Proactive Scene Decomposition and Reconstruction

Baicheng Li[1,2] Zike Yan[3] Dong Wu[1,2] Hongbin Zha[1,2†]

[1]School of Intelligence Science and Technology, Peking University
[2]National Key Laboratory of General Artificial Intelligence [3]AIR, Tsinghua University

libc@pku.edu.cn yanzike@air.tsinghua.edu.cn riserwu@stu.pku.edu.cn zha@cis.pku.edu.cn

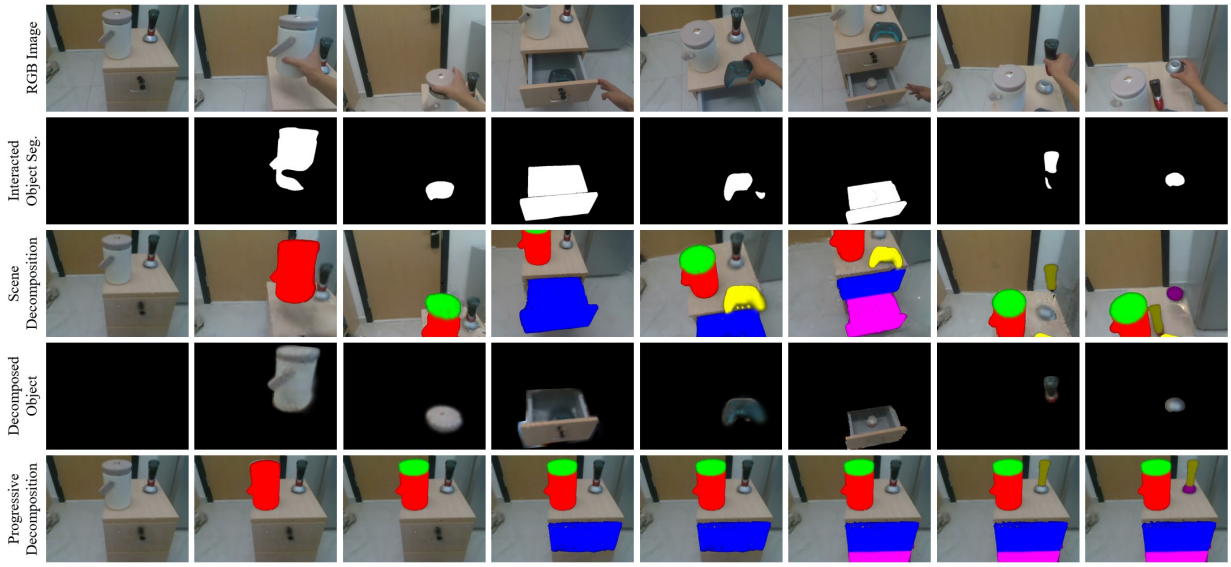

Figure 1. We introduce a dynamic SLAM system to tackle the proactive scene decomposition and reconstruction from egocentric live streams (first row). Only objects that are under proactive interactions will be decomposed (second) to maintain a consistent granularity. This fashion leads to photorealistic modeling of the environment (third to fifth rows), enabling progressive scene decomposition (fifth row) and robust object tracking (third and fourth rows).

Abstract

Human behaviors are the major causes of scene dynamics and inherently contain rich cues regarding the dynamics. This paper formalizes a new task of proactive scene decomposition and reconstruction, an online approach that leverages human-object interactions to iteratively disassemble and reconstruct the environment. By observing these intentional interactions, we can dynamically refine the decomposition and reconstruction process, addressing inherent ambiguities in static object-level reconstruction. The proposed system effectively integrates multiple tasks in dynamic environments such as accurate camera and object pose estimation, instance decomposition, and online map updating, capitalizing on cues from human-object interactions in egocentric live streams for a flexible, progressive alternative to conventional object-level reconstruction methods. Aided by the Gaussian splatting technique, accurate and consistent dynamic scene modeling is achieved with photorealistic and efficient rendering. The efficacy is validated in multiple real-world scenarios with promising advantages.

1. Introduction

Understanding the ever-changing environment is vital but fundamentally challenging for the vision and robotics communities. While many existing methods attempt to solve this problem by breaking down complex environments into manageable and semantically meaningful components, they often rely on passive data acquisition and pre-defined models to tackle 4D reconstruction or dynamic SLAM problems. While these methods can be effective in certain scenarios, they often struggle to capture the true dynamic nature of the environments, resulting in incomplete or inaccurate models. A key issue is the lack of consideration for human activity, which is often the dominant force in shap-

ing the dynamics of most real-world environments. Consequently, the methods miss out on valuable contextual insights derived from human-object interactions, which can provide critical context for understanding spatiotemporal relationships in a dynamic setting. By leveraging cues from these interactions, we can iteratively decompose independently moving regions, creating a more flexible and adaptive approach to modeling dynamic environments.

In this paper, we formalize a new task of *proactive scene decomposition and reconstruction*: an online process of dynamically disassembling and reassembling the environment given observations of ongoing human-object interactions. Traditional object-level scene reconstruction methods [2, 4, 14, 26, 38] have primarily focused on static environments, where the major challenge lies in the inherent ambiguity of decomposition. Besides, the intersected areas between objects lead to incomplete observations. Attempts have been made to either enforce consistency across views [2, 26] or perform inpainting for surface completion [16, 43]. However, as illustrated in Fig. 1, both the completion and the static decomposition are ill-posed. For instance, should we separate the drawer from the cabinet, or treat the contents of the drawer as part of the cabinet? What does the inside view look like if the drawer is closed during the data capture? We argue that effective decomposition should not be static, as the granularity is highly context-dependent. Instead, the decomposition process should be progressive and guided by interaction. Unlike 4D reconstruction or dynamic SLAM, which aim to handle arbitrary scene dynamics, we restrict the task to recovering a compositional scene representation from first-person live streams under intentional interactions, allowing for more controllable scene decomposition and accurate reconstruction by best exploiting the interaction cues.

Simultaneously addressing both scene decomposition and reconstruction in dynamic environments leads to a complicated system with the need for accurate camera pose estimation, object pose estimation, instance decomposition, and the fusion of past observations into a globally consistent map. The integration of these modules is usually difficult, as they are unstable and sensitive to outliers. However, as we demonstrate in the following section, hand-object interactions offer a stable and controllable definition of the compositional granularity as the individually moving part, allowing for progressively accurate object masks to be generated in a dynamic context. With this approach, the pose estimation task, both for the camera and the objects, becomes simplified, effectively reducing them to locally static problems that are trivial to solve, leading to accurate and holistic modeling of the environment.

For the proposed new task, we also introduce an online algorithm. Compared to offline methods, the online approach enables users to receive timely feedback, providing guidance during capturing and laying the foundation for incremental map updates. For streaming inputs, our method performs online camera pose tracking, object pose tracking, scene decomposition, and reconstruction. We fully leverage the interaction information available in egocentric inputs to achieve a high-quality object-level decomposed map reconstruction. To summarize, our main contributions include:

- We introduce a new task of proactive scene decomposition and reconstruction, aiming to decompose and reconstruct the environment based on online human-object interactions, offering a flexible alternative to conventional object-level reconstruction methods, allowing adaptive and progressive processes in response to the interaction cues.
- We propose an online dynamic SLAM system for proactive scene decomposition. Guided by the interaction priors, our system achieves more accurate scene decomposition, pose estimation, and reconstruction in an online fashion.
- We effectively combine the temporal constraints of foundation models and spatiotemporal consistency for modeling scene dynamics. The integration of both constraints along with fixed granularity induced by interactions enables our algorithm to achieve promising local homogeneity.

2. Related Work

2.1. Object-decomposed radiance fields

Recent advances in radiance fields have garnered widespread attention due to the photorealistic rendering results. As a global representation, decomposing the radiance field into individual components is one natural extension for downstream tasks that require local editing and reasoning, such as scene editing [43] and realistic simulation [27, 36]. [20] introduces a neural rendering technique and decomposes dynamic scenes with scene graphs. [35] designs a novel two-pathway architecture, where the scene branch encodes the geometry and appearance of the background, and the object branch encodes prior-conditioned learnable representations. ObjectSDF [31] and ObjectSDF++ [32] establish a connection between the semantics of each object and the corresponding geometry, enabling the creation of object-compositional neural implicit surfaces guided by RGB images and their corresponding instance masks. However, these methods generally require ground-truth instance masks and object association information as inputs.

To address the object decomposition problem, Panoptic Lifting [26] and Contrastive Lifting [2] adopt the linear assignment and contrastive learning to achieve object separation in 3D radiance fields given image segmentation predictions across views. With the emergence of Segment Any-

thing (SAM) [11] and the video segmentation models like SAM2 [23], the training of object-level radiance fields can be supervised directly from the predicted masks [4, 14, 38]. However, these methods often encounter issues due to ambiguous segmentation granularity. D2NeRF [33] and NeuralDiff [28]] attempt to decouple dynamic scenes through a simple motion segmentation, which can be defined precisely as the moving part of the environment. Additionally, some methods [30, 37] focus on modeling dense deformation fields to capture the spatiotemporal information of the scene. However, most existing methods neglect the strong cues inherited in the interaction between agent and environment. We share a similar motivation with a concurrent work of EgoGaussian [41], which leverages the hand-object interaction in egocentric videos for spatial-temporal modeling of dynamic environments and tracking rigid object motion. In contrast to the offline optimization process, our method extends the paradigm further by exploiting the instant feedback and temporal continuity within the streaming data to enable progressive scene decomposition and online holistic reconstruction of the dynamic environment.

2.2. Agent-in-the-loop scene understanding

Besides the passive scene understanding, agent-in-the-loop exploits the agent engagement to actively perceive and analyze the environment. For instance, Roboexp [10] introduces action-conditioned scene graphs, where robots accumulate information through active interactions to capture the geometry and the structure of the surroundings. Similarly, Nagarajan and Grauman [18] introduce affordance landscapes, enabling robots to learn about the actions that can be performed within a 3D environment. The approach helps robots recognize the potential for interaction with novel objects and enhance their ability to adapt to new environments. The cluttered environment always poses challenges for object recognition and segmentation. In [31], robust object recognition is achieved by combining perception with interaction. A similar system is adopted in Autoscanning [34] to couple scene reconstruction with proactive object analysis. In [17], scene segmentation is improved by selectively attending to certain areas, highlighting the role of fixation in active segmentation. Recent work has also delved into uncertainty-aware segmentation. Yu and Choi [39] propose a self-supervised method for interactive object segmentation through singulation-and-grasping approach. This demonstrates how robots can learn from their interactions without requiring explicit supervision, a significant step toward autonomous scene understanding. Fang et al. [7] explore how robots reduce the uncertainty of object segmentation through physical actions. This concept is also adopted in RISEG [22] by exploiting body frame-invariant features and robot interaction to correct inaccurate segmentation.

We refer readers to [3], a comprehensive review on interactive perception. This work sets a foundation for how robots can use their actions to improve scene understanding and vice versa. Besides the robot-in-the-loop scene understanding, human interventions also help to reduce the perception ambiguities. Many works study the visual perception in an egocentric video given hand-object interaction priors [6, 19, 40]. There are also studies working on 3D object decomposition through live annotations. iLabel [42] exploits the shared embedding space of jointly optimized neural fields, enabling efficient scene labeling given sparse clicks. Similarly, in Total-decom [15], extensive involvement of human labeling is reduced to enforce real-time control of quality and granularity of the scene decomposition with minimal interaction. These works collectively contribute to the advancement of agent-in-the-loop scene understanding, while we take a step further to directly perform online scene reconstruction and decomposition given an egocentric live stream of hand-object interaction. The scene is progressively decomposed and reconstructed in a unified SLAM system to jointly optimize scene radiance, camera motion, object poses, and instance segmentation.

2.3. Dynamic SLAM

The presence of dynamic objects introduces significant challenges to camera tracking and mapping as common consistency across views is assumed under static scenarios. The aim of dynamic SLAM is to remove features that violate the cross-view consistency constraints, ensuring precise camera tracking and reliable static map reconstruction. Relevant methods are commonly divided into two categories. The first approach utilizes warping or reprojection, as in [5, 21, 25], to detect inconsistencies in visual appearance or spatial geometry, thereby identifying dynamic regions in images. The second approach [8] leverages prior knowledge, such as semantic categories, to determine whether an object is dynamic. Some methods further combine these two approaches. DynaSLAM [1] and DRG-SLAM [29] remove features that belong to pre-defined categories or bypass geometric constraints. In contrast, SLAMANTIC [24] and CFP-SLAM [9] use projection to verify observations within pre-defined categories, selectively removing only those features that exhibit inconsistencies.

Our method can also be seen as a combination of these two approaches, leveraging both inconsistencies between observations and the map, and prior knowledge from user interactions to identify dynamic objects. However, unlike these existing methods, which primarily focus on reconstructing the static part of the environment, our approach not only reconstructs the static background but also decouples and reconstructs all interacted objects. This allows us to obtain a more informative and holistic understanding of the dynamic environment.

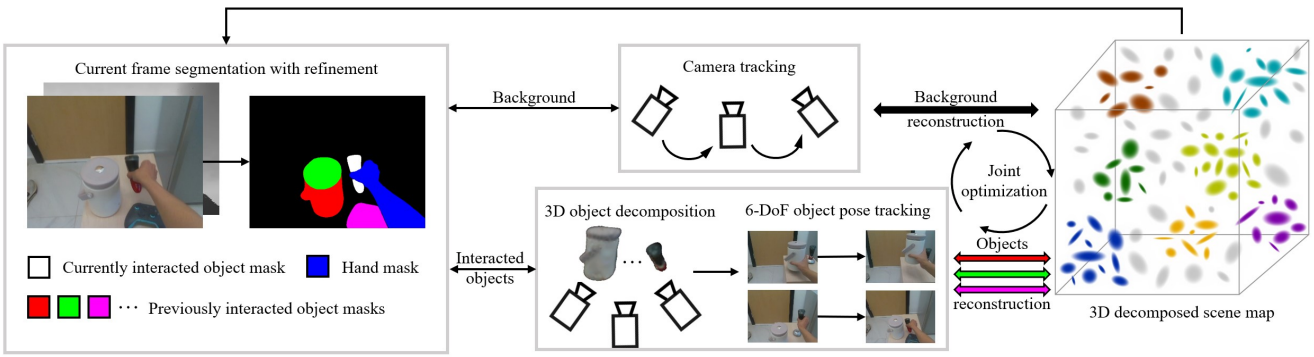

Figure 2. Overview of our method. With well-defined decomposition granularity induced by motion, the online system achieves reliable camera tracking and scene reconstruction, allowing progressive decomposition and robust instance tracking.

3. Overview

We aim to achieve online scene decomposition and reconstruction from egocentric RGB-D videos. The system takes streaming observations as inputs, where hand-object interactions are proactively carried out. The proposed system assumes that all objects moved by interactions exhibit approximately rigid body motion. The primary output is a scene representation that consists of background areas G_B along with the decomposed instances G_{O_i} $(i = 1, \ldots, n)$ as $\{G_B, G_{O_1}, \ldots, G_{O_n}\}$. The decomposition is carried out progressively, where a new instance will be initialized once the hand-object interaction is identified.

In practice, we maintain a Gaussian-based representation for better photorealism. To ensure fast convergence and prevent overfitting during the online optimization, the Gaussian primitive is parameterized by RGB color $\mathbf{c} \in \mathbb{R}^3$, center position $\mu \in \mathbb{R}^3$, isotropic variance $r \in \mathbb{R}$, and opacity $o \in \mathbb{R}$. Each object, once decomposed, is tracked and reconstructed independently. View-dependent color, depth, silhouette images and instance segmentation can be rendered using the Gaussian splatting technique as follows:

$$\hat{\mathbf{C}}[u,v] = \sum_{i \in N} \mathbf{c}_i f_i[u,v] \prod_{j=1}^{i-1} (1 - f_j[u,v]), \tag{1}$$

$$\hat{\mathbf{D}}[u,v] = \sum_{i \in N} d_i f_i[u,v] \prod_{j=1}^{i-1} (1 - f_j[u,v]), \tag{2}$$

$$\hat{\mathbf{S}}[u,v] = \sum_{i \in N} f_i[u,v] \prod_{j=1}^{i-1} (1 - f_j[u,v]), \tag{3}$$

$$\hat{\mathbf{I}}[u,v] = \sum_{i \in N} k_i f_i[u,v] \prod_{j=1}^{i-1} (1 - f_j[u,v]), \tag{4}$$

where k_i is the ID of the decomposed instance G_{O_i} the Gaussian belongs, $f_i[u,v]$ is computed as:

$$f[u,v] = o \exp\left(-\frac{\|[u,v] - \mu_{2D}\|^2}{2r_{2D}^2}\right), \tag{5}$$

$$\mu_{2D} = K\frac{\tilde{E}_t \mu}{d}, \quad r_{2D} = \frac{fr}{d}, \quad d = (\tilde{E}_t \mu)_z, \tag{6}$$

where $\tilde{E}_t$ represents the relative pose of the corresponding object with respect to the camera at time t.

The overarching goal of the scene decomposition and reconstruction is the joint optimization of camera pose, object pose, Gaussian parameters, and the assignment of instance labels:

$$L = \lambda_p L_p + \lambda_d L_d + \lambda_{ID} L_{ID}, \tag{7}$$

where L_p, L_d, L_{ID} are the expected L1 loss of color, depth, and instance segmentation given pixels within the mask M.

As illustrated in Fig. 2, the key to the problem, as also indicated in Eq. (7), is the decomposition that enforces the joint optimization of the map and poses as independent tasks for each decomposed instance under a local static assumption. We will show as follows that the priors originated from hand-object interactions and the spatiotemporal consistency maintained within the map jointly assures the accurate decoupling of objects from the background progressively, facilitating both pose estimation and scene reconstruction.

4. The Proactive Mapping System

We formulate the online scene decomposition and reconstruction under proactive hand-object interactions as an object-decomposed dynamic SLAM problem. The proposed system includes four modules: prompted segmentation, camera and object pose estimation, mask refinement, and decomposed scene reconstruction. The system undergoes iterative optimization that progressively decomposes instances under interactions and updates the locally independent maps.

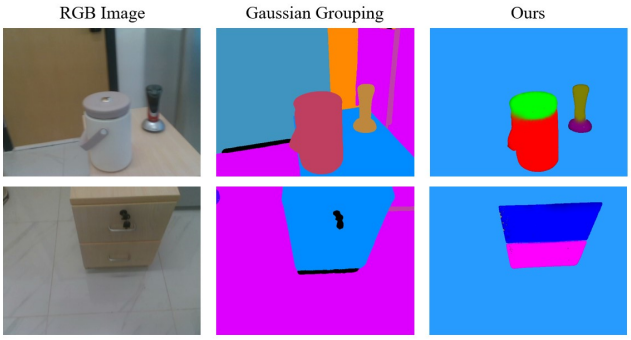

Figure 3. Comparison of scene decomposition with Gaussian Grouping [38]. Gaussian Grouping relies on the pre-defined segmentation granularity of the foundation models, whereas our method achieves adaptive decomposition given live interaction.

4.1. Prompted segmentation

As clarified above, we aim to maintain a fixed granularity for scene decomposition, defining it as the independently moving part. To achieve this, we extract information from the 3D scene map to determine the prompts for the segmentation module, effectively controlling the decomposition granularity. Specifically, as multi-view consistency only holds under static assumptions, the motion leads to inconsistency between the rendering results from the map and the instant observation. Similar to [12], we render a depth map based on the estimated camera pose and compute the differences compared to the observed depth map. Pixels with significant differences are regarded as inconsistent regions. Note that inconsistency may not only be caused by motion, but also by inaccurate pose estimation and map parameters. We adopt a filtering mechanism to divide the image into uniform grids and quantify the proportion of inconsistent pixels within each grid. A grid will be marked as dynamic if the portion exceeds a certain value:

$$\frac{\sum_{(u,v)\in S_{grid}} \mathbb{1}\left(\hat{D}[u,v] - D[u,v] > t_d\right)}{|S_{grid}|} > t_p, \qquad (8)$$

where t_d and t_p are hyperparameters for thresholding.

Subsequently, we detect connected marked grids to form coherent inconsistent area, then extract its centroid as SAM2 prompt for segmenting the interacted object. To validate segmentation accuracy, we concurrently execute hand localization via YOLO and SAM2, then verify spatial adjacency between the segmented object mask and hand region in both RGB and depth domains. This dual-space proximity check enforces physical interaction constraints to ensure logical segmentation results.

Note that SAM2 is a video segmentation model, where the mask decoder does not merely take encoded prompt features as input. A memory bank of previous observations is maintained and cross-attended with the prompted features for predicting the segmentation. Therefore, the prompted segmentation will only occur when a new instance undertakes the hand-object interactions for the first time. The following frames will track the activated instance through the cross-attended memory bank. To label the entire instance in the 3D space instead of merely the corresponding Gaussian primitives associated with the current view, we need to propagate the mask back to all past keyframes. We empirically find that the encoded prompt features can be directly utilized for segmenting previous keyframes as they remain temporally consistent across views for the same instance.

4.2. Camera and object pose estimation

In the online reconstruction setting, for each input frame, we estimate both the camera pose and the poses of all interacted objects, which are prerequisites for global scene reconstruction and object-level refinement. The optimization is performed over rotation and translation parameters, corresponding to the camera and interacted objects, respectively, and is guided by Eq. (7), with weights $\lambda_p^{ctrack}, \lambda_d^{ctrack}, \lambda_{ID}^{ctrack}$ for camera tracking and $\lambda_p^{otrack}, \lambda_d^{otrack}, \lambda_{ID}^{otrack}$ for object tracking.

A key difference between the two lies in the masking approach. Both employ a silhouette mask M_S to exclude previously unobserved pixels during optimization. However, to mitigate the interference of human and interacted object motion on camera localization, the camera pose estimation further incorporates the previously mentioned human mask M_h and interacted object mask M_o from the current frame, along with the rendered mask $M_{\hat{o}}$ of the interacted object from the scene map.

Once object tracking is complete, we follow the Gaussian Splatting-based SLAM approach to densify previously unobserved regions using depth information. For background areas, we directly initialize new Gaussians at the corresponding positions, while for regions of the interacted object, we warp the positions back to their expected locations based on the current estimated object pose.

4.3. Segmentation refinement

Though SAM2 achieves promising results for video segmentation, the spatial consistency is not well exploited due to the image domain inputs. As illustrated in Fig. 4, we notice typical failure modes during the proactive interactions. Benefiting from the unified framework to keep track of the entire sets of instances within the environment, the dense SLAM system is complementary to handle these failures.

One typical issue is that objects may be partially or fully outside the camera's field of view due to factors like camera angles or hand occlusion. Thanks to the photorealistic and efficient rendering of Gaussian primitives, we can assign a virtual camera to check if the instance is fully within

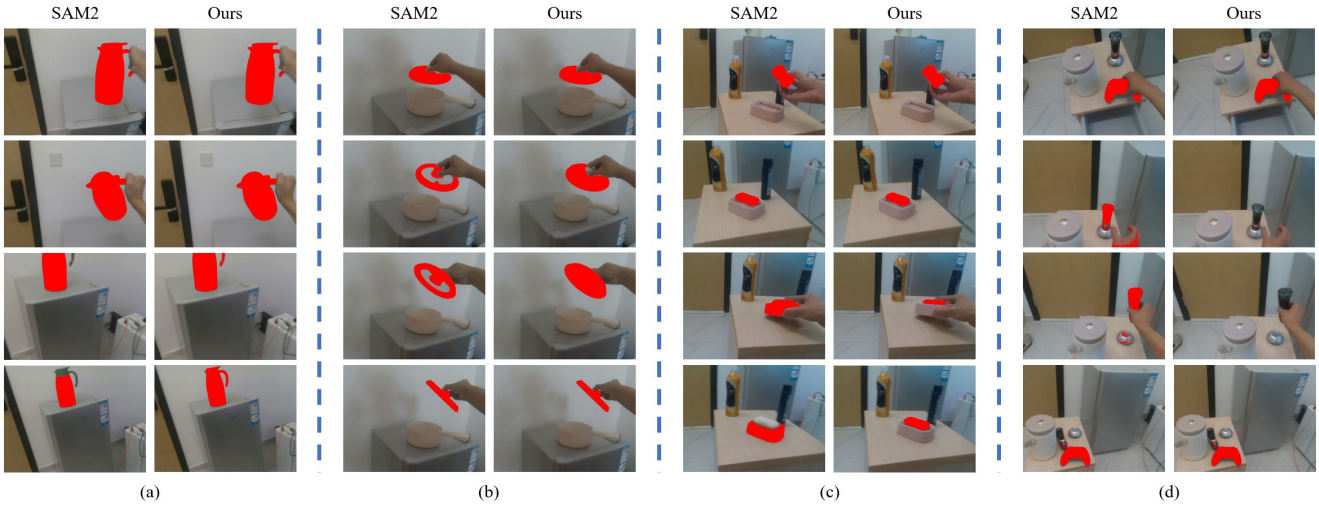

Figure 4. Comparison of segmentation results between SAM2 [23] and our method given instances under proactive interactions.

the field of view. As illustrated in Fig. 4(a), the kettle lid is erroneously excluded from the segmentation when the kettle reappears in the frame. To address this, we design a flexible-length memory bank to ensure that at least one complete observation of the object is retained in the memory queue. Based on the current state of the memory bank, we dynamically adjust the length of the memory queue to best retain the most complete observations. This strategy effectively mitigates segmentation errors caused by incomplete or occluded observations in specific periods of the video sequence.

Another issue is the inter-frame segmentation inconsistencies. The reliance on 2D information constrains the segmentation accuracy and temporal consistency. As illustrated in Fig. 4(b,c), we first check whether the previously identified inconsistent area is covered by the predicted mask. We then perform rigid object pose tracking and verify whether the rendered region of the interacted object matches the mask. Additional positive or negative point prompts will be added if the mask fails to cover the areas adequately or if it excessively overlaps with the rendered region. The conditions for mask refinement are defined as:

$$\left(\frac{S_{M\cap M_i}}{S_{M_i}} < t_{m_1}\right) \vee \left(\frac{S_{M\cap M_{\hat{o}}}}{S_{M_{\hat{o}}}} < t_{m_2}\right) \vee \left(\frac{S_{M\cap M_{\hat{o}}}}{S_{M}} < t_{m_3}\right). \tag{9}$$

Moreover, experiments show that SAM2 often produces noisy segmentation when the object is absent in certain frames. To prevent excessive refinement prompts, we first check for such cases. Assuming constant camera speed, if the object was absent in the previous frame and expected to stay out of view, any noisy segmentation is discarded, and the object mask is automatically set to zero.

4.4. Decomposed Scene Reconstruction

In the previous section, we outlined the process for obtaining accurate segmentation results for the interacted object in the 2D image. Now, we will focus on how to utilize these 2D segmentation results to construct and optimize our 3D decomposed map.

Progressive decomposition. Upon detecting a new instance that undergoes interactions, we decouple it from the original 3D map and represent it separately using a dedicated set of Gaussians. This process begins by extracting the object's mask from the current frame and propagating it to the past keyframes using the prompted segmentation method described earlier. With segmentation results available from multiple viewpoints, we project each Gaussian onto the 2D camera plane of these frames based on the estimated camera poses. As shown in Eq. (10), Gaussians $\tilde{g}$ that frequently appear within the mask are considered part of the object, and they are decoupled into an independent set, assigned a new object ID.

$$\frac{\sum_{f\in\mathcal{F}_{\text{valid}}} \mathbb{1}\left(P(\tilde{g}) \in M_r\right)}{|\mathcal{F}_{\text{valid}}|} > t_{3d}, \tag{10}$$

where $\mathcal{F}_{\text{valid}}$ represents the set of keyframes containing a complete observation of the interacted object, and M_r represents the refined mask in these keyframes.

Joint optimization. Once the camera and object poses are estimated, the next step is to perform global optimization to refine both object decomposition and reconstruction. The optimization is guided by a global objective function, as defined in Eq. (7), where the scene's color, depth, and object ID serve as supervision signals with weights λ_p^{map}, λ_d^{map} and λ_{ID}^{map}. For each training iteration, keyframes are sampled from the keyframe buffer and trained alongside the cur-

rent frame. All parameters, including camera poses, object poses, and Gaussian parameters, are jointly optimized, except for object poses outside interaction periods. This joint optimization serves a role similar to bundle adjustment in SLAM, where keyframe replay mitigates catastrophic forgetting and improves the consistency and quality of the reconstructed decomposed map.

5. Experiments

Our method aims to fully leverage the proactive interaction information contained in egocentric videos, while simultaneously achieving scene decomposition and reconstruction. To validate the effectiveness of our approach, we conducted evaluations from multiple perspectives. We tested our method on the HOI4D dataset [13], which is an egocentric dataset containing hand-object interactions. However, since the sequences in HOI4D contain only a small number of interacted objects, they do not adequately demonstrate our method's ability to accurately decompose scenes. Therefore, we propose a more challenging dataset, named the MHOI dataset, which contains ten egocentric RGB-D video sequences, each involving proactive interactions with 3 to 8 different objects. We conducted experiments on both datasets.

5.1. Experimental Setup

The experiments are performed on a desktop PC with an Intel i9-12900K CPU and an NVIDIA RTX 4090 GPU.

In our experimental implementation, we set the parameters as follows: loss function coefficients $\lambda_p^{ctrack} = \lambda_p^{otrack} = 0.5$, $\lambda_d^{ctrack} = \lambda_d^{otrack} = \lambda_p^{map} = \lambda_d^{map} = 1.0$, $\lambda_{ID}^{ctrack} = 0.0$, $\lambda_{ID}^{otrack} = \lambda_{ID}^{map} = 2.5$; thresholds for identifying interacted objects: $t_d = 0.3$, $t_p = 0.5$; thresholds for segmentation refinement: $t_{m_1} = t_{m_2} = 0.9$, $t_{m_3} = 0.7$; threshold for decoupling interacted objects from the background: $t_{3d} = 0.8$. For each incoming frame, we perform camera tracking and 6-DoF object tracking. Every 10 frames, we conduct joint optimization, and every 30 frames, we store the corresponding frame as a keyframe.

5.2. Segmentation and Decomposition

As shown in Fig. 1, our method achieves accurate segmentation of interacted objects and progressive scene decomposition. Unlike most current object-level scene reconstruction methods that rely on images captured under static conditions, our approach utilizes proactive interaction to clearly define segmentation granularity and eliminate its ambiguity. Fig. 3 clearly illustrates this point. Without leveraging motion information, Gaussian Grouping fails to separate components such as the thermal container and its lid, or the cabinet and its drawers. In contrast, our method accurately decouples each moving unit individually, which is more beneficial for downstream tasks like robotic manipulation.

In our workflow, we use depth inconsistencies as cues to generate prompts, applying SAM2 to obtain the mask of the interacted object and perform mask association. However, depending solely on the results from SAM2 is unreliable. Fig. 4 illustrates the significant role of our mask refinement by presenting several common failure cases of SAM2. As discussed in Sec. 4.3, to address the scenario depicted in Fig. 4 (a), we designed a flexible-length memory bank to mitigate the negative impact of problematic memory features on segmentation, thereby enabling a complete segmentation of the entire kettle. In Fig. 4 (b) and (c), errors are observed in the masks obtained by SAM2's mask association. To correct these segmentation errors, we compare the obtained masks with the inconsistency area and the rendered mask after object tracking, and then add new prompts to refine the results accordingly. Fig. 4 (d) demonstrates how SAM2 often produces noisy and incorrect segmentation when the object is entirely out of frame. To avoid adding excessive prompts, we determine the object's status based on its 3D position. If the object is projected outside the frame at a given moment, we directly assign a mask with all zeros.

We also conduct quantitative evaluations of decomposition on the HOI4D dataset, using the four sequences shown in Fig. 5. It can be clearly observed in Tab. 1 that our rendered masks are more accurate compared to those directly provided by SAM2, especially in Sequence 3, where the interacted object is a structurally complex pair of scissors. Without refinement, SAM2 often segments only the tip of the scissors, failing to capture the entire object.

Method	Seq 1	Seq 2	Seq 3	Seq 4
SAM2	0.913	0.884	0.318	0.941
Rendered Mask (Ours)	**0.925**	**0.920**	**0.835**	**0.947**

Table 1. Comparison of mask quality (mIoU) across sequences on HOI4D dataset.

Method	HOI4D			MHOI
	ATE	PSNR (s)	PSNR (d)	ATE
Co-SLAM	0.172	17.35	–	0.221
SplaTaM	0.156	18.61	–	0.293
NeuDySLAM	0.094	25.15	–	0.189
Ours	**0.076**	**29.12**	**27.58**	**0.093**

Table 2. Comparison with recent SLAM methods in terms of camera localization accuracy (ATE [m]) and rendering quality. Unlike NeuDySLAM and other dynamic SLAM methods that only reconstruct the static scene, our method also reconstructs the interacted objects that are in motion.

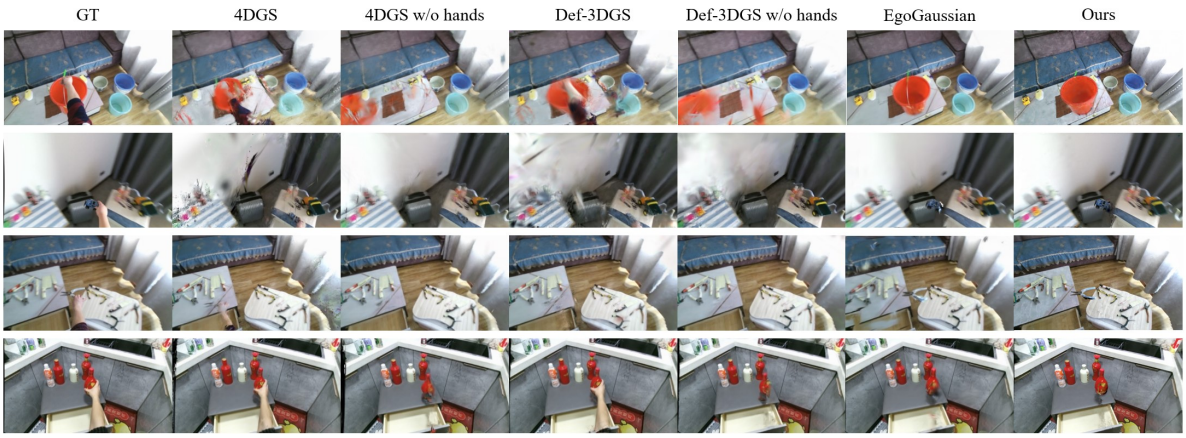

Figure 5. Qualitative comparison of rendering results on HOI4D dataset.

	Method	Static			Dynamic			Iterations
		SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	
Offline	4DGS [30]	0.88	25.33	0.13	0.89	25.34	0.13	30k
	4DGS w/o hands	0.94	28.69	**0.08**	0.94	27.33	0.10	30k
	Def-3DGS [37]	0.90	25.85	0.11	0.90	25.71	0.12	30k
	Def-3DGS w/o hands	0.94	28.09	**0.08**	0.94	26.92	0.10	30k
	EgoGaussian [41]	**0.96**	**30.99**	**0.08**	**0.95**	**30.33**	**0.09**	30k
	Ours	**0.96**	29.12	**0.08**	0.92	27.58	0.10	∼4k

Table 3. Quantitative comparison of novel view synthesis results with 4DGS, Def-3DGS, and EgoGaussian.

5.3. Camera Tracking and Scene Reconstruction

Camera tracking. We compared our method with three recent SLAM approaches: Co-SLAM and SplaTaM, which are based on a static scene assumption, and NeuDyS-LAM [12], the state-of-the-art NeRF-based dynamic SLAM method. As shown in Tab. 2, our approach achieves superior localization accuracy on both datasets. This difference is especially pronounced on the MHOI dataset, which contains multiple interacted objects. For NeuDySLAM, We attribute this to the fact that it masks out all objects that have experienced motion when optimizing the camera pose, whereas our method only masks out objects that are currently moving due to interaction. Therefore, when there are numerous interacted objects, NeuDySLAM discards too many useful features, resulting in a decrease in accuracy due to the lack of available features.

Scene reconstruction. We use novel view synthesis results to evaluate the quality of the reconstructed map. Fig. 1 and Fig. 5 show the qualitative results of our method on the MHOI and HOI4D datasets, respectively. Our method, as shown, produces high-quality rendering for both the background and interacted object parts. Moreover, it also achieves accurate object tracking[1]. These results demonstrate the effectiveness of our decomposed scene reconstruction. Notably, our method also possesses the capability to accurately reconstruct articulated objects and estimate their kinematics.[1].

We also perform quantitative evaluations on the HOI4D dataset, using the experimental settings from EgoGaussian [41] and including some results in Tab. 3 based on the original EgoGaussian experiment. Specifically, four sequences are selected to evaluate the rendering quality of both dynamic and static parts separately. We compare our method against 4DGS [30], Def-3DGS [37], and EgoGaussian. The first two methods are designed for non-rigid motion, whereas EgoGaussian and our method are targeted at scenes involving hand interactions with rigid objects. For a fairer comparison, modifications are made to the other two methods so that hands can be masked out.

As to the quantitative results, our method outperforms 4DGS and Def-3DGS while achieving metrics close to EgoGaussian, yet requires only a small fraction of the optimization iterations used by the other methods. Additionally, they rely on accurate camera poses as input, whereas our method performs its own camera tracking as a SLAM system.

6. Conclusion

In this paper, we introduce the task of proactive scene decomposition and reconstruction, which aims to adaptively decompose and reconstruct dynamic environments on the fly based on human-object interactions. To tackle the problem, we propose an online dynamic SLAM system that iteratively refines the map representation and the corresponding composition through interaction cues. Our approach is verified through experiments in camera pose estimation, object decomposition, and scene reconstruction, achieving high-quality and accurate modeling of dynamic environments. The results confirm the effectiveness of our system in capturing and representing the dynamic nature of the environment through proactive interactions.

[1] Visualization results are provided in the supplementary material

Acknowledgement

We gratefully acknowledge the anonymous reviewers and AC for their valuable comments and suggestions. This work is supported by NSFC (U22A2061, 62176010) and 230601GP0004.

References

[1] Berta Bescos, José M Fácil, Javier Civera, and José Neira. Dynaslam: Tracking, mapping, and inpainting in dynamic scenes. IEEE Robotics and Automation Letters, 3(4):4076–4083, 2018. 3

[2] Yash Bhalgat, Iro Laina, João F Henriques, Andrea Vedaldi, and Andrew Zisserman. Contrastive lift: 3d object instance segmentation by slow-fast contrastive fusion. In Advances in Neural Information Processing Systems (NIPS), 2024. 2

[3] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S Sukhatme. Interactive perception: Leveraging action in perception and perception in action. IEEE Trans. Robotics, 33 (6):1273–1291, 2017. 3

[4] Xiaokang Chen, Jiaxiang Tang, Diwen Wan, Jingbo Wang, and Gang Zeng. Interactive segment anything nerf with feature imitation. arXiv preprint arXiv:2305.16233, 2023. 2, 3

[5] Jiyu Cheng, Yuxiang Sun, and Max Q-H Meng. Improving monocular visual slam in dynamic environments: An optical-flow-based approach. Advanced Robotics, 33(12): 576–589, 2019. 3

[6] Tianyi Cheng, Dandan Shan, Ayda Hassen, Richard Higgins, and David Fouhey. Towards a richer 2d understanding of hands at scale. In Advances in Neural Information Processing Systems (NIPS), 2023. 3

[7] Xiaolin Fang, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Embodied uncertainty-aware object segmentation. In IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2024. 3

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pages 2961–2969, 2017. 3

[9] Xinggang Hu, Yunzhou Zhang, Zhenzhong Cao, Rong Ma, Yanmin Wu, Zhiqiang Deng, and Wenkai Sun. Cfp-slam: A real-time visual slam based on coarse-to-fine probability in dynamic environments. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 4399–4406. IEEE, 2022. 3

[10] Hanxiao Jiang, Binghao Huang, Ruihai Wu, Zhuoran Li, Shubham Garg, Hooshang Nayyeri, Shenlong Wang, and Yunzhu Li. Roboexp: Action-conditioned scene graph via interactive exploration for robotic manipulation. In Conf. on Robot Learning, 2024. 3

[11] 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 Intl. Conf. on Computer Vision (ICCV), pages 4015–4026, 2023. 3

[12] Baicheng Li, Zike Yan, Dong Wu, Hanqing Jiang, and Hongbin Zha. Learn to memorize and to forget: A continual learning perspective of dynamic slam. In European Conference on Computer Vision, pages 41–57. Springer, 2024. 5, 8

[13] 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, pages 21013–21022, 2022. 7

[14] Yichen Liu, Benran Hu, Chi-Keung Tang, and Yu-Wing Tai. Sanerf-hq: Segment anything for nerf in high quality. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 3216–3226, 2024. 2, 3

[15] Xiaoyang Lyu, Chirui Chang, Peng Dai, Yang-tian Sun, and Xiaojuan Qi. Total-decom: Decomposed 3d scene reconstruction with minimal interaction. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 20860–20869, 2024. 3

[16] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 20669–20679, 2023. 2

[17] Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. Active segmentation with fixation. In Intl. Conf. on Computer Vision (ICCV), pages 468–475. IEEE, 2009. 3

[18] Tushar Nagarajan and Kristen Grauman. Learning affordance landscapes for interaction exploration in 3d environments. In Advances in Neural Information Processing Systems (NIPS), pages 2005–2015, 2020. 3

[19] Supreeth Narasimhaswamy, Huy Anh Nguyen, Lihan Huang, and Minh Hoai. Hoist-former: Hand-held objects identification segmentation and tracking in the wild. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 2351–2361, 2024. 3

[20] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 2856–2865, 2021. 2

[21] Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguere, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 7855–7862. IEEE, 2019. 3

[22] H. H. Qian, Y. Lu, K. Ren, G. Wang, N. Khargonkar, Y. Xiang, and K. Hang. Riseg: Robot interactive object segmentation via body frame-invariant features. In IEEE Intl. Conf. on Robotics and Automation (ICRA), 2024. 3

[23] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714, 2024. 3, 6

[24] Matthias Schorghuber, Daniel Steininger, Yohann Cabon, Martin Humenberger, and Margrit Gelautz. Slamantic-

leveraging semantics to improve vslam in dynamic environments. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019. 3

[25] Raluca Scona, Mariano Jaimez, Yvan R Petillot, Maurice Fallon, and Daniel Cremers. Staticfusion: Background reconstruction for dense rgb-d slam in dynamic environments. In 2018 IEEE international conference on robotics and automation (ICRA), pages 3849–3856. IEEE, 2018. 3

[26] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 9043–9052, 2023. 2

[27] Marcel Torne, Anthony Simeonov, Zechu Li, April Chan, Tao Chen, Abhishek Gupta, and Pulkit Agrawal. Reconciling reality through simulation: A real-to-sim-to-real approach for robust manipulation. In Robotics: Science and Systems (RSS), 2024. 2

[28] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. Neuraldiff: Segmenting 3d objects that move in egocentric videos. In Intl. Conf. on 3D Vision (3DV), pages 910–919. IEEE, 2021. 3

[29] Yanan Wang, Kun Xu, Yaobin Tian, and Xilun Ding. Drg-slam: A semantic rgb-d slam using geometric features for indoor dynamic scene. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 1352–1359. IEEE, 2022. 3

[30] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20310–20320, 2024. 3, 8

[31] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In European Conf. on Computer Vision (ECCV), pages 197–213. Springer, 2022. 2, 3

[32] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In Intl. Conf. on Computer Vision (ICCV), pages 21764–21774, 2023. 2

[33] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D$\hat{2}$nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. In Advances in Neural Information Processing Systems (NIPS), pages 32653–32666, 2022. 3

[34] Kai Xu, Hui Huang, Yifei Shi, Hao Li, Pinxin Long, Jianong Caichen, Wei Sun, and Baoquan Chen. Autoscanning for coupled scene reconstruction and proactive object analysis. ACM Trans. Graphics, 34(6):1–14, 2015. 3

[35] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In Intl. Conf. on Computer Vision (ICCV), pages 13779–13788, 2021. 2

[36] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pages 1389–1399, 2023. 2

[37] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 20331–20341, 2024. 3, 8

[38] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In European Conf. on Computer Vision (ECCV), pages 162–179. Springer, 2025. 2, 3, 5

[39] Houjian Yu and Changhyun Choi. Self-supervised interactive object segmentation through a singulation-and-grasping approach. In European Conf. on Computer Vision (ECCV), pages 621–637. Springer, 2022. 3

[40] Chuhan Zhang, Ankush Gupta, and Andrew Zisserman. Helping hands: An object-aware ego-centric video recognition model. In Intl. Conf. on Computer Vision (ICCV), pages 13901–13912, 2023. 3

[41] Daiwei Zhang, Gengyan Li, Jiajie Li, Mickaël Bressieux, Otmar Hilliges, Marc Pollefeys, Luc Van Gool, and Xi Wang. Egogaussian: Dynamic scene understanding from egocentric video with 3d gaussian splatting. In Intl. Conf. on 3D Vision (3DV), 2025. 3, 8

[42] Shuaifeng Zhi, Edgar Sucar, Andre Mouton, Iain Haughton, Tristan Laidlow, and Andrew J Davison. ilabel: Revealing objects in neural fields. IEEE Robotics and Automation Letters, 8(2):832–839, 2022. 3

[43] Zhide Zhong, Jiakai Cao, Songen Gu, Sirui Xie, Liyi Luo, Hao Zhao, Guyue Zhou, Haoang Li, and Zike Yan. Structured-nerf: Hierarchical scene graph with neural representation. In European Conf. on Computer Vision (ECCV), pages 184–201. Springer, 2024. 2