EgoGaussian: Dynamic Scene Understanding from Egocentric Video with 3D Gaussian Splatting

Daiwei Zhang[1] Gengyan Li[1,2] Jiajie Li[1] Mickaël Bressieux[1]
Otmar Hilliges[1] Marc Pollefeys[1,3] Luc Van Gool[1,4,5] Xi Wang[1]

[1]ETH Zürich [2]Google [3]Microsoft [4]KU Leuven [5]INSAIT, Sofia

Abstract

Human activities are inherently complex, and even simple household tasks involve numerous object interactions. To better understand these activities, it is crucial to model their interactions with the environment captured through dynamic changes. The recent availability of affordable head-mounted cameras and egocentric data offers a more accessible and efficient means to understand dynamic human-object interactions in 3D environments. However, most existing methods for human activity modeling either focus on reconstructing 3D models of hand-object or human-scene interactions or on mapping 3D scenes, neglecting dynamic interactions with objects. The few existing solutions often require inputs from multiple sources, including multi-camera setups, depth-sensing cameras, or kinesthetic sensors. To this end, we introduce EgoGaussian, the first method capable of simultaneously reconstructing 3D scenes and dynamically tracking 3D object motion from RGB egocentric input alone. We leverage the uniquely discrete nature of Gaussian Splatting and segment dynamic interactions from the background. Our approach employs a clip-level online learning pipeline that leverages the dynamic nature of human activities, allowing us to reconstruct the temporal evolution of the scene in chronological order and track rigid object motion. Additionally, our method automatically segments object and background Gaussians, providing explicit 3D representations for both static scenes and dynamic objects. EgoGaussian shows significant improvements in terms of both dynamic object and background reconstruction quality compared to the state-of-the-art. We also qualitatively demonstrate the high quality of the reconstructed models.

1. Introduction

Human activities are inherently complex and performing simple household tasks involves numerous interactions with objects. For example, making a coffee in the morning involves multiple steps: taking a mug from a shelf, placing it under the coffee machine, pressing a button for the preferred type of coffee, and adding milk or sugar. Even this seemingly simple task includes various object interactions and movements. To better understand human activities and behaviors, it is important to be able to model these dynamic interactions with the environment. The recent availability of affordable head-mounted cameras [43, 54] and egocentric data [11, 19, 20, 37] offers a more accessible and efficient means to understand dynamic human-object interactions in 3D environments. Toward this goal, we tackle the challenging task of reconstructing 3D scenes and dynamic interactions of objects from RGB egocentric videos.

Most existing methods for modeling human-object interactions either focus on reconstructing 3D hand-object [16, 32, 61, 70] or human-scene interaction models [1, 25, 27, 31, 72, 76] or on mapping 3D scenes [57]. These approaches often neglect dynamic interactions with objects, resulting in static representations with motion-induced artifacts, commonly known as the "ghost effect". The few existing solutions often require inputs from multiple sources, including multi-camera setups [36], depth-sensing cameras [62], or kinesthetic sensors [24]. While these methods achieve 3D reconstruction, they do not consider changes caused by interactions and thus fail to capture the dynamics depicted in egocentric videos.

In this paper, we go beyond prior works to tackle the task of dynamic scene reconstruction from RGB egocentric videos. Our proposed method EgoGaussian simultaneously reconstructs 3D scenes and dynamically tracks 3D object motions within them. Our key insight is that the uniquely discrete nature of Gaussian Splatting makes it especially suitable for spatial segmentation, allowing objects to be trained separately from the background. Given that human activities involve continuous motion over time, we identify critical contact points in time and distinguish dynamic interactions from static captures that only contain camera movements. We propose a clip-level online learning pipeline that leverages the dynamic nature of human activities, allowing us to reconstruct the temporal evolution of the scene in chronological order and track rigid object motion.

To reconstruct the dynamic scenes from an egocentric video, EgoGaussian first obtains hand-object segmentation using an off-the-shelf method and derives camera poses through structure-from-motion. By leveraging the natural trajectories of interactions, we partition the input video into static and dynamic clips. The static clips are used to reconstruct the background scenes and initialize the shapes of the object that will be interacted with. Subsequently, we refine the object's shapes and track their motion through the dynamic clips. We empirically show that EgoGaussian achieves better reconstruction of dynamic scenes than the state-of-the-art. We quantitatively evaluate our method on two in-the-wild egocentric video datasets following the evaluation protocol for novel-view synthesis. We also qualitatively demonstrate the high quality of the reconstructed scenes and the tracked object shapes and their motion.

Our main contributions can be summarized as follows:

- We present a novel method that accurately reconstructs 3D scenes and dynamic object motion within them from RGB egocentric videos.
- We leverage the dynamic nature of interactions that consist of transitions between static and dynamic phases, which facilitates the reconstruction of the static scenes, the object shapes, and the tracking of their motion.
- Through both qualitative and quantitative evaluation, we demonstrate that our method outperforms previous approaches and provides better 4D reconstruction that captures the dynamic object interactions.

Video results and code are publicly available at https://zdwww.github.io/egogs.github.io/

2. Related Work

Hand-Object Segmentation. Many works have studied hand-object interaction in egocentric vision from different aspects. One significant area of focus is segmentation, specifically obtaining image segmentation masks of hands and the objects they hold. Ren et al. [48] proposed a motion-based approach to robustly segment both hand and object using optical flow and domain-specific cues from egocentric video.

Concurrent with the emergence of deep neural networks-based hand-object segmentation is the scaling-up of egocentric data that includes pixel-level annotations and involves diverse daily activities [11, 12, 19]. VISOR[13] annotates videos from EPIC-KITCHENS[11, 12] dataset and provides masks for 67k hand-object relations covering 36 hours of videos. EgoHOS [74] further introduces the notion of a dense contact boundary to explicitly model the interaction and a context-aware compositional data augmentation technique to generate semantically consistent hand-object segmentation on out-of-distribution egocentric videos. Cheng et al. [9] produces a rich, unified 2D output of interaction by converting predicted bounding boxes to segments with Segment Anything (SAM) [29]. Our method takes egocentric videos with hand-object segmentation masks as input and creates dynamic 3D models.

Hand-Object Reconstruction. Another highly related direction is to reconstruct the hand-object interaction, featuring 3D pose estimation for hands and objects. Recent works often jointly reconstruct hands and objects to favor physically plausible interactions [16, 32, 61, 70, 71, 78]. These approaches can be grouped into two categories. One assumes a known 3D object model and fits that model into 2D image [6, 10, 32, 56, 67]. For example, RHO [6] adapts an optimization-based approach that's able to reconstruct hands and objects from single images in the wild, by leveraging 2D image cues and 3D contact priors to provide constraints.

Recent works eliminate the need for a known 3D model and directly reconstruct 3d object shapes from the input [16, 47, 71]. However, they either require multiview input [47], specific hand-object interaction supervision [71], or can only reconstruct simple object shapes [16]. Current shape-agnostic methods struggle in in-the-wild scenarios [16, 78]. In contrast, our method does not require prior knowledge and obtains 3D object shapes through differentiable 3D Gaussian-based rendering.

Static Scene Modeling. In the past few years, the domain of static scene modeling has garnered considerable attention. Mildenhall et al. [40] introduce the groundbreaking Neural Radiance Fields (NeRF), which utilizes a large Multilayer Perceptron (MLP) to represent 3D scenes and renders via volume rendering technique. However, their method queries the MLP at hundreds of points for each ray, resulting in slow training and rendering speed. Additionally, the original NeRF's performance can diminish in scenes with highly dynamic elements due to its static, volumetric nature. Therefore, some subsequent works have tried to enhance the quality by (1) mitigating existing problems, such as aliasing [3, 4, 8] and reflection [23, 58] (2) incorporating image processing [38, 41] (3) employing per-image transient latent codes [39, 52], and (4) introducing supervision of expected depth with point clouds [14, 65]. There also exist some other follow-up works aiming to improve the speed, for example, by caching precomputed MLP results [26, 73], employing well-designed data structures [7, 55], removing the neural network [18], or utilizing multi-resolution hash encoding [42]. Yet, most of these methods still use ray marching, which involves sampling millions of points and slows down real-time rendering.

Recently, Kerbl et al. [28] propose a different approach in the modeling and rendering of complex static 3D scenes - 3D Gaussian Splatting (3DGS). They model static scenes with Gaussians whose position, opacity, shape, and color

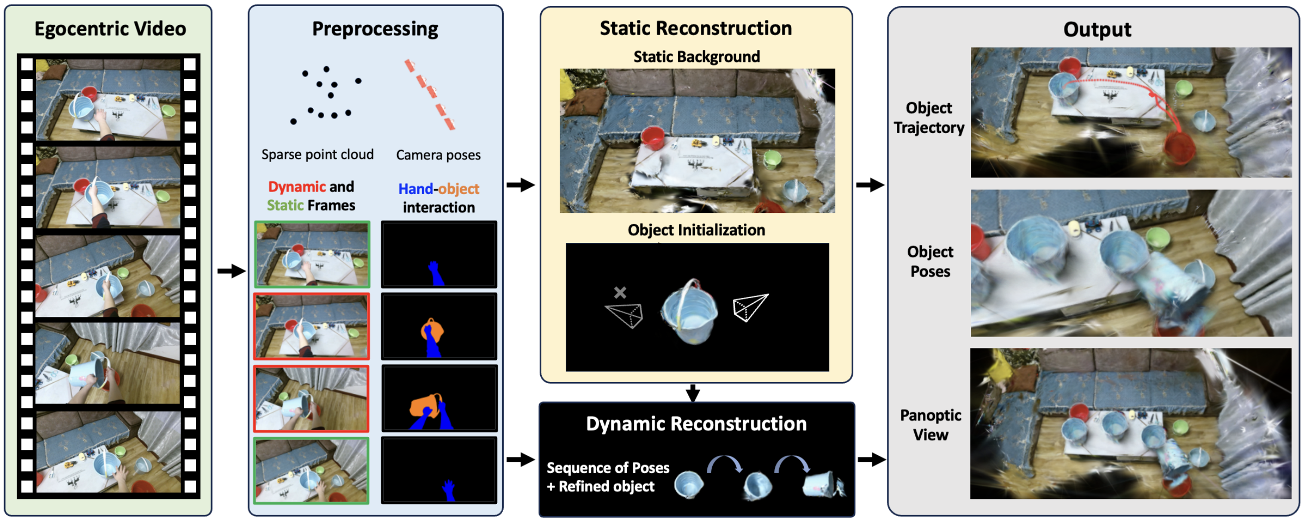

Figure 1. **EgoGaussian Pipeline.** Given an egocentric video input, our framework first estimates camera poses via structure-from-motion and obtains hand-object segmentation masks using an off-the-shelf approach. We also partition the video input into static and dynamic clips in the preprocessing step. The static clips are used to reconstruct the background scenes and initialize the shapes of the object that will be interacted with. Subsequently, we refine the object's shapes and track their motion through the dynamic clips.

are learned through a differentiable splatting-based renderer, achieving real-time rendering speed.

Dynamic Scene Modeling. Motivated by the success of NeRF [40] in static scene modeling, numerous studies have adopted neural representations to model dynamic scenes. One strategy to extend 3D into 4D scenes is by using time stamps as an additional conditioning factor. [2, 64]. Another set of 'dynamic NeRF' works [5, 17, 49, 53, 59] involves employing 4D space-time grid-based representations. Representing the 3D scene at a certain timestamp as a canonical space and then explicitly modeling deformation fields to warp 3D points into the canonical space is a common strategy as well [15, 34, 44, 46]. Another strategy is to combine the two approaches, using a conditional neural volume together with a deformation field [30, 45]. However, these methods all suffer from the same issues as static NeRFs, in that they require raymarching and despite advances in performance, still are not sufficiently fast for real-time rendering.

Similarly, many dynamic extensions to 3D Gaussian splatting were also proposed [35, 36, 63, 68]. The most common approach is to learn for every timestep a set of deformations for each Gaussians. This can be done explicitly, or implicitly using a deformation which is evaluated for each Gaussian. This results in substantially faster training and rendering speed, with comparable levels of rendering quality. Although these methods result in decent quality renders, upon closer inspection all of them result in noticeably blurrier results than are possible with static reconstructions, especially when strong motion is involved.

3. Methodology

Figure 1 summarizes our method, EgoGaussian, for dynamic scene reconstruction from RGB egocentric videos. Our central idea is to identify dynamic objects and train them separately from the background. To do so, we develop a framework that fully integrates the dynamic characteristics of object interactions by temporally segmenting the video into static and dynamic clips and applying existing hand-object interaction modeling techniques to achieve 2D spatial distinction. Specifically, EgoGaussian first obtains camera poses and hand-object segmentation masks and the segmentation masks are further used to partition the videos into static and dynamic clips (Sec. 3.2). The static clips are used to initialize the static background and object shapes (Sec. 3.3), while the dynamic clips are used to track object motion and gradually refine their shapes (Sec. 3.4).

3.1. Preliminary: 3D Gaussian Splatting

We use 3D Gaussian Splatting (3D-GS) as our modeling structure because it provides an explicit 3D scene representation with a set of point-cloud-like 3D Gaussians. Each Gaussian is characterized by a position (mean) μ, a covariance matrix Σ, an opacity, and color features $\mathbf{c}$. The Gaussians are defined using the standard multivariate Gaussian distribution $G(\mathbf{x}) = e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$, providing a flexible optimization framework and a compact 3D scene representation.

3D Gaussian Splatting utilizes a differentiable point-based α-blending rendering to compute the color C of pixel $\mathbf{x}_p$. Specifically, it adapts a typical neural point-based approach and blends N ordered points overlapping the pixel: $C(\mathbf{x}_p) = \sum_{i \in \mathcal{N}} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j)$, where α_i is calculated by evaluating a 2D Gaussian with covariance Σ, projected from the 3D Gaussian, and then multiplied with its opacity; and $\mathbf{c}_i$ is the color of each Gaussian. The original 3D-GS implementation treats color as a directional appearance component represented via spherical harmonics (SH). For simplicity, we disable the view-dependent color by setting the maximum SH degree to 0.

3.2. Data Preprocessing

As pointed out by previous work [21], 3D-GS tends to overfit to training views and generate excessive floaters when there are scene inconsistencies among 3D views. In order to eliminate such inconsistencies, our idea is to identify any objects that move at all in the scene and separate them from the static background. In this preprocessing step, we adopt an initialisation method similar to other 3D-GS approaches and process the input to separate dynamic contents from static ones.

Initialisation for 3D-GS. Following existing 3D-GS methods, we first use COLMAP [50, 51] to estimate camera poses. COLMAP's SfM also creates a sparse point cloud corresponding to the camera poses estimated, and we use them as an initialization for 3D Gaussian Splatting. For egocentric dataset where camera poses are available, e.g., EPIC Fields [57] provides estimated camera poses for EPIC-KITCHENS [11], we employ them directly.

Separation of Dynamic and Static. To separate dynamic and static information in 2D frames, we use off-the-shelf approaches to obtain segmentation masks of hand-object interaction. Specifically, we use EgoHOS [74] to generate hand masks and Track-Anything model [66] for object masks and human body masks. Furthermore, these masks are dilated by 2 pixels for better robustness. The onset and offset frames of each hand-object interaction are estimated throughout the video. We then partition the egocentric video along the temporal axis into *static* and *dynamic* clips according to the onset and offset of interactions.

3.3. Static Reconstruction

Given a static video clip, our goal is to obtain a 3D reconstruction that distinguishes between the static background and dynamic objects involved in interactions within the dynamic clips. We begin by training a static representation of the scene and then identify dynamic objects using information extracted from the adjacent dynamic clips.

Initial Static Reconstruction. We have a set of T observations/frames from the static clip $S = \{\mathbf{I}_t, \mathbf{M}_{\text{body},t}, \theta_t | t = 1, \dots, T\}$ as input, where $\mathbf{I}_t$ is an input RGB egocentric frame, $\mathbf{M}_{\text{body},t}$ is the binary hand/body segmentation mask where pixel value $= 0$ represents body part and pixel value $= 1$ is for rest of the frame, and θ_t is the corresponding camera parameters for frame t. We follow a similar optimization pipeline as the original 3DGS [28], including both pruning and densification but use a masked version of the loss function:

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 (\mathbf{I}_{\text{input}}, \mathbf{I}_{\text{render}}) + \lambda \mathcal{L}_{\text{D-SSIM}} (\mathbf{I}_{\text{input}}, \mathbf{I}_{\text{render}}),$$

with the gradients zeroed out according to the mask $\mathbf{M}_{\text{body}}$. Similar to SuGaR [22], after around 30K iterations, we append an additional entropy loss on the opacity α of Gaussians, i.e.

$$\mathcal{L}_{\text{entropy}_\alpha} = -\alpha \log(\alpha) - (1 - \alpha) \log(1 - \alpha),$$

as a way to enforce Gaussians to be either fully transparent or completely opaque and train for another 10K iterations while disabling pruning and densification. Instead, we prune the transparent Gaussians once at the end of this phase of training. This produces a set of 3D Gaussians $\mathcal{G}$ reconstructing the scene captured by this static clip, which includes both the static background and any objects that might move during dynamic portions of the video.

Dynamic Object Identification. To identify the set of Gaussians associated with dynamic objects, we extract interaction information from nearby dynamic clips and aim to generate masks for objects that have moved or will move. We automatically generate such masks by selecting a random point from the object mask during interaction within the dynamic clips and using it as a prompt for the Track-Anything model [66]. The initial static reconstruction from the previous step allows us to lift 2D masks to a dense 3D reconstruction of the scene, and conversely, project 3D Gaussian points to 2D. Experimental results show that object masks from just N static frames adjacent to the dynamic clips are sufficient to obtain reliable 3D segmentation of the object. For example, consider a static clip S with T frames immediately preceding a dynamic clip where an object is moved by the camera wearer. We obtain segmentation masks for the object in the last N frames of this static clip, i.e., $\{\mathbf{M}_{\text{obj},T-N}, \dots \mathbf{M}_{\text{obj},T}\}$, where a pixel value of $= 1$ indicates the targeted object and $= 0$ represents the rest of the frame. We set $N = 5$ in our experiments.

Similar to Gaussian Grouping [69], though more streamline, an additional trainable parameter of label l is then attached to each Gaussian and initialized to a very small value. This label can then be rendered similar to the RGB value as: $L(\mathbf{x}_p) = \sum_{i \in \mathcal{N}} l_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j)$. This produces a segmentation $\mathbf{L}_{\text{r}}$, upon which we can apply a binary cross entropy loss using the object masks $\mathbf{M}_{\text{obj}}$:

$$\mathcal{L}_{\text{BCE}_l} = -\left[\mathbf{M}_{\text{obj}} \cdot \ln (\sigma(\mathbf{L}_{\text{r}})) + (1 - \mathbf{M}_{\text{obj}}) \cdot \ln (1 - \sigma(\mathbf{L}_{\text{r}}))\right]$$

where $\sigma(X) = \frac{1}{1+e^{-X}}$ is the sigmoid activation function. Training the 3D Gaussians with respect to this loss function while freezing all parameters except for l allows us to

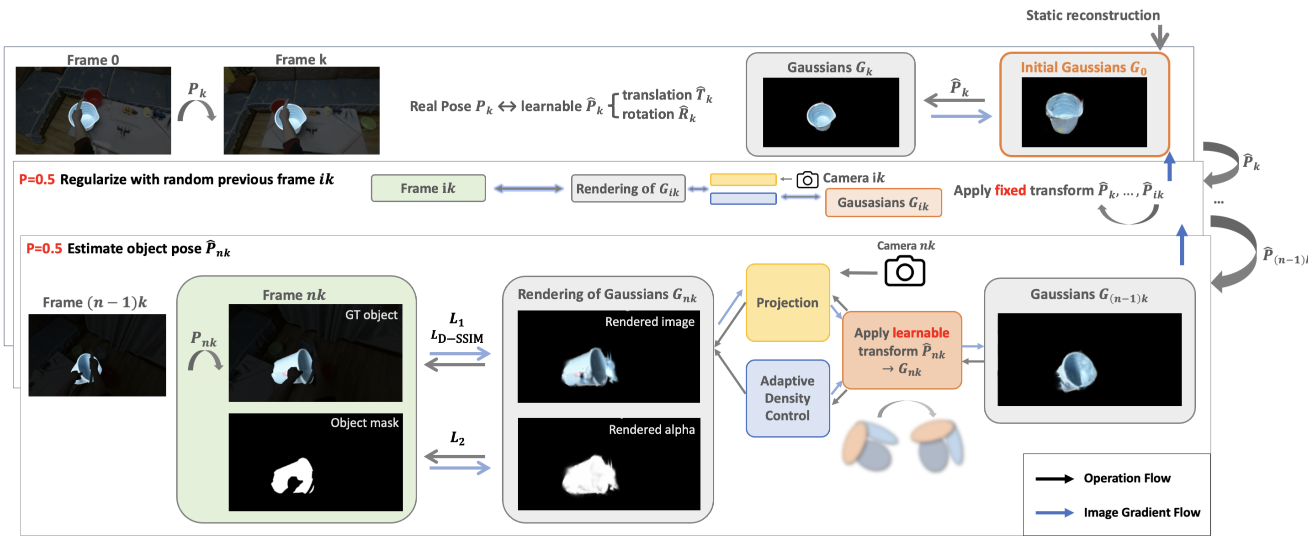

Figure 2. **Dynamic Object Modeling Pipeline**. We use a sequential pipeline with regularization from previous frames, allowing us to estimate object poses and iteratively refine their shapes simultaneously.

separate the Gaussians into object $\mathcal{G}_{\text{obj}}$ and background $\mathcal{G}_{\text{bg}}$ based on thresholding l and such 3D segmentation can be used to render targeted object masks from more viewpoints. More details and graphical illustration are in the Supp.

3.4. Dynamic Reconstruction

Given the reconstructed background $\mathcal{G}_{\text{bg}}$ and the initial object $\mathcal{G}_{\text{obj}}$ using the static clips, we refine the object shapes, track their motion through the dynamic clips and update the background with new information revealed by the movement of the dynamic objects.

Object Pose Estimation. We use the set of object Gaussians $\mathcal{G}_{\text{obj}}$ as an initial estimate of the object appearance and estimate its pose for every k-th frame in the dynamic clip. Specifically, we estimate for a targeted frame f_t a corresponding relative pose $\mathbf{P}_t$ from a previous state at f_{t-k}. We further decompose the pose $\mathbf{P}_t$ into a 3D translation vector T_t and a rotation matrix $\mathbf{R}_t$. Unlike previous dynamic 3D-GS methods [36, 63], EgoGaussian applies one set of transformation parameters to the whole collection of object Gaussians $\mathcal{G}_{\text{obj}}$ as a whole, treating it as a single rigid object.

We optimize the rotation $\mathbf{R}_t$ using the 6D continuous rotation representation $\tilde{\mathbf{R}}_t$ proposed by [77]. To ensure the transformation is rigid, an estimated 3×3 rotation matrix must be $\mathbf{R}_t \in SO(3)$. Hence for each target frame f_t, we apply the estimated translation and rotation parameters to the 3D center of each Gaussian:

$$\mathbf{X}_{\mathcal{G}_{\text{obj},t}} = \mathbf{X}_{\mathcal{G}_{\text{obj},t-k}} \cdot g(\tilde{\mathbf{R}}_t) + \mathbf{t}_t,$$

where $\mathbf{X}_{\mathcal{G}_{\text{obj},t}}$ is the 3D coordinates of object Gaussians $\mathcal{G}_{\text{obj},t}$ at time t and $\mathbf{X}_{\mathcal{G}_{\text{obj},t-k}}$ at time $t-k$. $g(\cdot)$ is a function defined in [77] that transforms the 6D representation of rotation to a standard 3×3 rotation matrix. Such object rotation is also applied to the anisotropic covariance of each 3D Gaussian to regularize its alignment with the object surface:

$$\Sigma' = \mathbf{R}\Sigma\mathbf{R}^T = (\mathbf{R}_t\mathcal{R})\mathcal{S}\mathcal{S}^T(\mathbf{R}_t\mathcal{R})^T,$$

where the covariance matrix Σ is decomposed into the scaling vector $\mathcal{S}$ and the Gaussian's rotation matrix $\mathcal{R}$ constructed from quaternion.

Object Shape Refinement. In order to better reconstruct the shape of the object, we apply a silhouette loss onto the computed alpha value using the following equation: $A(\mathbf{x}_p) = \sum_{i \in \mathcal{N}} \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j)$, which effectively equates the RGB rendering equation without color.

Dynamic Object Reconstruction. Our final dynamic object reconstruction loss is then:

$$\mathcal{L}_{\text{obj}} = \mathcal{L}_1\left(\mathbf{I}_{\text{obj}}, \mathbf{I}_{\text{render},\mathcal{G}_{\text{obj}}}\right) + \lambda\mathcal{L}_2\left(\mathbf{M}_{\text{obj}}, \mathbf{A}_{\text{render},\mathcal{G}_{\text{obj}}}\right),$$

where $\mathbf{I}_{\text{obj}}$ is cropped from $\mathbf{I}_{\text{input}}$ with the object segmentation mask $\mathbf{M}_{\text{obj}}$. $\mathbf{I}_{\text{render},\mathcal{G}_{\text{obj}}}$ is rendered from $\mathcal{G}_{\text{obj}}$ so it contains the object only and black background. $\mathbf{A}_{\text{render},\mathcal{G}_{\text{obj}}}$ is the rendered alpha. We experimentally observe that 0.5 is a suitable λ value. Additionally, we lower the learning rate on the Gaussians parameters, such as position and color, by a factor of 10 to prioritize the learning of object poses.

As we optimize the time-dependent pose parameters T_t and $\tilde{\mathbf{R}}_t$ one frame at a time, the Gaussians can easily overfit to the current frame. To address this, we train not only on the current frame, instead for every training iteration we train either on the current frame or a random previous frame with a probability of 0.5.

Static Scene Update. The static scene $\mathcal{G}_{\text{bg}}$ is reconstructed from the static clips, where parts of the background are obscured by objects that interact within the dynamic clips. To utilize the visible information from the dynamic clips, we retrain $\mathcal{G}_{\text{bg}}$ with the dynamic objects masked out.

Combination of Static Scene and Dynamic Objects. As a final step, we combine the object model $\mathcal{G}_{\text{obj}}$ with the full background model $\mathcal{G}_{\text{bg}}$. In practice, we note that at this stage, there are often floaters belonging to the background that obscure parts of the object. To eliminate these, we perform a final fine-tuning stage using all training frames and Gaussians. As we focus here on optimizing how the background and dynamic object interact and fit with each other, we again freeze the estimated per-frame object pose. This produces then the full scene reconstruction, including per-frame data of the object pose.

Training details. During the Gaussian splatting optimization process, the opacity is frequently set to zero in order to prune floaters. However, this would produce a very noisy signal for the pose optimization. As such, we instead alternate between optimizing the rigid object pose, and densifying/pruning the Gaussians. We first train for 4k iterations on every dynamic frame, optimizing the object poses without pruning or densification. We then freeze the object poses, and train another 4k iterations to better incorporate visual information with the estimated object poses, and finally train another 4k iterations optimizing the poses without densification or pruning. For all 12k iterations, all Gaussian parameters such as color and position are continuously optimized. After iterating through the whole dynamic clip with M frames, we obtain a coarse object pose for each frame $P = \{T_t, \mathbf{R}_t | t = 1, \ldots, M\}$. Finally, we perform one final round of joint training using all frames, with 6k iterations of pose estimation, 6k iterations of pruning/densifications, and finally another 6k iterations of pose estimation. This ensures that our object model is more equally fit onto all frames, rather than focused on the last seen ones.

4. Experiments

We compare our method with existing baselines for dynamic scene reconstruction where the goal is to reconstruct both static 3D scenes and dynamic objects from RGB egocentric videos. To quantitatively assess the quality of the reconstructed 4D scene, we follow the evaluation protocol of the novel view synthesis task using two different egocentric video datasets. We then present qualitative results of the reconstructed dynamics and conduct ablation studies on two key aspects of the proposed method.

4.1. Novel View Synthesis

Datasets. We evaluate our method on two commonly used egocentric video datasets HOI4D and EPIC-KITCHENS.

HOI4D [33] is a large-scale egocentric video dataset of human-object interactions consisting of 20 second-long videos. From this dataset, we randomly select 4 videos involving rigidly moving objects. Among these, 2 videos contain mostly translations, while the other 2 include both translations and rotations. Compared to the original dataset, we downsample the image resolution to one-quarter of its original size, resulting in a resolution of 480 x 270 pixels. Each video has a framerate of 15 FPS.

EPIC-KITCHENS [11] is a large-scale dataset featuring in-the-wild egocentric videos of human-object interactions in native kitchen environments. We again randomly select 4 video clips involving rigidly moving objects. Of these clips, 2 contain mostly translations, while the other 2 include both translations and rotations. Similar to the HOI4D dataset, we also downsample the image resolution to 455 x 256. The average length of these clips is 10.43 seconds with 60 FPS.

Evaluation protocol. For each video, we train using every second frame and evaluate on the rest. Although we are still able to correctly track and even reconstruct the moving object with fewer training frames (i.e. frames with larger step size), the rapid motion that comes with egocentric videos means that neither our method nor any baselines are able to properly reconstruct the background. We show examples with different step sizes in the Supp.

Metrics. To assess the performance of our model, we use the peak signal-to-noise ratio (PSNR), the structural similarity index (SSIM)[60], and the VGG-based perceptual similarity metric LPIPS [75]. As we aim to reconstruct the background and object without the actor, we mask out the arm and body of any actors within the scene when computing these metrics and only evaluate the quality of the object and background reconstruction.

Baselines. We compare our method with two state-of-the-art dynamic 3D-GS-based methods, Deformable 3DGS [68] and 4DGS [63], both of which apply deformation fields to model monocular dynamic scenes. However, both approaches are unable to properly utilize masks of image regions which should not be modeled, in this case the body of the camera wearer. To ensure a fair comparison, in addition to using the publicly available implementations of the compared methods, we modify them to support masking out gradients on the segmented human body, similar to our approach.

Results. Table 1 and Figure 3 compares our method with existing dynamic 3D-GS methods and their modified versions. We observe that EgoGaussian consistently outperforms existing methods across all evaluation metrics on two datasets, while the two SOTA methods perform comparably. Though the integration of gradient masks increases their performance, our method still achieves superior results across all evaluations in dynamic frames.

Method	HOI4D						Epic-Kitchen					
	Static			Dynamic			Static			Dynamic		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
4DGS [63]	0.88	25.33	0.13	0.89	25.34	0.13	0.84	26.84	0.20	0.79	22.54	0.24
4DGS w/o hands	*0.94*	*28.69*	*0.08*	*0.94*	*27.33*	*0.10*	**0.87**	**28.90**	**0.16**	0.80	23.13	0.23
Def-3DGS [68]	0.90	*25.85*	*0.11*	*0.90*	25.71	0.12	*0.86*	*27.35*	0.18	0.81	*23.15*	0.22
Def-3DGS w/o hands	*0.94*	28.09	*0.08*	*0.94*	26.92	*0.10*	*0.86*	27.63	*0.17*	*0.82*	*23.27*	*0.21*
Ours	**0.96**	**30.99**	**0.08**	**0.95**	**30.33**	**0.09**	0.85	*28.33*	0.19	**0.88**	**28.34**	**0.17**

Table 1. **Comparison with SOTA dynamic Gaussian Splatting methods.** We evaluate our method and two other SOTA baselines along with their modified versions with hands excluded from modeling on the HOI4D and EK datasets. The best and second best results are **bolded** and *italicized* respectively. We show the evaluation results on static and dynamic frames separately.

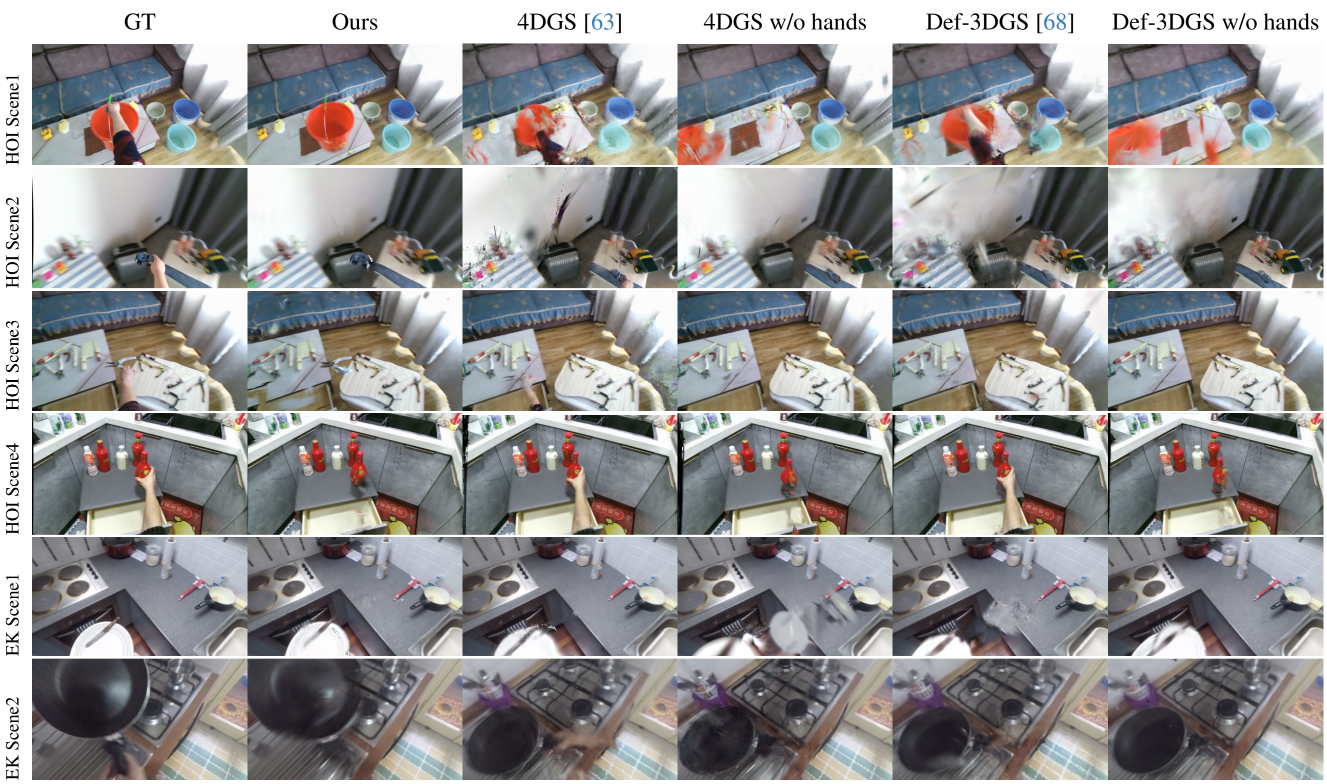

Figure 3. **Qualitative comparison with SOTA.** We show reconstructions produced by our method and SOTA baselines (4DGS [63] and Deformable 3DGS [68]) along with their modified versions from both HOI4D and EPIC-KITCHENS. Our reconstruction achieves greater accuracy, whereas baseline approaches fail to handle dynamic interactions even when hands are excluded during training.

4.2. Dynamic modeling

Figure 4 shows the estimated object trajectories and novel views rendered from arbitrary viewpoints. Demonstration videos and more visualizations are included in the Supp.

4.3. Ablation study

Estimate poses with larger time gap. We show that our chronological pose estimation schema is also able to model the dynamic object with larger time gaps k, by training on every 6 frames instead of every 2 frames. We can then estimate the state of object at each timestamp t through interpolation of transformation matrix constructed from relative pose $\mathbf{P}_t$ from $t-k$ to t. As seen in Table 2, although PSNR drops, we are still nonetheless able to produce an accurate reconstruction. Note that the metrics are computed only on the dynamic object itself.

Without full scene fine-tuning. We show the necessity of fine-tuning the static background and dynamic object as described in Section 3.4 jointly by comparing how our method performs when the background and object are only trained in isolation without fine-tuning on all frames or on the combined scene. As can be seen in Table 3 without full scene fine-tuning, quality drops significantly. This is partially due to 3D-GS being unable to distinguish between transparency

Method	HOI4D						Epic-Kitchen					
	Static			Dynamic			Static			Dynamic		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
With Original Step Size	**0.98**	**31.82**	**0.03**	**0.98**	**29.79**	0.04	**0.97**	**28.87**	**0.05**	**0.97**	**31.33**	**0.05**
With Larger Step Size	**0.98**	28.61	**0.03**	**0.98**	28.35	**0.03**	0.94	24.01	0.06	0.92	26.68	0.08

Table 2. **Ablation study** of step size over the object on object reconstruction.

Method	HOI4D						Epic-Kitchen					
	Static			Dynamic			Static			Dynamic		
	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↑	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓
With Fine-tuning	**0.96**	**31.52**	**0.07**	**0.95**	**30.29**	**0.09**	**0.94**	**34.22**	**0.10**	**0.88**	**28.30**	**0.17**
Without Fine-tuning	0.87	23.90	0.15	0.86	23.03	0.17	0.78	21.58	0.24	0.79	21.04	0.24

Table 3. **Ablation study** of full scene fine tuning.

Figure 4. **More qualitative results of the reconstructed dynamic scenes.** The left figure shows the object trajectory inferred from interpolated object poses. The right figure shows the rendering from arbitrary viewpoints.

Figure 5. Example of black artifacts

and blackness. By jointly training both object and background, we eliminate this uncertainty and such black artifacts. We show examples of this in Figure 5.

5. Conclusion and Discussion

We introduced EgoGaussian, a novel egocentric reconstruction method that is able to reconstruct rigid objects along with accompanying motion from egocentric data. We show significant improvements in terms of both dynamic object and background reconstruction quality compared to the state-of-the-art.

Although our method is able to well reconstruct rapid, rigid object motion, there are still a number of important limitations. As we require labels for both onset and offset of objects motion, as well as object masks, we require several additional offline data-preprocessing steps. Our method fundamentally relies on multiview-stereo data in order to reconstruct the scene geometry. As such, our method can encounter overfitting if the datasets have limited viewpoint coverage or lack features that can be tracked across time or view (e.g. uniformly textured walls). We utilize pixel-wise gradients to estimate per-frame object motion. As such, if the object goes out of frame, or there is too much motion between two frames, our method can lose track of the object entirely. The frame-by-frame dynamic object modelling requires significantly longer training compared to most previous dynamic 3D-GS-based methods.

Both motion labels and mask segmentations are parallel avenues of ongoing research, and as improvements are made, can easily be integrated into our pipeline. Similarly, as our method fundamentally trains using static or rigid sets of Gaussians, future improvements in sparse 3D Gaussian reconstruction could also be integrated into our pipeline. Although pixel-wise gradients prove to be sufficient in our setting, non-local methods of supervision such as optical flow would be an interesting avenue of research in order to improve robustness for large motions or when objects move out of frame. Such forms of supervision could also be utilized in order to estimate per-frame object pose more rapidly, allowing for reduced training time.

Acknowledgment This research was partially funded by the Swiss National Science Foundation Advanced Grant 216260: Beyond Frozen Worlds – Capturing Functional 3D Digital Twins from the Real World, and by the German Federal Ministry of Education and Research through the ExperTeam4KI funding program for UDance (Grant No. 01IS24064).

References

[1] Joao Pedro Araújo, Jiaman Li, Karthik Vetrivel, Rishi Agarwal, Jiajun Wu, Deepak Gopinath, Alexander William Clegg, and Karen Liu. Circle: Capture in rich contextual environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21211–21221, 2023. 1

[2] A. Bansal and M. Zollhoefer. Neural pixel composition for 3d-4d view synthesis from multi-views. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 290–299, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 3

[3] 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. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844, 2021. 2

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 2

[5] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2023. 3

[6] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12417–12426, 2021. 2

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

[8] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[9] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. *Advances in Neural Information Processing Systems*, 36:30453–30465, 2023. 2

[10] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020. 2

[11] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018. 1, 2, 4, 6

[12] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. 2

[13] Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, Sanja Fidler, David Fouhey, and Dima Damen. Epic-kitchens visor benchmark: Video segmentations and object relations. *Advances in Neural Information Processing Systems*, 35:13745–13758, 2022. 2

[14] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[15] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314, 2020. 3

[16] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Muhammed Kocabas, Xu Chen, Michael J Black, and Otmar Hilliges. HOLD: Category-agnostic 3d reconstruction of interacting hands and objects from video. 2024. 1, 2

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

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

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1, 2

[20] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez,

Gedas Bertasius, David Crandall, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C. V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives, 2023. 1

[21] Qiao Gu, Zhaoyang Lv, Duncan Frost, Simon Green, Julian Straub, and Chris Sweeney. Egolifter: Open-world 3d segmentation for egocentric perception. *arXiv preprint arXiv:2403.18118*, 2024. 4

[22] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *arXiv preprint arXiv:2311.12775*, 2023. 4

[23] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. *CoRR*, abs/2111.15234, 2021. 2

[24] Vladimir Guzov, Julian Chibane, Riccardo Marin, Yannan He, Torsten Sattler, and Gerard Pons-Moll. Interaction replica: Tracking human-object interaction and scene changes from human motion. *arXiv preprint arXiv:2205.02830*, 2022. 1

[25] Mohamed Hassan, Partha Ghosh, Joachim Tesch, Dimitrios Tzionas, and Michael J Black. Populating 3d scenes by learning human-scene interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14708–14718, 2021. 1

[26] Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. *ICCV*, 2021. 2

[27] Chun-Hao P. Huang, Hongwei Yi, Markus Höschle, Matvey Safroshkin, Tsvetelina Alexiadis, Senya Polikovsky, Daniel Scharstein, and Michael J. Black. Capturing and inferring dense full-body human-scene contact. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13274–13285, 2022. 1

[28] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 2023. 2, 4

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[30] Tobias Kirschstein, Shenhan Qian, Simon Giebenhain, Tim Walter, and Matthias Nießner. Nersemble: Multi-view radiance field reconstruction of human heads. *ACM Trans. Graph.*, 42(4), 2023. 3

[31] Jenny Lin, Xingwen Guo, Jingyu Shao, Chenfanfu Jiang, Yixin Zhu, and Song-Chun Zhu. A virtual reality platform for dynamic human-scene interaction. In *SIGGRAPH ASIA 2016 virtual reality meets physical reality: Modelling and simulating virtual humans and environments*, pages 1–4. 2016. 1

[32] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 1, 2

[33] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 6

[34] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 3

[35] Zhicheng Lu, Xiang Guo, Le Hui, Tianrui Chen, Ming Yang, Xiao Tang, Feng Zhu, and Yuchao Dai. 3d geometry-aware deformable gaussian splatting for dynamic view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 3

[36] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 1, 3, 5

[37] Zhaoyang Lv, Nickolas Charron, Pierre Moulon, Alexander Gamino, Cheng Peng, Chris Sweeney, Edward Miller, Huixuan Tang, Jeff Meissner, Jing Dong, et al. Aria everyday activities dataset. *arXiv preprint arXiv:2402.13349*, 2024. 1

[38] Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V. Sander. Deblur-nerf: Neural radiance fields from blurry images. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12851–12860, 2021. 2

[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 *CVPR*, 2021. 2

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

[41] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. NeRF in the dark: High dynamic range view synthesis from noisy raw images. *CVPR*, 2022. 2

[42] 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. 2

[43] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20133–20143, 2023. 1

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

[45] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), 2021. 3

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10313–10322, 2020. 3

[47] Wentian Qu, Zhaopeng Cui, Yinda Zhang, Chenyu Meng, Cuixia Ma, Xiaoming Deng, and Hongan Wang. Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15100–15111, 2023. 2

[48] Xiaofeng Ren and Chunhui Gu. Figure-ground segmentation improves handled object recognition in egocentric video. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3137–3144. IEEE, 2010. 2

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

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

[51] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 4

[52] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *ArXiv*, abs/2007.02442, 2020. 2

[53] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[54] Kiran Somasundaram, Jing Dong, Huixuan Tang, Julian Straub, Mingfei Yan, Michael Goesele, Jakob Julian Engel, Renzo De Nardi, and Richard Newcombe. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1

[55] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 2

[56] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019. 2

[57] Vadim Tschernezki, Ahmad Darkhalil, Zhifan Zhu, David Fouhey, Iro Laina, Diane Larlus, Dima Damen, and Andrea Vedaldi. Epic fields: Marrying 3d geometry and video understanding, 2024. 1, 4

[58] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR*, 2022. 2

[59] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19649–19659, 2022. 3

[60] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 6

[61] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Muller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. *CVPR*, 2023. 1, 2

[62] Yu-Shiang Wong, Changjian Li, Matthias Nießner, and Niloy J Mitra. Rigidfusion: Rgb-d scene reconstruction with rigidly-moving objects. In *Computer Graphics Forum*, pages 511–522. Wiley Online Library, 2021. 1

[63] 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. 3, 5, 6, 7

[64] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9416–9426, 2020. 3

[65] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5428–5438, 2022. 2

[66] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. Track anything: Segment anything meets videos, 2023. 4

[67] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 2

[68] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 3, 6, 7

[69] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. *arXiv preprint arXiv:2312.00732*, 2023. 4

[70] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 1, 2

[71] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *ICCV*, 2023. 2

[72] Hongwei Yi, Chun-Hao P. Huang, Shashank Tripathi, Lea Hering, Justus Thies, and Michael J. Black. MIME: Human-aware 3D scene generation. In *IEEE Conference on Com-*

puter Vision and Pattern Recognition (CVPR), pages 12965–12976, 2023. 1

[73] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5732–5741, 2021. 2

[74] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022. 2, 4

[75] 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. 6

[76] Kaifeng Zhao, Shaofei Wang, Yan Zhang, Thabo Beeler, and Siyu Tang. Compositional human-scene interaction synthesis with semantic control. In *European Conference on Computer Vision*, pages 311–327. Springer, 2022. 1

[77] Yi Zhou, Connelly Barnes, Lu Jingwan, Yang Jimei, and Li Hao. On the continuity of rotation representations in neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[78] Zhifan Zhu and Dima Damen. Get a grip: Reconstructing hand-object stable grasps in egocentric videos. *arXiv preprint arXiv:2312.15719*, 2023. 2