and Table 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. Temporal Interaction Capture System (TRICS)

Our goal is to capture long-duration sequences of table-scale objects and interactions to enable further research in high-fidelity dynamic neural fields. To achieve this, we need a hardware and software system that can capture high-framerate, high-resolution video and have the capability to synchronize and calibrate these sensor streams. While commercial products exist for this purpose, they are expensive and do not meet our requirements. We therefore designed and built our own hardware and software solution which we call the **Temporal Interaction Capture System (TRICS)**.

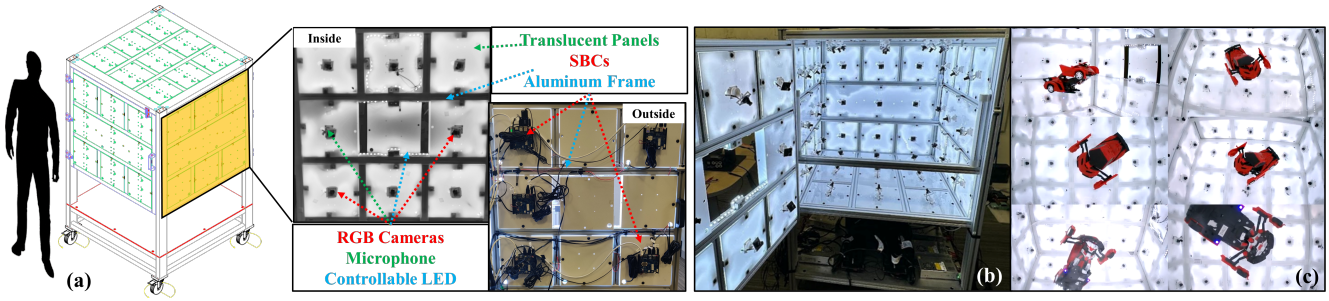

Figure 3. (a) TRICS is a refrigerator-sized aluminum frame that supports a 1 m^3 capture volume mounted on wheels for mobility. Each side wall of the capture volume is divided into a 3×3 grid, with each grid square containing sensors, LEDs, single-board computers (SBCs), and light diffusers. (b) Two walls of the capture volume act as doors for easy access to the capture volume. (c) We can acquire 360° RGB views of dynamic objects and intricate hand-object interactions in this capture volume (6 views shown).

Figure 3 shows our hardware system for capturing synchronized data.

TRICS Hardware: Our system uses a mobile aluminum frame, housing a 1 m^3 capture volume outfitted with sensor panels across a 3x3 grid on each of its six sides (Figure 3 (a)). These panels consist of RGB cameras, microphones, and LED light strips, which together create a versatile and uniformly light environment. For 360° capture, we installed a transparent shelf on which objects are placed. The system is designed to handle large data output through a custom communication setup that compresses and transmits data to a high-capacity control workstation. This design, combining portability, comprehensive capture capabilities, and efficient data management, allows for dynamic, 360° view capturing with low latency.

TRICS Software: While our hardware allows the capture of large-scale rich multi-view data, controlling and calibrating the cameras, and synchronizing and managing data requires specialized software. For camera and microphone synchronization, we adopt network-based synchronization [2] with an accuracy of 2–3 ms. For camera calibration, during each capture session, we affix transparent curtains with ArUco markers to the walls. Using COLMAP [54, 55], we generate camera poses for the 53 cameras. The camera poses are further refined using I-NGP's [44] dense photometric loss for improved reconstruction quality. Finally, we also built software for efficiently transferring terabytes of data from the control workstation to cloud storage. The DiVa-360 dynamic dataset contains synchronized long-duration videos of both moving objects and intricate hand interactions. Our goal is to make this dataset useful for learning long-duration dynamic neural fields of appearance - existing methods [3, 10, 16, 30, 31, 38, 58, 63, 65] have been limited to only short durations, usually around 10 seconds. Instead of just using 10 s clips of the sequences, we fully benchmark all sequences in our dynamic dataset that contains 21 object sequences categorized by different motion types, 25 hand-object interaction sequences including human daily activity, and 8 long-duration sequences with rich information. In total, DiVa-360 dynamic dataset contains **17.4 M** image frames of 53 dynamic scenes over **2738** seconds. In addition, our data also contains masks for foreground-background segmentation. Although not the focus of this work, we optionally provide synchronized audio and text descriptions for all sequences. 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 (see Figure 4). To be representative of real-world motions, we chose objects with different types of motion (see supplementary Section 3): (1) Slow motion: objects that perform slow, continuous motions, *e.g.*, music box and rotating world globe. (2) Fast motion: objects that move or transform drastically, *e.g.*, remote control cars and dancing toys. (3) Detailed motion: objects that perform precise small motions, *e.g.*, a clock. (4) Repetitive motion: objects that repeat the same motion pattern, *e.g.*, Stirling engine and toys that sway left and right. (5) Random motion: objects that perform indeterministic motions, *e.g.*, a toy that creates random patterns within a sphere.

Interactions: In addition to dynamic objects, we also include 25 hand-object interaction scenes representing intricate real-world activities (see supplementary Section 3). The interactions included are hand activities commonly observed in everyday life (see Figure 4), such as flipping a book, replacing a toy's batteries, and opening a lock. The purpose of these hand-centric interaction data is to check whether the dynamic neural fields can generalize well to more complicated motion with occlusion from hands. 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 (see supplementary

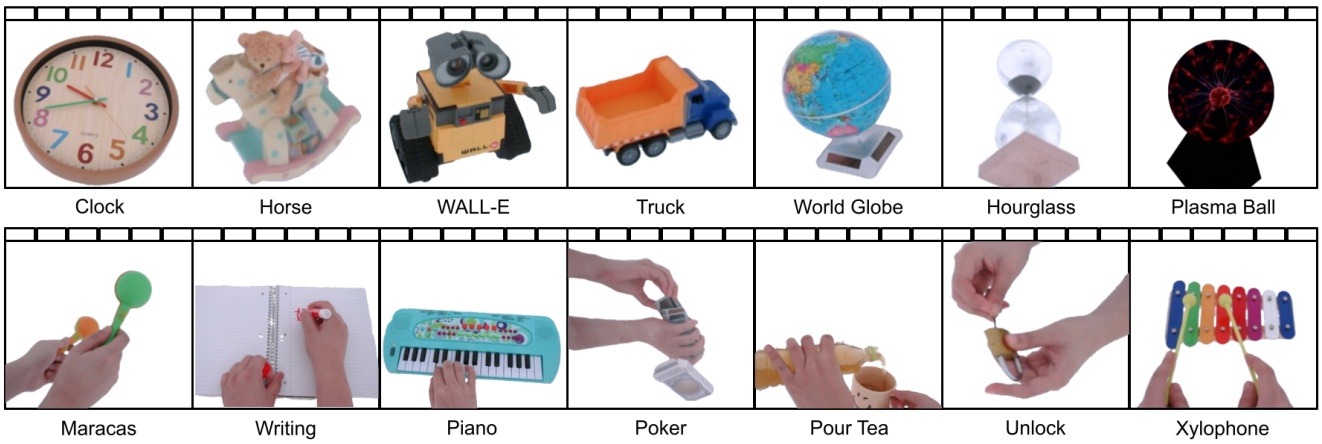

Figure 4. DiVa-360 covers diverse object and hand-object interaction data. Our object sequences represent a variety of motion types, while our hand-object interaction data contain intricate and realistic motions.

Section 3). 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: A major challenge with neural radiance fields is the segmentation of foreground objects from background clutter. Manually segmenting every frame is infeasible due to the quantity and view inconsistency. Therefore, we developed a segmentation method using I-NGP [44]. As preparation, we manually segment the foreground object in the first frame of one scene and train an I-NGP model on segmented images to refine coarse camera poses extracted from COLMAP [54, 55]. The refined pose is used for all downstream tasks. For each frame, we fit an I-NGP model that optimizes camera poses, lens distortion, and image latent vector. The model's bounding box is then progressively reduced to remove background clutter. We then render trained I-NGP as binary masks. To further refine the masks, we removed connected components smaller than a threshold. Segmenting with this method is possible because all objects are placed around the center of TRICS. 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, we analyze the effect of critical parameters on these methods, and justify the need for our dataset. These experiments were performed on Nvidia GPUs (RTX 3090, A5000) and involved **over 500 GPU-days** of training and inference.

5.1. Benchmark Comparisons

Our goal is to compare and contrast state-of-the-art methods for dynamic neural field reconstruction on our dataset. Specifically, we choose to compare three methods: (1) Per-Frame I-NGP (PF I-NGP) [44], a NeRF model which we train on individual frames in all 54 sequences, (2) MixVoxels [63], a state-of-the-art dynamic neural radiance field which uses variation fields to decompose scenes into static and dynamic voxels, and (3) K-Planes [16] which encourages natural decomposition through planar factorization with L1 regularization for space-time decomposition. We plan to include very recent work such as [38, 65] in the future.

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 best cameras for training and hold out 6 cameras for testing. Specifically, we eliminate the cameras from the bottom row of the side panels due to reflection caused by the glass panel in TRICS and randomly select one camera from each panel as test views. We note that current neural radiance fields cannot handle reflections from the glass panel on which we place the objects for 12 views in our data. Considering that not every NeRF model supports the camera distortion factor, we undistort the images with OpenCV [7] and crop the images to the same size (1160×550) after undistortion.

Metrics: We use (a) Peak Signal-to-Noise Ratio (PSNR), (b) Structural Similarity Index Measure (SSIM) [52, 62], and (c) Learned Perceptual Image Patch Similarity (LPIPS) [78] to measure the rendering quality, and Just Objectionable Difference (JOD) [39] to measure the visual difference between rendered video and ground truth, along with per-frame training/rendering time (in seconds) for 6 testing views.

Baseline	PSNR↑	SSIM↑	LPIPS↓	JOD↑	Train (s/f)↓	Render (s/f)↓
PF I-NGP [44]	**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 [63]	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 [16]	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 3. 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.

Results: We quantitatively compare the three methods in Table 3. Surprisingly, although PF I-NGP is trained on each frame individually without directly utilizing temporal information, its reconstruction quality is better than both MixVoxels and K-Planes in terms of PSNR, SSIM, and LPIPS. However, PF I-NGP suffers from temporal inconsistency, which is especially obvious for static parts (see Figure 7 and supplementary Figure 21). Furthermore, MixVoxels only requires 2.7-4.7 MB storage space per time step, and K-Planes requires 2 MB, both of which are over six times smaller than PF I-NGP's 29 MB. 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. Besides, we also notice that MixVoxels struggles to capture the dynamic components of the scenes, leading to blurry and noisy reconstruction (see Figures 2 and 7). 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

The goal of this section is to identify whether the dynamic neural fields are sensitive to temporal information and spatial information. For the experiment, we select sequences longer than 30 seconds from the object and interaction dataset. We then use the first 30 seconds (900 frames) of these sequences for the following experiments.

Temporal Information: In theory, temporal information can improve the performance of learning-based methods [19, 56]. However, benchmark results in Section 5.1 demonstrate that PF I-NGP outperforms MixVoxels and K-Planes. To further investigate how sensitive these methods are to the temporal information, we split the 30-second long sequences into 2, 3, 6, and 12 chunks and train one dynamic NeRF model per chunk. Figure 5 shows that Mixvoxels performs slightly better when trained with less temporal information (more chunks), but its performance remains roughly the same across different numbers of chunks. The

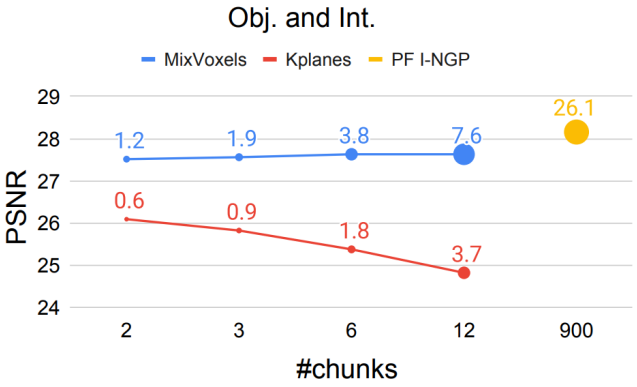

Figure 5. The rendering quality across different numbers of chunks with object and interaction data. The circle dot presents the storage space of the models in GB. MixVoxels prefers less temporal information, while K-Planes prefers more temporal information.

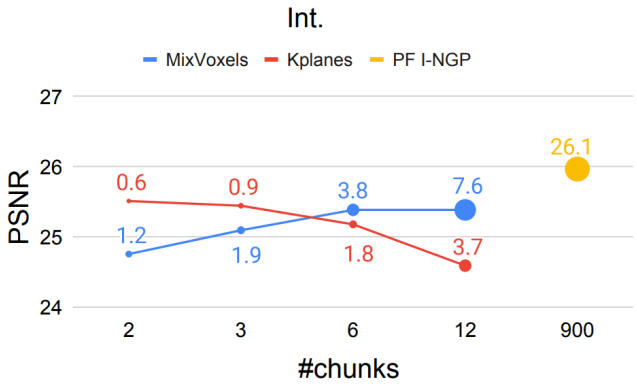

Figure 6. The rendering quality across different numbers of chunks with interaction data only. The circle dot presents the storage space of the models in GB. K-Planes outperforms MixVoxels with more temporal information on complex motion data.

dynamic branch of MixVoxels may not have sufficient capacity to handle more dynamic samples. Unlike MixVoxels, K-Planes is more sensitive to sequence lengths and performs better with more temporal information. Through the rendering results, we found that fewer chunks also mitigate the overfitting problem of K-Planes on DiVa-360 (see supplementary Figure 12). One interesting finding is that although

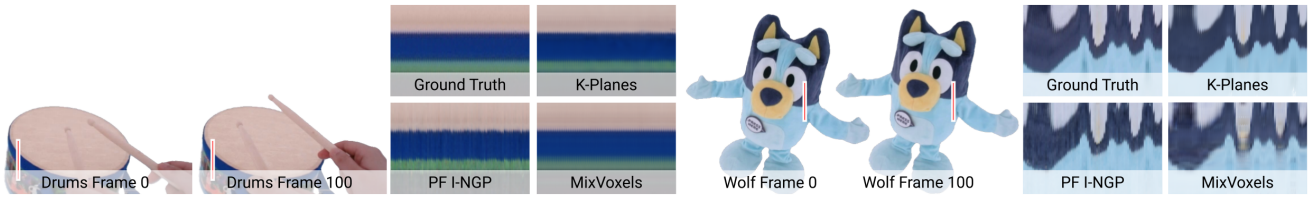

Figure 7. Visualization of temporal consistency in the same views by concatenating pixels from the same line across time steps. If a method is temporally consistent, its figure should be smooth horizontally and similar to the ground truth. PF I-NGP is less consistent across time, especially for static parts (e.g., drum), while MixVoxels is noisier on the dynamic parts (e.g., wolf's body).

MixVoxels outperforms K-Planes with object and interaction data, K-Planes outperforms MixVoxels with 2 and 3 chunks setting on interaction data (see Figure 6). This indicates that K-Planes can better handle the more complex motions in the interaction data when provided with more temporal information.

Spatial Information: Intuitively, neural fields trained with higher-resolution images should result in better reconstruction quality. To test this hypothesis, we conduct comparisons of model performances across different resolutions. In this experiment, we train the three methods on lower resolutions by downscaling the training set from 1160×550 (undistorted images) to 674×320 and 464×220. After training, we evaluate the trained models by rendering test views at the original resolution (see Table 4 and supplementary Figure 13). To our surprise, we found that the performance of PF I-NGP remains almost the same, and rendering results have apparent fine-grained details. Furthermore, MixVoxels and K-Planes perform better at lower resolution, with MixVoxels performing the best at 674×320 and K-Planes performing the best at 464×220. Both methods suffer from similarly blurry details across all resolutions. This interesting result contradicts our hypothesis, and there could be several reasons for it. First, NeRFs have shown impressive spatial interpolation ability leading to only a minimal drop in performance with reducing resolution. Second, under the same training setting, NeRFs will revisit the training samples more frequently when trained on lower-resolution images and thus reconstruct these samples better. Finally, dynamic NeRFs need to spend much more of their capacities to capture moving objects which could result in insufficient capacity to capture fine-grained details. In conclusion, we suspect that current dynamic methods cannot efficiently utilize spatial information in high-resolution images as they are biased toward motion and misses the high-frequency details presented in the images. In addition, we note that human perception of these images may not match the observed quantitative results [6].

Spatial and Temporal Information: In the previous experiment, we only have one control variable, spatial resolution or temporal length. It is unclear whether the same conclusion will hold if we change spatial resolution and temporal length simultaneously. Hence, in this experiment, we change spatial resolution and temporal length simultaneously while maintaining a similar size of the 3D volume (width x height x temporal length). Table 5 shows MixVoxels performs worst, and K-Planes performs best with the lowest spatial resolution and longer temporal length setting (464×220, 900 frames). This matches our findings in the previous experiments. If we check Table 4 and Table 5 together, the phenomenon is more obvious. The PSNR of MixVoxels drops from 27.75 to 27.2 and 26.91 to 26.1 for spatial resolution 674×320 and 464×220, respectively, after including more temporal information. The PSNR of K-Planes increases from 25.57 to 26.04 and 26.03 to 26.19 for spatial resolution 674×320 and 464×220, respectively, after including more temporal information.

Baseline	**1160×550**	**674×320**	**464×220**
PF I-NGP [44]	28.16	28.19	28.15
MixVoxels [63]	27.63	27.75	26.91
K-Planes [16]	25.38	25.57	26.03

Table 4. The PSNR of each baseline across different resolutions. During testing, we interpolate the images to 1160×550 resolution. The performance of PF I-NGP remains similar. MixVoxels and K-Planes get slightly better performance with the low-resolution training set.

Baseline	**1160×550, 6 ch.**	**674×320, 2 ch.**	**464×220, 1 ch.**
MixVoxels [63]	**27.63**	**27.20**	26.10
K-Planes [16]	25.38	26.04	**26.19**

Table 5. The PSNR of each baseline across different spatial resolutions and temporal lengths. MixVoxels reaches the worst performance, and K-Planes reaches the best performance with the lowest spatial resolution and fewest chunks setting (464×220, 1 chunk).

5.3. Dataset Justification

In this section, we justify the need for our 360° views with 53 cameras, and other design choices.

Number of Cameras: For evaluating the number of cameras, we compare three settings: (1) *All-view*, which follows the original setting with all cameras, (2) *Forward*, which only uses the cameras from two adjacent side panels, resulting in 10 cameras, (3) *Multi-view*, which uses two cameras per side panel and one camera from top and bottom panel, resulting in 10 cameras.

Both quantitative and qualitative results (see Table 6 and supplementary Figure 15-16) demonstrate that *All-view* outperforms *Multi-view*, indicating that more cameras improve the rendering quality of NeRFs. In addition, *All-view* and *Multi-view* outperform *Forward* when tested on an occluded view, suggesting that multi-view 360° is better than forward-facing settings for benchmarking.

Baseline	All-view	Forward	Multi-view
PF I-NGP [44]	28.16 / 27.24	24.77 / 23.99	24.07 / 25.37
MixVoxels [63]	27.63 / 27.43	20.51 / 15.80	23.65 / 24.02
K-Planes [16]	25.38 / 24.55	23.23 / 22.40	22.64 / 22.89

Table 6. The PSNR of testing views / occluded views across different settings of the capture system. The PSNR of *All-view* is higher than *Multi-view*. Hence, more cameras can help NeRFs. The PSNRs of *All-view* and *Multi-view* are higher than *Forward* on occluded view, indicating that multi-view 360° settings are better than forward-facing settings.

Foreground-Background Segmentation Method: In Section 4, we mention that we use I-NGP to segment each frame. Although the segmentation model can be replaced with improved models in the future, we believe that our current method is suitable for DiVa-360, especially due to its multiview consistency. To validate the performance of I-NGP segmentation, we compare it against Segment Anything (SAM) [28] in terms of segmentation quality and multiview consistency. For this benchmark, we manually segment one frame of 6 random views from all scenes as ground truth and compute mean intersection over union (mIoU) for images segmented by SAM and our method.

According to Table 7, I-NGP segmentation reaches better mIoU and lower average standard deviation over six views on DiVa-360. In addition, the visualization results (see supplementary Figure 18 and 19) also support the statement that the performance of I-NGP segmentation is more multiview consistent.

Baseline	Obj. and Int.	Obj.	Int.
I-NGP Seg. [44]	0.926 / 0.048	0.962 / 0.016	0.901 / 0.071
SAM [28]	0.919 / 0.086	0.955 / 0.042	0.885 / 0.118

Table 7. The mean intersection over union (mIoU) / average standard deviation of mIoU over six views. I-NGP segmentations outperform SAM on DiVa-360. In addition, a lower standard deviation indicates more equal quality across views.

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 TRICS capture system for synchronized long-duration data capture, which also acts as a rich multimodal data capturing system (see supplementary Section 2). 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, imbalance capacity, temporal information, and spatial information.

Limitations and Future Work: Although TRICS can also act as a multimodal capturing system, our current metrics and evaluation are limited to images – in future work, we will consider metrics for audio and text (see supplementary Section 8). TRICS cannot capture scenes larger than table-scale – we plan to expand the capture system to larger volumes in the future. Due to the training speed of the existing state-of-the-art methods, we cannot include more baselines or metrics for longer videos. Hence, we hope to include more long sequences and baselines in the future [65, 71].

Societal/Ethical Impact: Our dataset does not reveal any private information and presents limited means for misuse. However, future extensions of our work could contain private information that can be misused. Another possible impact is the environmental cost since the total GPU running days of training, rendering, and experiments are at least 500 days. Thus, we will release pretrained weights after publishing so that others may avoid re-training.

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. 4

[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, 4

[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, 3

[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] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6228–6237, 2018. 7

[7] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 5

[8] 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. 3

[9] 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

[10] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *CVPR*, 2023. 1, 3, 4

[11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[13] 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

[14] 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, 3

[15] 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

[16] 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, 5, 6, 7, 8

[17] 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, 3

[18] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022. 1, 3

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 6

[20] 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. 1

[21] 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. 1

[22] 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

[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 *The IEEE International Conference on Computer Vision (ICCV)*, 2015. 1

[24] 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. 3

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

[26] 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. 3

[27] 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, 3

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 8

[29] 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

[30] 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, 4

[31] 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, 3, 4

[32] 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, 3

[33] 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. 3

[34] 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

[35] 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. 1

[36] 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, 3

[37] 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, 3

[38] 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, 3, 4, 5

[39] Rafał K Mantiuk, Gyorgy Denes, Alexandre Chapiro, Anton Kaplanyan, Gizem Rufo, Romain Bachy, Trisha Lian, and Anjul Patney. Fovvideovdp: A visible difference predictor for wide field-of-view video. *ACM Transactions on Graphics (TOG)*, 40(4):1–19, 2021. 5

[40] 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. 3

[41] 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

[42] 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

[43] 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, 3

[44] 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, 5, 6, 7, 8

[45] 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, 3

[46] 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, 3

[47] 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. 3

[48] 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, 3

[49] 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. 1

[50] 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. 3

[51] 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

[52] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 5

[53] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 3

[54] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 4, 5

[55] 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. 4, 5

[56] 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. 6

[57] 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. 3

[58] 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, 3, 4

[59] 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

[60] 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

[61] 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. 3

[62] Paul Upchurch, Noah Snavely, and Kavita Bala. From a to z: Supervised transfer of style and content using deep neural network generators. *CoRR*, abs/1603.02003, 2016. 5

[63] 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, 5, 6, 7, 8

[64] 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

[65] 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, 3, 4, 5, 8

[66] 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

[67] 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

[68] 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

[69] 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, 3

[70] 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

[71] 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. 3, 8

[72] 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, 3

[73] 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

[74] 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, 3

[75] 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, 3

[76] 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. 1

[77] 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. 3

[78] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[79] 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. 3