Replay: Multi-modal Multi-view Acted Videos for Casual Holography

Roman Shapovalov* Yanir Kleiman* Ignacio Rocco* David Novotny
Andrea Vedaldi Changan Chen† Filippos Kokkinos Ben Graham Natalia Neverova
Meta UT Austin† *equal contribution

https://replay-dataset.github.io/

Abstract

We introduce Replay*, a collection of multi-view, multi-modal videos of humans interacting socially. Each scene is filmed in high production quality, from different viewpoints with several static cameras, as well as wearable action cameras, and recorded with a large array of microphones at different positions in the room. Overall, the dataset contains over 4000 minutes of footage and over 7 million timestamped high-resolution frames annotated with camera poses and partially with foreground masks. The Replay dataset has many potential applications, such as novel-view synthesis, 3D reconstruction, novel-view acoustic synthesis, human body and face analysis, and training generative models. We provide a benchmark for training and evaluating novel-view synthesis, with two scenarios of different difficulty. Finally, we evaluate several baseline state-of-the-art methods on the new benchmark.*

1. Introduction

A staple of science fiction is to relive past events and memories as holograms. With technological advances in virtual, mixed and augmented reality, this vision is ever closer to become a reality. We can now think of recording an event with a pair of AR glasses instead of a camera, and relive it later as a 360° re-projection in a real or virtual space. However, there are still major technical hurdles before this can be done reliably and with sufficient quality.

High-fidelity 3D reconstruction remains one of the primary obstacles. Given a casual recording from a single sensor like a pair of AR glasses, it is in general not possible to reconstruct the content in 3D. Monocular data, or even data collected from cameras with a short baseline, simply does not contain sufficient information for 360° reconstruction. For example, in such setup, it is not possible to observe simultaneously the front and back of an object. Furthermore, reconstructing appearance is not enough: any engaging user experience also requires to reconstruct sounds, so the problem is inherently multi-modal.

Consumer holography requires to compensate for the intrinsic limitations of a casual data capture setup via machine learning. However, despite the success of neural rendering [40], even the bests method struggle to reconstruct complex, long dynamic content from a monocular sensor. Furthermore, none of these approaches tackles multi-modal reconstruction yet.

Here, we suggest that further progress in casual holography, and in general in the reconstruction and generation of realistic 4D (3D + time) multi-modal content, is severely hampered by the lack of suitable datasets. We address this gap by introducing *Replay*, a new large dataset to study the problem of multi-modal new-view synthesis for long captures of acted dynamic content. Replay contains long scenes in a natural indoor environment (living room, dining room, *etc*.), where multiple people are interacting with props and with each other and performing a variety of activities such as exercising, playing games, or chatting. Each scene is several minutes long, and is filmed in 4K resolution with 8 static DSLR cameras and 3 head-mounted GoPro cameras that capture the scene from all view points, allowing the evaluation of scene reconstruction from the view points that significantly differ from the source video. For each scene, we also provide a semi-static *fly-around* sequence, where the actors pause and remain still while the head-mounted camera operators walk around them. In addition, the scene is recorded with a large array of microphones to allow novel view acoustic synthesis [9]. All sensors are temporally calibrated, and cameras are also color- and view-calibrated as well. In addition, metadata such as foreground segmentation masks is provided for some of the scenes. The data is collected with actors' consent, addressing privacy concerns, and will be public for non-commercial research.

This paper focuses on the visual component of Replay; the audio part of the dataset is introduced and used for novel-view *acoustic* synthesis by Chen et al. [9]. The Replay videos, in turn, constitute a notable step up compared to existing datasets for static and dynamic novel view synthesis; so far methods have been evaluated on short sequences with a limited range of view points. For exam-

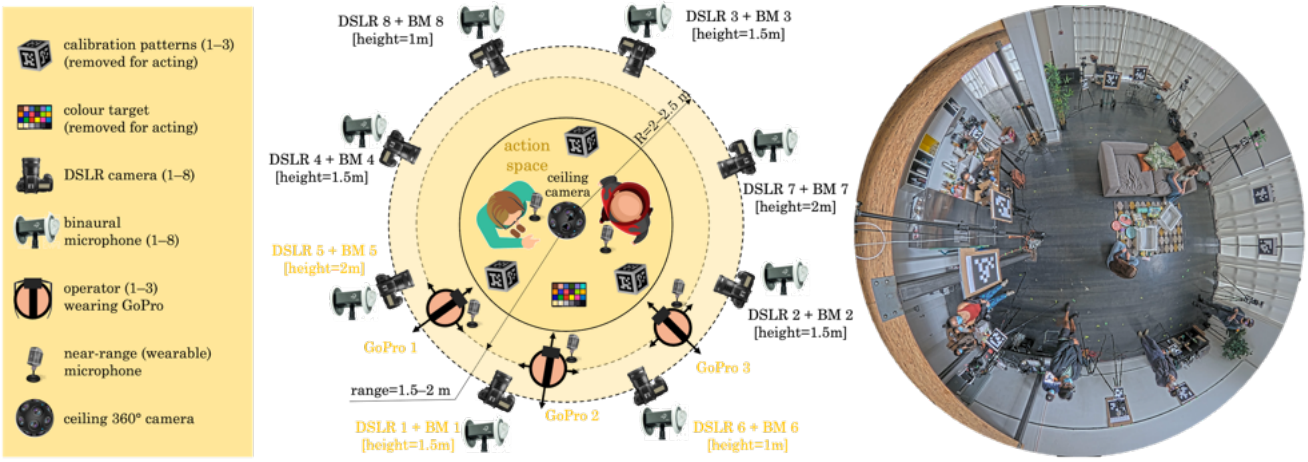

Figure 1. Recording setup (left) and a frame from the ceiling camera during capture (right). Actors wearing near-range microphones are located in the centre of the scene; they are surrounded by a ring of static DSLR cameras paired with binaural microphones. Operators wearing GoPros and microphones are standing in front of the actors at around 2 metres distance during the acting phase, and during the fly around phase they go around the scene filming semi-static actors. The colour of sensor labels reflects their usage in new-view synthesis benchmarks (Section 5): fly around is trained using GoPro-2 frames while evaluated on all DSLRs; we define an 'acting benchmark' using the 6 frontal sensors (gold), of which DSLR-1 is held out for testing, and the other 5 sensors are used for training.

ple, the popular Dynamic Scene Dataset [68], which is often used to evaluate dynamic new-view synthesis, contains short scenes (≈5 sec) sampled at low FPS (30 frames in total), and captured by static cameras where the farthest two cameras are about one meter apart. Other datasets such as ZJU-Mocap [41] and AIST++ [29], provide longer videos with a large variety of view points, but are human-centric and contain people with an empty background and no additional objects, which makes them less useful for evaluating full scene reconstruction. None of these datasets contain naturally-acted events with sounds.

Due to the richness of scenes, actors, sensors and modalities, Replay can be used to define a large variety of different tasks in multi-modal new-view synthesis. The most direct setting for casual holography is to reconstruct a scene from a single head-mounted camera; then, reconstruction quality can be assessed in a 360° manner by using the static DSLRs cameras or the other head-mounts for evaluation. However, tasks of various complexity can also be defined, such as reconstruction from any combination of static and moving cameras. Furthermore, the fly-around segments at the beginning of each sequence can be used to test reconstruction in a decidedly simpler (semi-static) setting, and to simplify the reconstruction of the dynamic part as well.

Using Replay we define two such benchmark tasks of increasing difficulty and assess various existing techniques on them. Specifically, we consider baselines representing different families of radiance-field models (NeRF [40], TensoRF [8]) and their extensions dealing with dynamic scenes (NeRF+time, HexPlane [5], Nerfies [43]).

2. Related Work

Datasets for dynamic new-view synthesis. With the explosion of neural rendering, many datasets for studying new-view synthesis of dynamic content were proposed. Focusing on humans, HumanEva [52], Human3.6M [23], AIST [60], AIST++ [29], and ZJU-Mocap [46] portray a single person in isolation, without context, performing scripted motion. In contrast, Replay contains groups of humans acting naturally in a familiar environment.

More complex multi-view data containing dynamic humans in context include the Immersive Light Field dataset [3], which contains sixteen scenes captured from approximately 46 calibrated cameras. The NVIDIA Dynamic Scene Dataset [68] contains eight videos captured with 12 (mostly front-facing) calibrated GoPro Black Hero 7 cameras. The UCSD Dynamic Scene Dataset [33] contains 96 videos collected in a similar manner to [68], but using 10 cameras. The Plenoptic Video dataset [30] provides 6 more scenes from 21 cameras. All such videos are complex and visually diverse, but they all capture a single short-duration activity (typically 1 or 2 minutes at most).

Some datasets are collected using domes, and therefore do not contain natural environments or moving cameras. Panoptic Studio [24] contains 3 hours of recordings of humans engaged in multiple social activities captured with roughly 500 cameras and depth sensors. NeuralDome [71] contains videos of a single human manipulating an object captured from 76 cameras, and additional sensor-based participant tracking data. Our Replay focuses on long sequences with professional actors in a familiar setting. Furthermore, the usage of head-mounts makes our dataset par-

ticularly well-suited for studying scene reconstruction from a single egocentric device, which is one of the most realistic settings for future applications in casual holography.

Finally, all datasets above focus on visual reconstruction, and are thus not multimodal like Replay. See also Table 1 for a schematic summary.

Reconstructing 3D dynamic humans. Reconstructing a 4D video remains a challenging problem, so many authors have focused on special cases, such as reconstructing individual humans. Much work has focused on modelling articulated human bodies, including Neural volumes [36], Relightables [20], Articulated Neural Rendering [48], A-NeRF [56], Neural Actor [35], H-NeRF [67], Neural Performer [28], Deep Dynamic Character [22], Human Re-rendering [50], Pixel Aligned Avatars [49], HumanNeRF [65], HiFi Human Avatar [72], Generative Neural Articulated RFs [2], Animatable NeIS [45]. Most of these works approach the problem by explicitly tracking the human body, usually by using SMPL [37] fits, and then modelling shape and appearance in a canonical, articulation-free, space. Other works, including Neural Head [19], Dynamic Head [64], Dynamic Neural Faces [14], MoRF [62], specialise in reproducing heads, and a few such as Artemis [38] explore other animals. Since our scenes contain several interacting humans and objects, these methods are not applicable to our problem because they focus on isolated reconstructions of specific object classes.

Reconstructing generic 4D videos. Several authors have considered the problem of reconstructing generic 4D videos. Some have proposed to capture directly the light-field, with no or partial understanding of scene geometry. Examples include [18, 6, 4, 73, 26, 1, 25, 54, 70, 55, 3, 57].

Other methods model shape more explicitly, often using dynamic generalizations of NeRF [40]. These are the most applicable to the Replay scenarios. Many of them, including D-NeRF [47], Deformable NeRF [42], Dynamic NVS [68], Nerfies [43], HyperNeRF [44], Neural Trajectory Fields [61], NR-NeRF [58], NSFF [31], NeRFlow [11], STaR [69], NeRFPlayer [53], Deformable Voxel Grid [21], TiNeuVox [12], DynIBaR [32], DeVRF [34] attempt to estimate a deformation field and thus explicitly model the motion in the scene, reducing the video to a single canonical reconstruction and the deformation field. While this is statistically parsimonious, and necessary for reconstruction when the number of input viewpoints is small, estimating a correct deformation field is difficult due to the underconstrained nature of the problem.

Other methods directly add time to the radiance field parametrisation (sometimes called NeRF+t), thus avoiding the challenge of explicitly estimating deformations. Examples include NERF-W [39], NeuralDiff [59], Video NeRF [66], Dynamic View Synthesis [15], Fourier PlenOctrees [63] and DyNeRF [30]. Finally two recent concurrent works, K-Planes [13] and HexPlane [5], extend the voxel grid decomposition introduced in EG3D [7] and TensoRF [8] to spatio-temporal 4D grids.

Many of these methods explicitly require a large number of viewpoints. In some cases, this requirement is indirect [16], in the sense that methods may also work from a monocular camera, but only if the camera motion dominates the scene motion.

3. The Replay Dataset

The full Replay dataset consists of 68 scenes of social interactions between people, such as playing board games, exercising, or unwrapping presents. Each scene contains about 5 minutes of acting following a few minutes of calibration stages, and is filmed with 12 cameras, static and dynamic. Audio is captured separately by 8 binaural microphones and additional near-range microphones for each actor and for each egocentric device. All sensors are temporally synchronized, undistorted, geometrically calibrated, and color calibrated.

In addition to the full dataset, we introduce the *Replay novel view reconstruction benchmark*, a curated subset of scenes with given training and test sets and supplemental information such as foreground/background masks. We run several state-of-the-art novel view reconstruction methods on this benchmark and report the results in Section 5.

Content. The videos depict human social interaction in a large variety of indoor settings and contexts. Examples include meeting friends, talking, sitting in a living room, making hand gestures, playing charades, exercising on a yoga mat, playing video games, playing board games, arranging ornaments, having a meal, unwrapping presents, and more. Each scene contains up to 4 actors, with a total of 42 actors of diverse age, gender, and ethnicity across the scenes. In particular, 21 of them are white, 11 are Asian, 5 are black, and 5 are mixed race.

Scene setup. In each scene, there are three human operators wearing wearable *egocentric* cameras that provide eye-level views of the scene. The focus on monocular and binocular wearable cameras and microphones is a unique feature of Replay, which enables evaluating methods targeting AR/VR applications where a scene captured by one wearable device may have to be rendered in a world-locked manner on another device, from a new viewpoint.

In addition, the scene is shot by 8 static DSLR cameras arranged in a full circle around the action, approximately 45° apart from one another. Each static DSLR camera has a binaural microphone attached to it, and each actor and egocentric camera operator is equipped with a near-range microphone. We also provide an auxiliary capture of the entire scene with a 360° ceiling camera. This capture is

Dataset	#Sc.	Viewpoints	Resolution	Motion	Angles	Dur.	#Act.	Motion types
Dynamic Scene Dataset	8	1 moving camera	1920×1080	Dynamic	Frontal	5sec	1–4	Simple body motions (facial, jump, etc.)
ZJU-Mocap	10	21 static cameras	1024x1024	Dynamic	360°	20sec	1	Simple body motions (punch, kick, etc.)
AIST++	1408	9 static cameras	1920x1080	Dynamic	360°	20-50sec	1	Dancing
Ours: flyaround	46	1 moving, 8 static	3840x2160	≈Static	360°	40-60sec	1–4	Dancing, chatting, playing video games,
Ours: acting	46	3 quasi static, 8 static		Dynamic		3-5min		unwrapping presents, playing ping pong

Table 1. Comparison with related datasets. For each dataset, we report the number of scenes (*#Sc.*, which may be recorded from multiple (*viewpoints*) at different resolution, contain varying amount and type of *motions*, filmed either from the frontal position or from around the scene. Our dataset uniquely has multiple actors per scene (*#Act*) and duration (*Dur.*) of several minutes. Our datased has a natural background (as opposed to a studio or dome), and provided foreground masks include not only actors but also objects they are interacting with.

not intended to be used as input, and is included to provide an overview of the scene for users of the dataset. Figure 1 shows a bird-eye view of the scene setup, as well as a schematic representation.

Phases. Scenes are divided into three logical phases: calibration, flyaround, and acting. The *Calibration* phase is part of the scene setup, and contains images of calibration patterns and of the digital clapper. The *Flyaround* phase shows the actors take their place in the scene and remain still. Then, the wearable camera operators walk around the scene while looking at the central action area. This provides a continuous 360° view of a scene which remains nearly static. This stage takes 40–60 seconds. Finally, during the *Acting* phase the actors perform for about 3–5 minutes.

To reduce the amount of data necessary to process for the benchmark, we limit the fly-around part of the scene to 40 seconds and the acting part to 60 seconds. We further segment the acting part into two 30-seconds segments, since we found that most state-of-the art-methods are incapable of handling longer sequences. However, we strongly encourage future users of the dataset to test reconstructing at lest an entire minute.

Sensors. As shown in Fig. 1, Replay uses several sensors. These are: Eight static DSLR cameras (Sony A7 III; 24″ lens; 30 FPS; 4K resolution); Eight 3Dio binaural microphones co-located with the DSLR cameras; A ceiling camera (AXIS M4308-PLE: wide angle, 30 FPS, 2880^2 resolution, circular frame); Three GoPro cameras (Hero 9; RAW model, 60 FPS, 4K resolution); And a near-range lavalier microphone for each camera operator and actor in the scene.

Publicly-available assets. We have made the pre-processed data available to researchers, delivering the following assets. For each imaging sensor s, we provide (1) a collection of video frames I_{st} indexed by t; (2) the distortion and intrinsic calibration parameters (ρ_s, K_s); (3) the camera pose π_{st} with respect to the scene reference frame; (4) and, for 10 of the scenes, foreground segmentation masks M_{st} for each frame, including furniture, actors, and objects they interact with. For each audio sensor a, we provide (1) a collection of audio frames A_{at} indexed by t; (2) the location π_{at} of the sensor (which usually coincides with a certain imaging sensor), except for the near-range microphones, which are mounted on the actors and camera operators, whose dynamic location is therefore difficult to estimate. All sensors are temporally synchronised; for this, we provide the time information τ_{st} and τ_{at} for each video and audio frame.

We also provide benchmark definitions (Section 5.1) and corresponding evaluation code.

4. Data collection

Collecting a dataset such as Replay is a major endeavour. We describe the key aspects of the data collection to better understand the properties of the dataset, and because they can be helpful for other researchers that wish to engage in a similar experimental activity.

4.1. Production

The data was produced with the help of a vendor who took care of finding locations, ordering hardware, hiring professional actors, running the filming, quality assurance, and assigning basic metadata. Production lasted for more than six months, and resulted in the collection of 119 scenes, of which 68 have been processed so far to be released. The vendor was instructed to calibrate sensors before recording each scene, as described below.

To be able to diversify recording locations and keep natural backgrounds, we required a relatively mobile capture setup that (contrary to a dome) required non-trivial calibration and synchronisation from scratch before each recording. The large amount of sensors of different types and setup phases significantly increased the likelihood of human and equipment failures, such as camera or microphone malfunction, accidental camera movement during filming, wrong focus or focal length, missing or low-quality setup step for a specific camera, *etc*. Secure data storage and transfer was also a challenge due to the data volume (60 GB per scene).

4.2. Processing

The captured data required substantial pre-processing, including, in order: intrinsic calibration, temporal synchro-

nisation, temporal segmentation, extrinsic calibration, photometric calibration, foreground/background segmentation. The principal challenges and solutions are discussed next.

Intrinisic calibration. The focal length, principal point, and lens distortion of each sensor was estimated by asking the vendor to show a moving ChARuco calibration board at least once to each sensor, and for latter recordings, before recording each scene. A ChARuco board is a commercially available checkerboard combined with ARuco tags [17], which allows disambiguating the pose of the board in camera coordinates. These estimations were then used initialise the cameras in the COLMAP SfM software [51], which then refined the intrinsic parameters through joint optimisation with camera poses; see below for details.

Temporal synchronisation. Accurate temporal synchronisation of the various sensors is crucial for training and evaluating a dynamic scene from multiple view points. While specific sensors support hardware-based synchronisation, this is inapplicable to our heterogeneous mix of sensors. Instead, the sensors are synchronised using their audio signature. We take a salient segment of the audio recorded by each sensor, and match it with the entire audio of a second sensor using cross correlation. To increase robustness, we take several segments from different positions in the audio file and use a majority vote to estimate the true offset between each two sensors. After the computation of the pairwise correspondence, we check the stability of the estimated offsets by analysing cycles of three or more sensors. That allows us to identify sensors which were not matched correctly, due to errors in the data collection, for example noisy cameras or cameras which did not record audio due to technical issues. In such cases, manual temporal synchronisation was required for specific sensors in a small portion of the shots.

Camera pose calibration. Camera poses were estimated using COLMAP [51] with several improvements for robustness and scalability. Specifically, camera intrinsics were initialising via ChARuco calibration (see above); we then fixed the principal point, but let COLMAP refine the focal length and distortion parameters. The environment was first reconstructed using the head-mounted camera from the fly-around phase of the capture at a low frame rate (3 FPS). This produced a sufficiently small number of frames ($\sim$1,000) with sufficient parallax for COLMAP to run successfully. Then, all the other frames in the capture were triangulated against this initial reconstruction after masking out image regions prone to contain dynamic objects (instances of person, cat, dog segmented using PointRend [27]). Finally, since the absolute (and thus relative) positions of the static cameras are constant, they are assumed to form a rig; we thus use bundle adjustment with rig constraints to reconstruct them. For environments that did not provide sufficient texture detail to robustly triangulate the DSLR poses, we further refined calibration of the DSLR cameras by showing to pairs of them a ChARuco board.

Photometric Calibration. We bring all frames from different sensors in the same sRGB colour space, despite differences in the sensor type, factory calibration, and possible processing steps over which we had no control. This was done by (1) using the color model of each sensor to map colors to linear space; (2) filming a reference colour chart from each camera and fitting a linear map to align colours between sensor pairs; (3) moving colours back to sRGB space.

Foreground segmentation. We provide high-quality foreground masks for part of the dataset. Fully automatic foreground segmentation in videos is still an extremely challenging research problem. We therefore adopted a human-in-the-loop approach, leveraging the state-of-the-art XMem [10] video segmenter. The model extends scribbles annotated by an operator to generate a segmentation mask for each video frame, and then automatically propagates masks to the subsequent frames. When an error is detected, the operator intervenes drawing additional scribbles to correct the current prediction, and then re-initiating the propagation process. Since the definition of foreground objects is open to interpretations, we made sure the same person worked on each scene, ensuring consistency within and across its videos. Although the process requires only a limited intervention, we still found it time consuming due to the long duration of the videos and the number of cameras per scene. Thus, only 10 scenes are annotated with foreground masks.

5. Experiments

After defining two benchmarks (Section 5.1), we conduct several experiments on Replay for the purpose of demonstrating its application to the development of new-view synthesis methods, as well as to assess current state-of-the-art neural rendering techniques (Section 5.3) on this challenging data (Section 5.4).

5.1. Benchmarks

We define two novel-view synthesis benchmarks using Replay: *flyaround*, with semi-static actors filmed from 360° trajectory, and the more challenging *acting*, where naturally behaving actors are filmed with frontal cameras.

Flyaround. This simpler benchmark allows evaluating standard reconstruction approaches such as neural radiance fields that cannot model time, alongside with those that model time dependency and shape deformations. The 40-second segments are extracted from each of the Replay sequences (see Section 3). Training frames are extracted from one dynamic wearable camera at 30 FPS (GoPro-2 in

Fig. 1), and all 8 static DSLR cameras are used for evaluation. This amounts to 1,200 training and 64 evaluation frames per scene. Despite the fact that there is no significant motion in the scene in these segments, the task is already quite challenging since, unlike typical NVS datasets, we require *extrapolating* beyond camera the trajectory, as the DSLRs are located far away from the trajectory of the wearable camera used for capture.

Acting. This is the most challenging setting since the actors are allowed to move freely, while the operators of wearable cameras stand fairly still in front of them. Monocular reconstruction in this context is still beyond the capability of state-of-the-art reconstruction methods, so in this instance we consider a multi-view reconstruction setup. We consider a 30-second segment sampled at 30 FPS, which is significantly longer than the data used for testing modern deformable NVS methods, so it stretches their limits. We hold out one static DSLR for evaluation, while using two relatively frontal DSLRs (DSLR-5 and DSLR-6 in Fig. 1) and 3 wearable cameras for training, resulting in 4,500 training frames and 50 randomly sampled evaluation frames per scene. While this scenario is more challenging because of the changing geometry of the scene, all training and evaluation sensors are located in front of the actors, so that the methods do not have to generalise to a wide range of viewing angles.

5.2. Metrics

We compare the methods using a range of metrics evaluating the faithfulness of the rendering, their perceptual quality, and the quality of the opacity mask (for the methods that produce it). To this end, we use the following metrics: Peak Signal-to-Noise Ratio (PSNR), Learned Perceptual Image-Patch Similarity (LPIPS), both evaluated in the foreground region only, and the Intersection over Union (IoU) between the produced opacity mask and ground-truth foreground mask. We compute the metrics at a fixed resolution of 960×540, which can be handled by all the methods.

Method	Flyaround (≈static)			Acting (dynamic)		
	PSNR	IOU	LPIPS	PSNR	IOU	LPIPS
NeRF [40]	21.85	**0.95**	**0.22**	20.22	0.93	0.23
NeRF+t	20.86	0.92	0.25	**21.28**	**0.94**	**0.22**
TensoRF [8]	20.58	0.92	0.22	17.26	0.78	0.42
HexPlane [5]	15.08	0.87	0.29	17.66	0.72	0.44
Nerfies [43]	**23.22**	N/A	0.61	18.08	N/A	0.71

Table 2. Quantitative results on the two proposed benchmarks. Please note that the numbers are not comparable across benchmarks: in *flyaround*, methods have to model a wider range of viewpoints. In *acting*, all training and evaluation cameras are frontal, but the methods have to model the dynamic geometry of the scene.

5.3. Baselines

We evaluate the dataset on a range of novel-view synthesis methods, including those modelling dynamic scenes.

NeRF and NeRF+t. Neural Radiance Field (NeRF) [40] fitting is a cornerstone of modern novel-view synthesis. The method learns the radiance field of the scene through an MLP Ψ that predicts the colour $c_i \in \mathbb{R}^3$ and density $o_i \in \mathbb{R}$ for points r_i along the rays $\mathbf{r}$ emitted from the camera center passing through each pixel:

$$[c_i, o_i] = \Psi(\gamma_{R_x}(r_i), \gamma_{R_d}(d_i)), \tag{1}$$

where d_i is a normalised vector pointing from the camera centre to r_i, γ_R is an order-R harmonic encoding, $R_x, R_d \in \mathbb{N}$ are hyperparameters. The predicted colours and opacities are then integrated along the ray using the emission-absorption raymarching function to get the final RGB value in the corresponding pixel. Unlike standard NeRF, we do not model view-dependent colours (*i.e.* $R_d = 0$) due to sparsity of input views: in *flyaround* setting, we noticed the model generalises poorly to DSLR camera poses that are located farther away from the scene centre than the wearable camera's trajectory; in *acting* setting, we found that 5 viewpoints per timestamp are not sufficient to fit view-directional colours reliably, i.e. we set $R_d = 0$. Since we are interested only in the foreground, we pre-process the images by masking out background pixels.

NeRF assumes that the scene is static and produces blurry renders even in case of a limited non-deliberate motion. Hence, we consider the temporal extension NeRF+t (used in various video reconstruction methods [39, 59, 66]):

$$[c_i, o_i] = \Psi(\gamma_{R_x}(r_i), \gamma_{R_t}(t)), \tag{2}$$

where t is a frame's timestamp normalised to $[0, 1]$ range. This model is thus tasked in modelling a 4D time-space.

Nerfies. Nerfies [43] extend the vanilla NeRF model to handle deformations, but does this in a different way than NeRF+t. Instead of treating t as an additional input dimension, Nerfies model the dynamics by considering a time-invariant rigid radiance field Ψ in *canonical space*, and a time-dependant deformation field Δ, which allows to convert points from posed to canonical space. Therefore, in order to compute the color of a pixel of an image I_k, the points r_i along the ray $\mathbf{r}$ are first offset to canonical coordinates by applying Δ, before obtaining the colors c_i and opacities o_i by evaluating the implicit function Ψ:

$$\begin{aligned} \bar{r}_i &= \Delta(\gamma_{R_x}(r_i), \phi_k) \\ [c_i, o_i] &= \Psi(\gamma_{R_x}(\bar{r}_i), \gamma_{R_d}(d_i)), \end{aligned} \tag{3}$$

where ϕ_k is an appearance code corresponding to image I_k. We found that the quality of Nerfies degrades with masking, so we train it on the unmasked videos. Note that we still report the foreground-only PSNR and LPIPS.

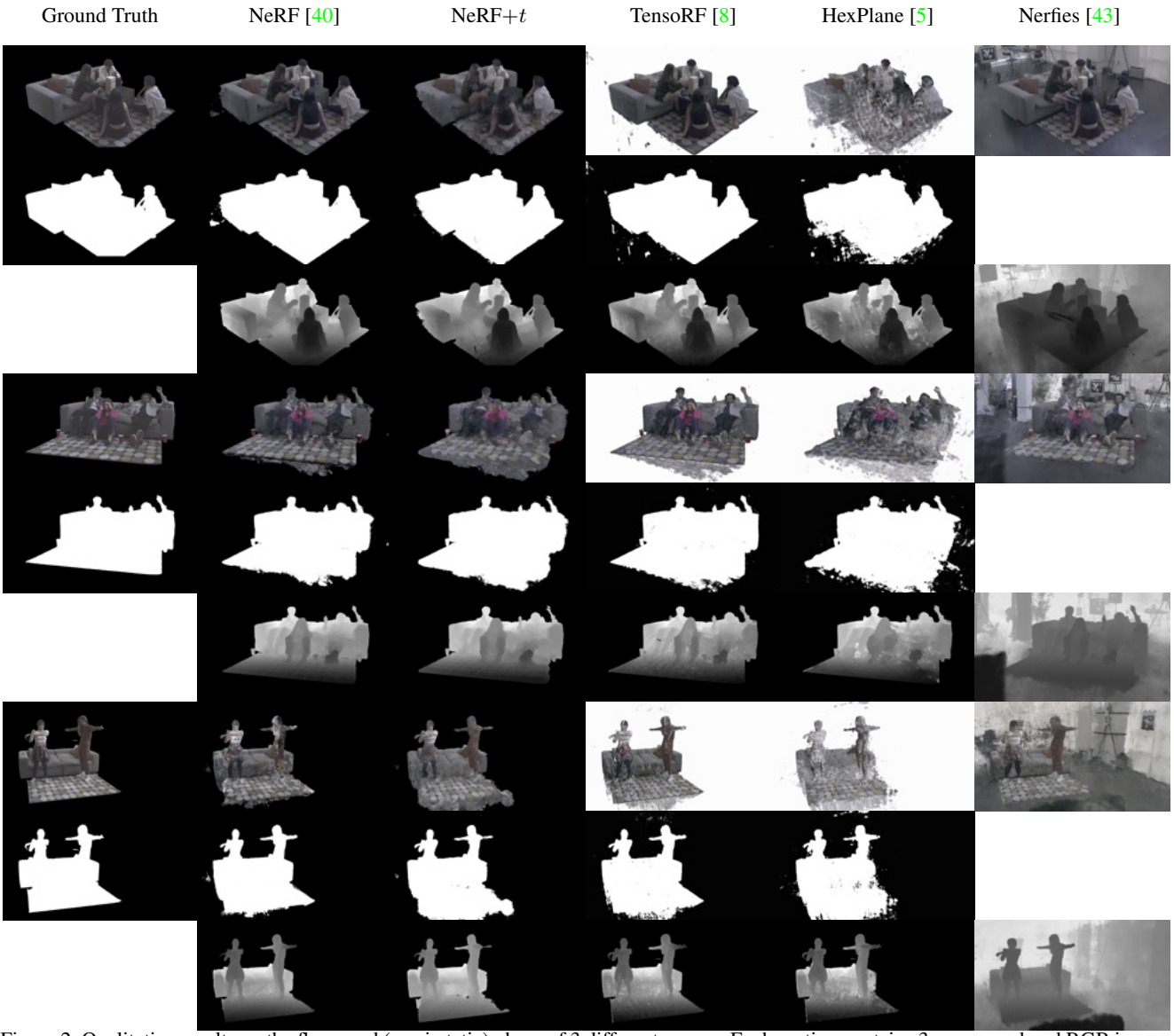

Figure 2. Qualitative results on the flyaround (semi-static) phase of 3 different scenes. Each section contains 3 rows: rendered RGB image, rendered opacity mask, and rendered depth map. Note that NeRFies is not trained to produce opacity masks, so we skip these renders.

TensoRF and HexPlane. TensoRF [8] shares with NeRF the underlying idea of volumetric rendering through emission-absorption ray marching. However, instead of modelling the radiance field with an MLP, TensoRF proposes to model it through a product of a set of 2D (M^{XY}, M^{YZ}, M^{XZ}) and 1D tensor components (v^X, v^Y, v^Z), which factorize the 3D density and color spatial fields in a memory- and computationally-efficient manner.

HexPlane [5] extends TensoRF by considering the factorization of the 4D space-time density and color fields (analogousy to how NeRF+t extends NeRF), and therefore including into the factorization 2D tensors M^{Xt}, M^{Yt}, M^{Zt} which span the temporal axis t along with each spatial axis x, y, or z.

5.4. Results

Flyaround. The result on a flyaround benchmark are shown in Figure 2 and summarised in the left half of Table 2. NeRF and NeRFies produce best results, with the latter better adapting to small movements present in the scene. While TensoRF shows results comparable to NeRF, its temporal extension, HexPlane, falls short to generalise to a wide range of viewing angles, presumably due to overfitting to the additional temporal dimension.

Acting. The results on acting benchmark are shown in Figure 3 and summarised in the right half of Table 2. Here, time-extension of NeRF shows the best quality, being able to learn the dynamic geometry to a better extent, while time-

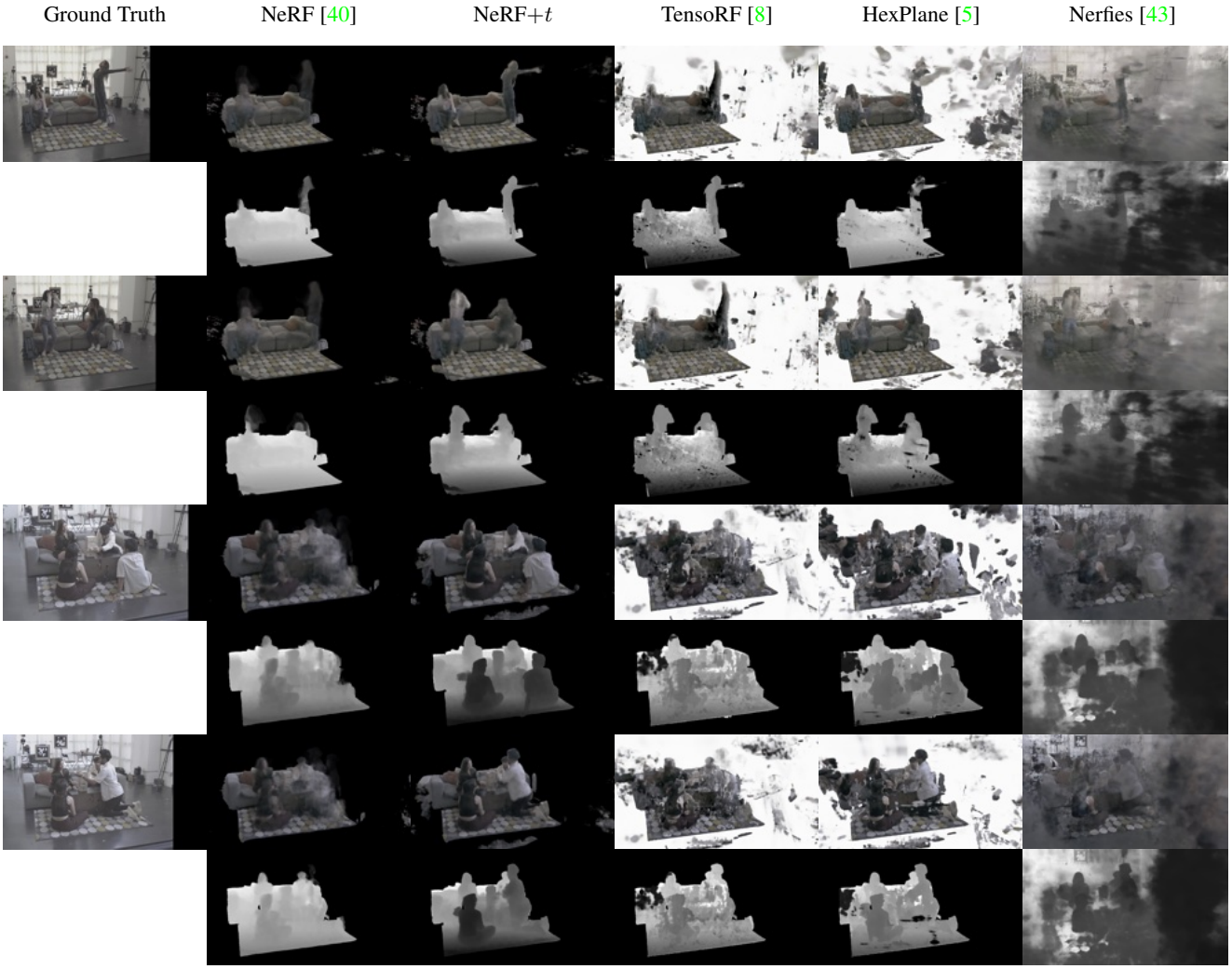

Figure 3. Qualitative results on the acting (dynamic) phase of 3 different scenes. Each section contains 2 rows: rendered RGB image and rendered depth map. Sections 1–2 and 3–4 come from the same scene; note that NeRF and TensoRF produce the same average render with ghosting artifacts because it is missing time input. NeRFies is not trained to produce opacity masks, so we skip these renders.

agnostic NeRF produces the ghost-looking shapes wherever an actor changed the pose. HexPlane, on the other hand, in spite of a better ability to model deformations, does not improve much over TensoRF, and NeRFies fails to reconstruct the geometry explaining the errors away with floaters.

6. Conclusion

We presented Replay, a collection of scenes captured with egocentric and scene-static sensors. We aim primarily to support research in new-view synthesis of dynamic and multi-modal content from egocentric sensors, including in particular reconstruction from a single viewpoint or a very narrow baseline. In the future, this technology will enable breakthrough applications such as personal holography. While this task is still too challenging for existing new-view synthesis methods, as generative AI matures, it will become possible to better hallucinate information missing in the capture, and Replay can spur further research in this direction. Furthermore, while in this paper we have only discussed the visual component of the data, Replay also contains fully-calibrated and synchronised audio information for research in multi-modal new-view synthesis, which is as of today largely unexplored.

Limitation. There are a few limitations of the released part of the dataset. First, we did not record stereo videos, which might become an important modality with the next generation of wearable devices. Second, the operators of dynamic cameras do not change location during acting, only moving their heads in a natural way, which represents a subset of AR/VR applications. Finally, all released scenes were filmed in the same room, albeit with various furniture and props. These limitations may be addresses in future work.

References

[1] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa G. Narasimhan. 4D visualization of dynamic events from unconstrained multi-view videos. In *Proc. CVPR*, 2020. 3

[2] Alexander W. Bergman, Petr Kellnhofer, Yifan Wang, Eric R. Chan, David B. Lindell, and Gordon Wetzstein. Generative neural articulated radiance fields. *arXiv.cs*, abs/2206.14314, 2022. 3

[3] Michael Broxton, John Flynn, Ryan S. Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul E. Debevec. Immersive light field video with a layered mesh representation. *Proc. SIGGRAPH*, 39(4), 2020. 2, 3

[4] Chris Buehler, Michael Bosse, Leonard McMillan, Steven J. Gortler, and Michael F. Cohen. Unstructured lumigraph rendering. In *Proc. SIGGRAPH*, 2001. 3

[5] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *arXiv.cs*, abs/2301.09632, 2023. 2, 3, 6, 7, 8

[6] Jinxiang Chai, Shing-Chow Chan, Heung-Yeung Shum, and Xin Tong. Plenoptic sampling. In *Proc. SIGGRAPH*, 2000. 3

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *Proc. CVPR*, 2022. 3

[8] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial radiance fields. In *arXiv*, 2022. 2, 3, 6, 7, 8

[9] Changan Chen, Alexander Richard, Roman Shapovalov, Vamsi Krishna Ithapu, Natalia Neverova, Kristen Grauman, and Andrea Vedaldi. Learning audio-visual dereverberation. In *CVPR*, 2023. 1

[10] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. *CoRR*, abs/2207.07115, 2022. 5

[11] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proc. ICCV*, 2021. 3

[12] 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 *Proc. SIGGRAPH Asia*, 2022. 3

[13] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. *arXiv.cs*, abs/2301.10241, 2023. 3

[14] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proc. CVPR*, 2021. 3

[15] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. *arXiv.cs*, abs/2105.06468, 2021. 3

[16] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *arXiv.cs*, abs/2210.13445, 2022. 3

[17] Sergio Garrido-Jurado, Rafael Munoz-Salinas, Francisco Jose Madrid-Cuevas, and Manuel Jesus Marin-Jimenez. Automatic generation and detection of highly reliable fiducial markers under occlusions. *Pattern Recognition*, 47(6):2280–2292, 2014. 5

[18] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proc. SIGGRAPH*, 1996. 3

[19] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *cvpr*, 2021. 3

[20] Kaiwen Guo, Peter Lincoln, Philip L. Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Ryan Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul E. Debevec, and Shahram Izadi. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), 2019. 3

[21] Xiang Guo, Guanying Chen, Yuchao Dai, Xiaoqing Ye, Jiadai Sun, Xiao Tan, and Errui Ding. Neural deformable voxel grid for fast optimization of dynamic view synthesis. In *Proc. ACCV*, 2022. 3

[22] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Trans. Graph.*, 40(4), 2021. 3

[23] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *PAMI*, 36(7), 2014. 2

[24] Hanbyul Joo, Tomas Simon, Xulong Li, Hao Liu, Lei Tan, Lin Gui, Sean Banerjee, Timothy Scott Godisart, Bart Nabbe, Iain Matthews, Takeo Kanade, Shohei Nobuhara, and Yaser Sheikh. Panoptic studio: A massively multiview system for social interaction capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 2

[25] Nima Khademi Kalantari, Ting-Chun Wang, and Ravi Ramamoorthi. Learning-based view synthesis for light field cameras. *Proc. SIGGRAPH*, 35(6), 2016. 3

[26] Takeo Kanade, Peter Rander, and P. J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multim.*, 4(1), 1997. 3

[27] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross B. Girshick. Pointrend: Image segmentation as rendering. In *Proc. CVPR*, 2020. 5

[28] Youngjoong Kwon, Dahun Kim, Duygu Ceylan, and Henry Fuchs. Neural human performer: Learning generalizable radiance fields for human performance rendering. In *Proc. NeurIPS*, 2021. 3

[29] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with AIST++: Music conditioned 3d dance generation. In *Proc. ICCV*, 2021. 2

[30] Tianye Li, Mira Slavcheva, Michael Zollhöfer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard A. Newcombe, and Zhaoyang Lv. Neural 3D video synthesis from multi-view video. In *Proc. CVPR*, 2022. 2, 3

[31] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proc. CVPR*, 2021. 3

[32] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. *arXiv.cs*, abs/2211.11082, 2022. 3

[33] Kai-En Lin, Lei Xiao, Feng Liu, Guowei Yang, and Ravi Ramamoorthi. Deep 3d mask volume for view synthesis of dynamic scenes. In *Proc. ICCV*, 2021. 2

[34] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *arXiv.cs*, abs/2205.15723, 2022. 3

[35] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.*, 40(6), 2021. 3

[36] Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. Neural volumes: learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4), 2019. 3

[37] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: a skinned multi-person linear model. *ACM Trans. on Graphics (TOG)*, 2015. 3

[38] Haimin Luo, Teng Xu, Yuheng Jiang, Chenglin Zhou, Qiwei Qiu, Yingliang Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Artemis: Articulated neural pets with appearance and motion synthesis. *ACM Trans. on Graphics (TOG)*, 2022. 3

[39] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *Proc. CVPR*, volume abs/2008.02268, 2021. 3, 6

[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 *Proc. ECCV*, 2020. 1, 2, 3, 6, 7, 8

[41] Mohamed Omran, Christoph Lassner, Gerard Pons-Moll, Peter V. Gehler, and Bernt Schiele. Neural body fitting: Unifying deep learning and model based human pose and shape estimation. In *3DV*, 2018. 2

[42] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B. Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Deformable neural radiance fields. *CoRR*, abs/2011.12948, 2020. 3

[43] 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 *iccv*, 2021. 2, 3, 6, 7, 8

[44] 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. *Proc. SIGGRAPH*, 40(6), 2021. 3

[45] Sida Peng, Shang-Wei Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *arXiv*, abs/2203.08133, 2022. 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. *CoRR*, abs/2012.15838, 2020. 2

[47] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. *arXiv.cs*, abs/2011.13961, 2020. 3

[48] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. ANR: articulated neural rendering for virtual avatars. *arXiv.cs*, abs/2012.12890, 2020. 3

[49] Amit Raj, Michael Zollhöfer, Tomas Simon, Jason M. Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pixel-aligned volumetric avatars. In *Proc. CVPR*, 2021. 3

[50] Kripasindhu Sarkar, Dushyant Mehta, Weipeng Xu, Vladislav Golyanik, and Christian Theobalt. Neural re-rendering of humans from a single image. *CoRR*, abs/2101.04104, 2021. 3

[51] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proc. CVPR*, 2016. 5

[52] Leonid Sigal, Alexandru O. Balan, and Michael J. Black. HumanEva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *IJCV*, 87(1-2), 2010. 2

[53] 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. *arXiv.cs*, abs/2210.15947, 2022. 3

[54] Pratul P. Srinivasan, Richard Tucker, Jonathan T. Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proc. CVPR*, 2019. 3

[55] Timo Stich, Christian Linz, Georgia Albuquerque, and Marcus A. Magnor. View and time interpolation in image space. *Comput. Graph. Forum*, 27(7), 2008. 3

[56] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Surface-free human 3d pose refinement via neural rendering. *arXiv.cs*, abs/2102.06199, 2021. 3

[57] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. *arXiv.cs*, abs/2112.09687, 2021. 3

[58] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proc. ICCV*, 2021. 3

[59] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2021. 3, 6

[60] Shuhei Tsuchida, Satoru Fukayama, Masahiro Hamasaki, and Masataka Goto. AIST dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In *Proc. ISMIR*, pages 501–510, 2019. 2

[61] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv.cs*, abs/2105.05994, 2021. 3

[62] Daoye Wang, Prashanth Chandran, Gaspard Zoss, Derek Bradley, and Paulo F. U. Gotardo. Morf: Morphable radiance fields for multiview neural head modeling. In *Proc. SIGGRAPH*, 2022. 3

[63] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier PlenOctrees for dynamic radiance field rendering in real-time. In *Proc. CVPR*, 2022. 3

[64] Ziyan Wang, Timur M. Bagautdinov, Stephen Lombardi, Tomas Simon, Jason M. Saragih, Jessica K. Hodgins, and Michael Zollhöfer. Learning compositional radiance fields of dynamic human heads. In *Proc. CVPR*, 2021. 3

[65] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. *arXiv.cs*, abs/2201.04127, 2022. 3

[66] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proc. CVPR*, 2021. 3, 6

[67] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Proc. NeurIPS*, 2021. 3

[68] 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 *Proc. CVPR*, 2020. 2, 3

[69] Wentao Yuan, Zhaoyang Lv, Tanner Schmidt, and Steven Lovegrove. STaR: Self-supervised tracking and reconstruction of rigid objects in motion with neural rendering. In *cvpr*, 2021. 3

[70] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Trans. Graph.*, 40(4), 2021. 3

[71] Juze Zhang, Haimin Luo, Hongdi Yang, Xinru Xu, Qianyang Wu, Ye Shi, Jingyi Yu, Lan Xu, and Jingya Wang. Neuraldome: A neural modeling pipeline on multi-view human-object interactions. *arXiv.cs*, abs/2212.07626, 2022. 2

[72] Hao Zhao, Jinsong Zhang, Yu-Kun Lai, Zerong Zheng, Yingdi Xie, Yebin Liu, and Kun Li. High-fidelity human avatars from a single RGB camera. In *Proc. CVPR*, 2022. 3

[73] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon A. J. Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *Proc. SIGGRAPH*, 23(3), 2004. 3