GSGTrack: Gaussian Splatting-Guided Object Pose Tracking from RGB Videos

Zhiyuan Chen[1], Fan Lu[1], Guo Yu[1], Bin Li[1], Sanqing Qu[1],
Yuan Huang[2], Changhong Fu[1], Guang Chen[1]*

[1]Tongji University, [2]Control science and engineering, Beijing Institute of Control Engineering

Abstract

Tracking the 6DoF pose of unknown objects in monocular RGB video sequences is crucial for robotic manipulation. However, existing approaches typically rely on accurate depth information, which is non-trivial to obtain in real-world scenarios. Although depth estimation algorithms can be employed, geometric inaccuracy can lead to failures in RGBD-based pose tracking methods. To address this challenge, we introduce GSGTrack, a novel RGB-based pose tracking framework that jointly optimizes geometry and pose. Specifically, we adopt 3D Gaussian Splatting to create an optimizable 3D representation, which is learned simultaneously with a graph-based geometry optimization to capture the object's appearance features and refine its geometry. However, the joint optimization process is susceptible to perturbations from noisy pose and geometry data. Thus, we propose an object silhouette loss to address the issue of pixel-wise loss being overly sensitive to pose noise during tracking. To mitigate the geometric ambiguities caused by inaccurate depth information, we propose a geometry-consistent image pair selection strategy, which filters out low-confidence pairs and ensures robust geometric optimization. Extensive experiments on the OnePose and HO3D datasets demonstrate the effectiveness of GSGTrackin both 6DoF pose tracking and object reconstruction.

1. Introduction

6DoF object pose tracking aims to continuously estimate the 3D position and orientation of target objects from consecutive video sequences. This provides consistent and accurate positional information for objects being manipulated, which is essential for applications such as robotic manipulation [14, 39] and control planning [30].

Early 6DoF object pose estimation or tracking approaches assume access to 3D models [28, 42] or category templates [13, 33] and rely on feature matching algorithms

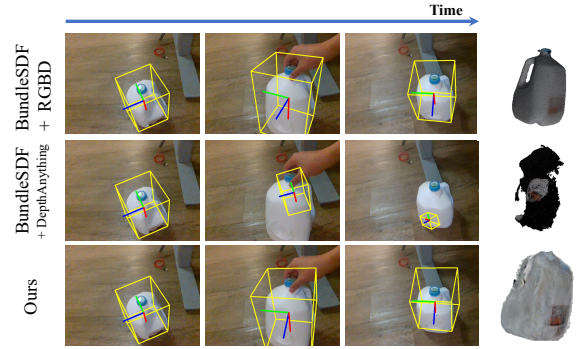*Corresponding author: guangchen@tongji.edu.cn

Figure 1. We are tackling a challenging problem: tracking 6DoF pose of unknown objects from RGB videos without accurate depth information. When applied to RGB videos with inaccurate estimated depth information [45], RGBD-based methods [40] degenerates quickly. In contrast, our method achieves robust tracking and reconstruction results.

to estimate the pose of the target object [29]. This reliance makes it extremely challenging for the models to generalize to novel, unseen objects. To achieve pose tracking of unknown objects, some studies have extended the concept of online localization from SLAM algorithms to pose tracking tasks [40, 44]. Given an RGBD video sequence, they project the 2D object into a 3D point cloud using accurate depth information, and employ point cloud registration to track the 6DoF pose. This pipeline inherently relies on accurate depth information for pose estimation. However, on most lightweight robots equipped with monocular vision systems, obtaining accurate depth information is usually not feasible, which poses significant challenges for the application of pose tracking algorithms [10].

To achieve 6DoF pose tracking using monocular RGB videos without depth information, a straightforward alternative is to use monocular depth estimation methods to obtain depth information [45, 52]. However, previous RGBD-based methods are fragile to depth noise [38, 40, 44]. As shown in Fig. 1, noisy depth information can lead to increased coarse pose errors in registration-based methods and introduce incorrect target information during pose optimization, resulting in a quick degeneration.

To this end, we propose a Gaussian Splatting Guided

object pose tracking framework, termed GSGTrack, which achieves RGB-based 6DoF object pose tracking by jointly optimizing pose and geometry. Specifically, we represent the object using 3D Gaussian Splatting (3DGS) [16], reformulating it as an online reconstruction pipeline that continuously reconstructs the object while guiding pose optimization through rendering losses. To enable accurate and robust object pose tracking even with inaccurate initial geometry, we design a graph-based geometric optimization method that jointly optimizes both poses and the 3D representation via an online geometric structure graph. However, the optimization process remains prone to noise in poses and geometry. To mitigate this, we design a differential silhouette loss and an outlier image pair pruning strategy, which leverages confidence metrics from pixel depth predictions, pose deviation from inertial data, and similarity between new and historical image geometry, enabling pruning of mismatched image pairs based on 3D geometric consistency.

To evaluate the proposed method, extensive experiments are conducted on two monocular RGB object pose tracking datasets, *i.e.*, OnePose dataset [21] and HO3D dataset [11]. The results demonstrate that the proposed method significantly outperforms existing approaches in terms of both accuracy and reconstruction quality. To summarize, our main contributions are as follows:

- We propose a novel framework for monocular RGB-basd 6DoF object pose tracking, which operates online and achieves robust pose tracking even with inaccurate geometric structures.
- We introduce a confidence-based pruning and optimization method for image pairs, effectively mitigating the impact of abnormal registration results on the global model.
- Extensive experiments show that GSGTrack significantly outperforms state-of-the-art methods on monocular RGB video object pose tracking datasets.

2. Related Work

6-DoF Object Pose Estimation and Tracking. Estimating the 6DoF pose of an object directly from RGB images is inherently an ill-posed problem, often requiring additional 3D information to resolve ambiguities. One approach introduces CAD models for offline training [15, 18, 27], but this limits generalization to novel objects. Some methods attempt to relax this with category templates [13, 33], yet their performance depends heavily on template accuracy, posing practical challenges. Other approaches leverages new view synthesis methods, such as Mask-RCNN [4], NeRF [19], or 3DGS [3, 22], to incorporate 3D shape information, achieving CAD-free pose tracking. However, they still require pre-captured reference views of the test object, which can be impractical in many scenarios. In 6DoF pose tracking, temporal information is used to estimate object poses across video frames. Some studies propose constructing 3D mod-

els from multi-view video frames to extend tracking to unknown objects [38, 40]. BundleSDF [40] is most similar to our approach, achieving pose tracking and reconstruction for unseen objects. Our method, however, integrates tracking and reconstruction with shape acquisition and optimization, enabling both accurate object pose tracking in RGB videos and improved appearance reconstruction. We further enhance reconstruction quality by incorporating generalized stereo matching [34] as a 3D prior.

Simultaneous Localization and Mapping Algorithms. RGB-SLAM algorithms primarily achieve camera pose estimation and scene map reconstruction from monocular RGB video sequences, addressing a problem similar to ours [17, 25]. However, RGB-SLAM algorithms are mainly applied to localization and mapping in large static scenes. Although some variants have extended SLAM to dynamic scenes [1, 2], these approaches typically mask dynamic objects from the scene, using the static portions to estimate camera poses and reconstruct the scene map. This limitation prevents them from handling the reconstruction of dynamic objects within a scene.Additionally, other research has introduced the concept of Object SLAM, where algorithms not only reconstruct the scene representation but also detect and recognize the semantics and basic appearance of objects [26, 46]. However, these algorithms cannot address scenarios involving dynamic interactions between objects and the environment, nor can they fully achieve 3D reconstruction of objects. In contrast, our method utilizes an innovative 3D Gaussian Splatting based representation technique . By integrating newly observed RGB images with existing Gaussian spheres, our approach generates a consistent 3D representation while simultaneously recovering 6DoF pose information of the object.

3D Reconstruction. Reconstructing 3D representations from 2D images has been widely studied with learning-based methods [5, 35, 47]. Recent advances in neural scene representations now allow high-quality 3D reconstructions [16, 24], though they typically assume known camera poses, limiting applicability. Some pose-free 3D reconstruction methods have emerged, but they focus mainly on static scenes, making them unsuitable for dynamic object interactions [9, 20, 48]. In particular, BundleSDF proposes an online approach that reconstructs 3D object meshes using SDF, relying on accurate depth information from RGBD images, but with limited appearance reconstruction [40]. In contrast, our method uses 3D Gaussian Splatting, supervised by RGB images, to achieve stronger appearance reconstruction. Another line of research leverages end-to-end feedforward neural networks for 3D scene reconstruction [50]. For example, Dust3R and similar methods [8, 34, 37] use generalized stereo networks to generate point clouds, while some 3D Gaussian Splatting approaches directly predict Gaussian attributes [51]. However, these
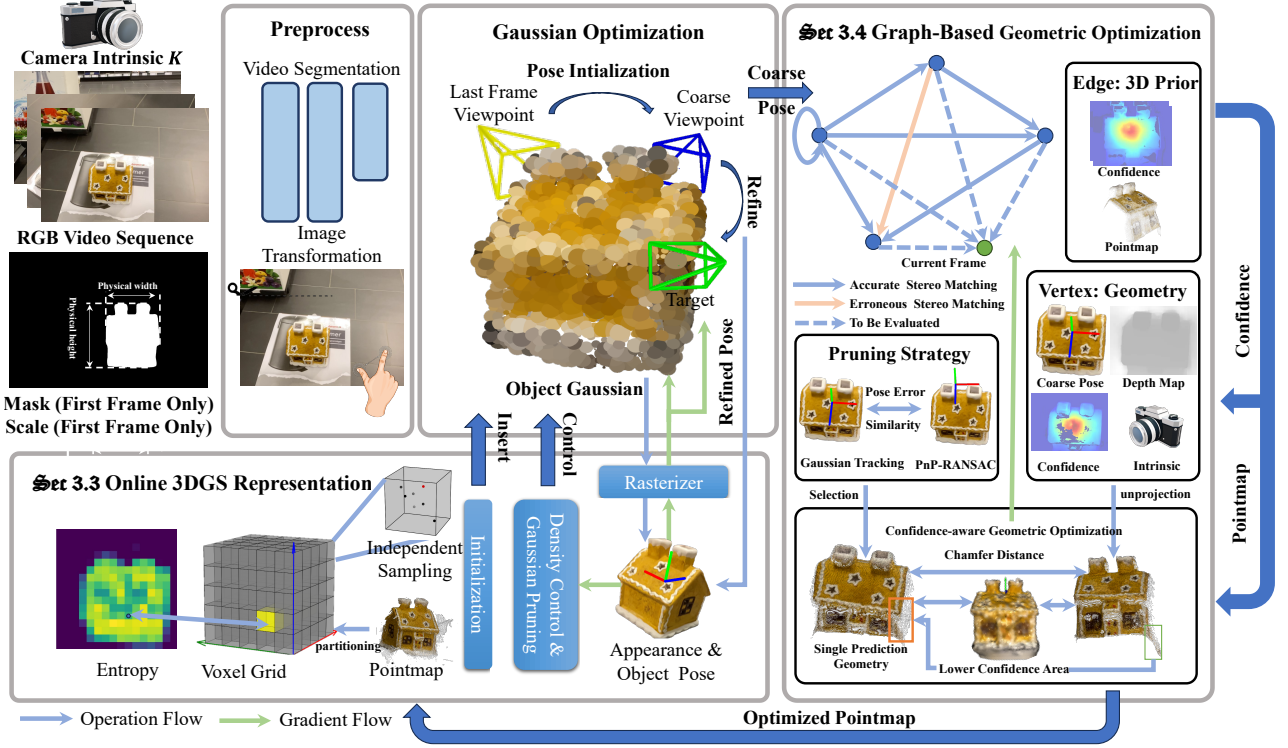
Figure 2. **Overview of our proposed GSGTrack.** To achieve accurate 6DoF object pose tracking without relying on precise depth information, we propose a joint optimization framework. Starting with a video sequence, we preprocess consecutive frames by generating object masks and estimating coarse geometry. Next, we introduce an online 3DGS representation that facilitates continuous object reconstruction from incoming video frames. Building on this 3D representation, we design a graph-based geometric optimization framework that refines both object pose and 3D structure through an online geometric structure graph. Additionally, we introduce an image pair pruning strategy and a confidence-aware geometric optimization technique to enhance the robustness and accuracy of the optimization process.

rely on accurate priors and are less effective in dynamic scenes. Our method instead employs a selective geometric optimization strategy, addressing challenges when priors are imprecise.

3. Methodology

3.1. Preliminary

Problem Formulation. In an object-centric dynamic scene, given a collected monocular RGB video sequence $F = \{F_0, F_1, \ldots, F_{n-1}\}$ $\left(F_t \in \mathbb{R}^{W \times H \times 3}\right)$, along with the segmentation mask o M_0 and the ground truth projected size $\mathbf{s}_0 = \begin{bmatrix} w_0 \\ h_0 \end{bmatrix} \in \mathbb{R}^2$ of the object *in the first frame only* as inputs, the goal of GSGTrack is to track the 6DoF pose of the object online while reconstructing a textured 3D model of the object.

Perliminary for 3DGS [16]. As mentioned before, we use 3D Gaussian Splatting (3DGS) as our basic object representation. 3DGS is a differential 3D representation for real-time neural rendering. Thanks to the explicit repre-

sentation, 3DGS enables fast 3D scene reconstruction and rendering, making it suitable as a basic 3D representation for object pose tracking. Specifically, 3DGS represents a scene as a set of anisotropic 3D Gaussian sphere. Each gaussian sphere is defined with a center μ_p, a covariance matrix Σ, a view-dependent color c, and a transparency α. For rendering, 3DGS project all 3D Gaussian spheres into 2D Gaussian distributions through a differentiable Gaussian splatting pipeline, and then blend the colors using fast alpha blending. The rendering process can be summarized as follows:

$$\boldsymbol{\mu}' = \pi\left(\boldsymbol{T} \cdot \boldsymbol{\mu}\right), \quad \boldsymbol{\Sigma}' = JW\boldsymbol{\Sigma}W^T J^T, \tag{1}$$

$$C = \sum_{i \in M} \mathcal{C}_i \alpha_i \prod_{j=1}^{i-1} \left(1 - \alpha_i\right), \tag{2}$$

where π is the projection operation, T is the camera pose of the viewpoint, W is the rotational part of T and J.

3.2. Joint Optimization Framework

To achieve accurate 6DoF object pose tracking and reconstruction under noisy geometric information, we propose a

joint optimization framework, which jointly optimizes object poses and 3D representation. Specifically, given consecutive video frames, we employ generalized stereo matching network [34] to estimate coarse geometric information. The estimated results include dense, pixel-aligned 3D pointmaps X_e^u and X_e^v in a shared local coordinate system O_e and also confidence maps C_e^u and C_e^v. Then, we introduce an Online 3DGS representation to enable continuous, real-time object reconstruction even with noisy poses and imprecise initial points (Sec. 3.3). To track object pose, for each video frame F_t, we first compute and optimize its 6DoF pose relative to the 3DGS by performing pose optimization where gradients are propagated solely to the pose parameters. This estimated pose serves as a coarse initialization for subsequent pose refinement and object reconstruction. Each frame is then integrated into an online geometric structure graph, where an image pair pruning strategy and a confidence-aware geometric optimization strategy are employed to fuse geometrically accurate historical frame information to estimate the geometry of the current frame (Sec. 3.4). Then, we simultaneously optimize the 3DGS and refine the object pose by incorporating both photometric and depth losses. Details of the gradient flow is provided in our supplementary material.

3.3. Online 3DGS Representation

Unlike traditional 3DGS, which reconstructs scenes from a fixed set of images, our approach performs reconstruction as an ongoing process, continuously incorporating new object views. To enable this, we developed an online 3DGS framework that supports dynamic Gaussian insertion and pruning under noisy pose and inaccurate points.

Gaussian Insertion. Due to the redundancy and high error rate in the initial pointmap [34], it is unsuitable for directly initializing Gaussian spheres. Conventional downsampling methods based on confidence [8] or random sampling often lead to gaps in the training results. To address this challenge, we propose a downsampling method that incorporates image complexity. Specifically, the 3D space is divided into a $K \times K \times K$ voxel grid, and the corresponding 2D image is partitioned into $K \times K$ squares. As shown in Eq. (3), we compute the image entropy E for each region to determine the number of sampling points $N \propto E$ for each voxel column.

$$E_{ij} = -\sum_{p=0}^{L-1} P_{ij}(p) \cdot \log_2 \left(P_{ij}(p) \right), \qquad (3)$$

where L represents the number of grayscale levels, and $P_{ij}(p)$ denotes the probability of a pixel having grayscale value p within the region $block_{ij}$.

Given that the pointmap is generated through unprojection from a dense 2D depth map, points in each voxel column typically fall within the same voxel. To ensure a rich

hierarchical structure in the sampled pointmap and to enhance the 3D representation capability of Gaussian spheres, we perform random pointmap interpolation along the negative direction of each point's normal vector within the pointmap. The maximum number of sampled points per voxel is set to $K/2$. Within each voxel, points are sampled randomly, with the sampling probability of each point proportional to its confidence level, which can be obtained form the generalized stereo matching network.

Gaussian Optimization. During the optimization of the 3DGS, we utilize RGB image to guide the photometric optimization of the rendering results. The photometric loss $\mathcal{L}_p$ can be represented as:

$$\mathcal{L}_p = \| I \left(\mathcal{G}, \boldsymbol{T} \right) - I_{gt} \|_1, \qquad (4)$$

where $I \left(\mathcal{G}, \boldsymbol{T} \right)$ represents the rendering result of the Gaussian model $\mathcal{G}$ from the viewpoint $\boldsymbol{T}$, while I_{gt} denotes the ground truth image.

We concurrently use the depth map from geometric optimization, which will be introduced in Sec. 3.4, to guide the geometric refinement of the 3D representation. This depth map is generated via alpha blending of depth data from Gaussian spheres. Unlike the depth losses in algorithms like SparseGS [43] and GS-SLAM [23], we address inherent errors in depth supervision signals by incorporating a depth confidence map derived from geometric structure optimization. The depth loss $\mathcal{L}_d$ can be formally represented as:

$$\mathcal{L}_D = \sum_{p \in \Omega} C_p \cdot (\sum_{i \in \mathcal{N}} z_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j)) - D_{gt}^p, \qquad (5)$$

where Ω represents the set of pixels, z_i represents the depth of the Gaussian sphere i, and α represents the transparency of the Gaussian sphere i.

Gaussian Pruning. To address geometric inaccuracies in initial pointmaps, inspired by [23], we prune erroneous points by employing a simple mask-based approach. Specifically, after each round of model training on F_t, we select several reference frames from past frames and remove any newly added Gaussian points that project outside the mask in the reference frames.

3.4. Graph-based Geometric Optimization

Geometric optimization seeks to achieve precise object pose tracking, even with imperfect geometric information. To accomplish this, we first perform object pose tracking and then construct an online geometric structure graph to further refine the tracked pose. In this process, we employ an image pair pruning strategy and confidence-aware geometric optimization to enhance the accuracy of the optimization.

Pose Initialization. A proper initial pose is essential for effective pose optimization and object reconstruction. Previous methods typically rely on point cloud registration to obtain the initial pose; however, this approach often fails

due to noise in the initial geometry. To address this, we propose an geometry-based strategy for object pose initialization. Specifically, we first compute the relative pose between the current frame F_t and the object's Gaussian representation $\mathcal{G}_{t-1}$, initializing the coarse pose $\tilde{\xi}_t$ with the pose ξ_{t-1} from the previous frame.

The process outlined above provides an initial pose estimation; however, this initial pose is often coarse and inaccurate. To refine the pose, we leverage image texture information for pose optimization. A straightforward approach is to minimize photometric loss $\mathcal{L}_p$; however, relying solely on photometric loss often causes pose optimization to converge on viewpoints with similar textures, resulting in significant errors. To address this, we propose incorporating silhouette loss to capture the object's geometric structure more accurately. While a typical approach is to calculate the Intersection over Union (IoU) between visible and segmented silhouettes [6, 49], this method can lack gradients when overlap is minimal. To overcome this limitation, we introduce a distance-based metric that weights the loss by computing each pixel's distance to the nearest silhouette edge. Our optimized silhouette loss is therefore formulated as:

$$\mathcal{L}_s = \frac{1}{|\Omega|} \sum_{p \in \Omega} \left(D_S(p) \cdot (1 - \tilde{S}(p)) + D_{\tilde{S}}(p) \cdot (1 - S(p)) \right),$$

(6)

where S and $\tilde{S}$ denote the binary masks for the ground truth and rendered image, D_S and $D_{\tilde{S}}$ correspond to the Euclidean distance transforms.

Online Geometric Structure Graph. As we mentioned before, we construct a geometric structure graph $\mathcal{H}$ for pose optimization. $\mathcal{H}$ is a directed graph, where each node v represents the geometric structure of the image frame F_v, including the 6DoF pose T_v and the 3D representation; each edge represents the results from the 3D generalized stereo matching network [34], including pixel-aligned 3D pointmaps X_e^u, X_e^v and confidence maps C_e^u, C_e^v of two frames. The graph $\mathcal{H}$ stores rich historical information to avoid tracking drift, besides, the storage of matching results also avoid repetitive computation.

Image Pair Pruning Strategy. To mitigate long-term tracking drift due to catastrophic forgetting, it is essential to store historical frame information and use multiple frames to jointly predict current frame geometry for accurate pose tracking and reconstruction. However, the 3D prior knowledge obtained through generalized stereo matching network predictions may be inaccurate. To address this issue, we propose an image pair pruning strategy, which removes mismatched results that do not satisfy geometric consistency from the global geometric structure graph. Specifically, we design three different strategies as following:

(1) Pose consistency-based pruning: To mitigate significant pose errors caused by object symmetry, we use the PnP-RANSAC algorithm to estimate the relative pose between the current frame and the reference frame within the local coordinate system. This estimated pose is then compared to the coarse pose obtained from tracking, with image pairs discarded if mismatches exceed the rotation threshold τ_r or the translation threshold τ_t.

(2) Geometry similarity-based pruning: To address geometry mispredictions in low-texture regions, we compare the predicted geometry of the reference frame with the actual reference structure at image pair nodes. This comparison uses the Chamfer Distance (CD) between point clouds to evaluate shape similarity and we filter out edges with low similarity to the reference nodes.

(3) Pixel credibility-based edge cropping: We calculate the confidence of edge (u, v) based on the confidence maps produced by the generalized stereo matching network:

$$\mu = \frac{1}{w \cdot h} \left(\sum_{i=1}^{w} \sum_{j=1}^{h} C_u^{(i,j)} \cdot M_u^{(i,j)} \times \sum_{j=1}^{h} C_v^{(i,j)} \cdot M_v^{(i,j)} \right),$$

(7)

where μ is the edge confidence, C is the confidence map, and M is the segmentation mask. A confidence threshold hyperparameter τ_c is introduced to prune edges that were not successfully matched.

Confidence-aware Geometric Optimization. In geometric graph pose optimization, we optimize only the current frame pose T_v to maintain consistency between the 3DGS and the graph poses, while historical keyframe poses are refined by the optimized Gaussian model. We simultaneously optimize the depth maps $D_i\{i \in [0, t]\}$ of all nodes and the edge transformation matrices T_{e2w} to align image pair points with the world coordinate system.

Our objective is to minimize the Chamfer Distance between graph nodes and geometry predicted by the generalized stereo matching network, thereby forming a consistent 3D representation. To reduce errors from inaccurate depth estimates, we incorporate the predicted confidence map into the loss function, weighting points by confidence. Thus, the geometry loss can be expressed as:

$$\mathcal{L}_{pg} = \sum_{e \in \mathcal{H}} \sum_{v \in \mathcal{E}_e} \sum_{i=1}^{hw} C_i^{v,e} \left\| \chi_i^v - T_{e2w} X_i^{v,e} \right\|,$$

(8)

where $\mathcal{H}$ denotes the structural graph, e represents the graph edges, $\mathcal{E}_e$ refers to the two nodes connected by a directed edge, C is the confidence map, χ_i^v is the pointmap of the node v, T_{e2w} is the transformation matrix from the edge coordinate system to the world coordinate system, and $X_i^{v,e}$ represents the pointmap of node v in edge e.

We employ the Gauss-Newton algorithm to optimize the structural graph, obtaining dense 2D-3D correspondences and the optimized object pose for the current frame. Geometric optimization is also performed on historical frames to correct potential errors in previous estimates.
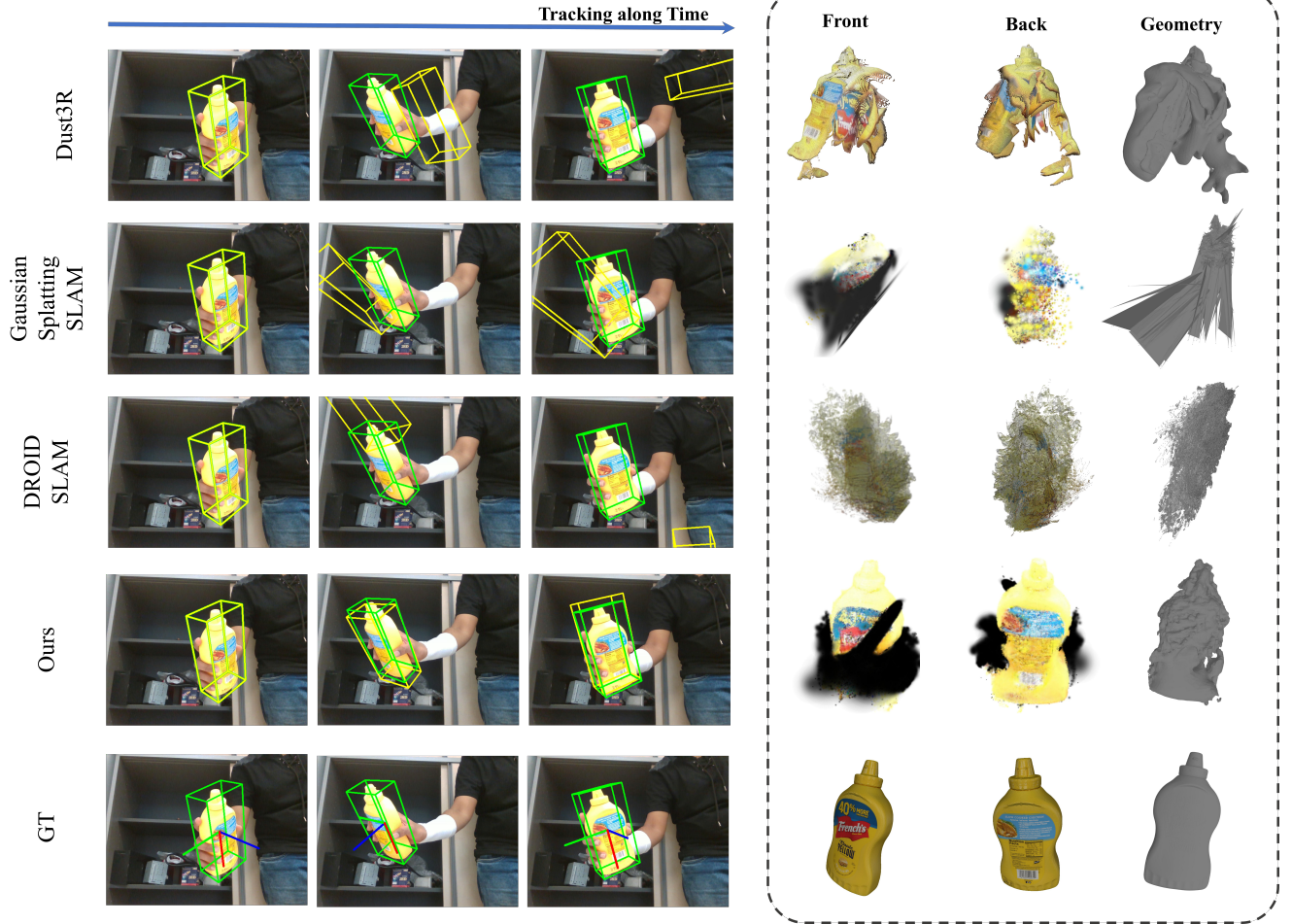
Figure 3. **Qualitative Comparison of GSGTrack and Baseline on HO3D.** Left: 6-DOF pose tracking with green and yellow boxes showing ground truth and estimated poses, respectively. Right: front and back views of reconstruction results, highlighting the object's geometric structure. Due to hand occlusions, black hand-shaped artifacts appear, obscuring parts of the object. Our reconstruction corrects the color divergence between ground truth and actual object colors seen in the video.

4. Experiments

4.1. Experimental Setup

Datasets. To evaluate our approach, we conduct extensive experiments on publicly available real-world datasets, OnePose [21] and HO3D [11], which include multiple dynamic, object-centered videos. The OnePose dataset provides 20FPS RGB videos of static objects at a resolution of 1920×1440. We utilize the SAM model to obtain object masks in the first frame and extracted scale information based on the provided 3D bounding box annotations. The HO3D dataset, on the other hand, contains RGBD videos centered on objects interacting with hands; we use only the RGB data for pose tracking and online reconstruction. We apply the annotated mask from BundleSDF bounding boxes and derive object scale estimates from ground-truth point clouds in the first frame.

Baselines. To comprehensively compare GSGTrack, we in-

clude various types of benchmarks. To better evaluate the pose tracking capabilities of algorithms on RGB images, we compare several algorithms, including Droid-SLAM [32], a deep learning-based approach that jointly optimizes camera poses and scene structure; Gaussian Splatting SLAM (GS-SLAM) algorithm [23], which serves as our primary comparative baseline. GS-SLAM is implemented based on 3DGS to achieve simultaneous localization and mapping. We utilize depth information generated through a generalized depth matching model [34] as input to assist Gaussian insertion in GS-SLAM. Droid-SLAM, on the other hand, inherently leverages optical flow and depth priors to assist sparse point cloud construction without requiring our depth prior information. Both algorithms are implemented using official open-source code.

To further compare reconstruction performance, we also evaluate against the Dust3R algorithm [34], which performs end-to-end 3D reconstruction based on generalized

Table 1. **Quantitative comparison on HO3D dataset.** We compared our method with baseline methods to evaluate the algorithm's capabilities in reconstruction and tracking.

Method	ADD-S(%)[0-0.3]m ↑					ADD(%)[0-0.3]m ↑				
	AP	MPM	SB	SM	Avg	AP	MPM	SB	SM	Avg
Dust3R [34]	17.32	24.97	24.25	32.12	24.67	9.64	15.51	16.51	19.76	15.36
Gaussian Splatting SLAM [23]	20.61	21.64	13.15	28.17	20.89	10.65	11.32	9.71	15.26	11.73
DROID-SLAM [32]	3.21	0.32	5.34	9.67	4.64	0.77	0.17	3.55	5.64	2.53
GSGTrack	**70.04**	**62.16**	**63.70**	**62.51**	**64.60**	**54.16**	**43.81**	**50.80**	**51.83**	**50.15**

Method	PSNR ↑					SSIM ↑					Reconstruction CD (cm) ↓				
	AP	MPM	SB	SM	Avg	AP	MPM	SB	SM	Avg	AP	MPM	SB	SM	Avg
Dust3R [34]	—	—	—	—	—	—	—	—	—	—	77.13	52.12	67.19	43.23	59.92
Gaussian Splatting SLAM [23]	18.57	20.13	17.89	20.50	19.27	0.79	0.82	0.77	0.82	0.80	85.14	69.49	80.23	60.40	73.82
DROID-SLAM [32]	—	—	—	—	—	—	—	—	—	—	150.33	130.80	81.86	100.87	115.97
GSGTrack	**26.70**	**24.83**	**25.20**	**27.04**	**25.92**	**0.97**	**0.96**	**0.95**	**0.97**	**0.97**	**23.93**	**15.72**	**21.39**	**19.20**	**20.06**

stereo matching priors. For fair comparison, we restrict Dust3R's image pairs to only use historical frames pointing to the current frame and prevent optimization of the historical frame's pose. Additionally, we compare the 3D reconstruction capabilities by using SfM [41] and 3DGS [16] algorithms on the static scene dataset, OnePose.

Metrics. We separately evaluate the quality of pose tracking and reconstruction accuracy. For 6-DOF object pose, we calculate the area under the ADD and ADD-S metrics curves (0 to 0.3m) using the actual object geometry [12, 42]. Given that the OnePose dataset lacks ground-truth object geometries, we indirectly measure the pose tracking accuracy for static objects by evaluating the precision of the camera trajectory [31]. For 3D reconstruction, PSNR and SSIM metrics are used to assess the quality of appearance [36]. Additionally, to evaluate the accuracy of object shape reconstruction, we compute the Chamfer Distance between the final reconstructed mesh and the ground-truth mesh defined in the reference coordinate system of the first video frame [40].

4.2. Implementation Details

For each video frame, we use object segmentation for scaling and cropping to focus on the object. Following 3DGS [16], both time-critical rasterization and gradient computation are implemented using CUDA. We implement the graph-based geometric optimization using PyTorch, with object-specific structural optimization carried out using the Adam optimizer. Coarse optimization runs for 300 iterations, followed by 125 pose refinement iterations per frame after pose estimation. All experiments use an NVIDIA GeForce RTX 3090. To ensure generalizability, We use the officially released network weights of the Dust3R algorithm, which are not trained on HO3D or OnePose datasets.

4.3. Results on the HO3D Dataset

The quantitative results of the comparison on the HO3D dataset are shown in Tab. 1. The proposed method demonstrates significant improvements in 6DoF object pose tracking and 3D reconstruction. For the DROID-SLAM algorithm, working in object-centered scenes reduces the availability of textures and geometric cues for tracking in the images. This environment significantly diminishes the reliability of the optical flow and depth priors on which the algorithm depends, resulting in an overall decline in performance. Both the Dust3R and GS-SLAM are enhanced with generalized stereo matching to adopt stronger 3D geometric priors. However, these algorithms build a globally optimized model that does not adequately handle noise and errors in the depth priors. Consequently, as the inference progresses over multiple frames, errors quickly accumulate within the global model, causing continual degradation in performance and eventually leading to tracking failure.

Fig. 3 presents qualitative comparisons with other methods. Despite various challenges (*e.g.*, severe hand occlusions, self-occlusions, frames lacking texture and geometric cues, and strong light reflections), our algorithm successfully tracks the object's 6DoF pose and achieves a significantly high-quality 3D appearance representation. Notably, the appearance of the reconstructed 3D object in our approach better aligns with the texture and color information of the source object in the scene, compared to the ground truth reconstructed from scans.

4.4. Results on the OnePose Dataset

The quantitative results on the OnePose dataset are shown in Tab. 2. This dataset consists of object-centered static scenes, where we use a video segmentation network to isolate the object. Our algorithm exhibits superior global tracking compared to previous SLAM algorithms, especially

Table 2. Quantitative comparison on the OnePose dataset.

Method	APE (cm)↓	RPE (cm)↓	PSNR↑	SSIM↑
Dust3R [34]	30.42	25.29	—	—
Gaussian Splatting SLAM [23]	10.28	9.62	19.27	0.85
DROID-SLAM [32]	8.57	6.94	—	—
SfM [41]+3DGS [16]	—	—	21.43	0.87
GSGTrack	**7.36**	**8.79**	**23.22**	**0.90**

Table 3. Ablation study of different settings of our methods.

Method	ADD-S% ≤0.3m↑	ADD% ≤0.3m↑	PSNR↑	SSIM↑
w/o Tracking	32.22	23.14	14.54	0.88
w/o Silhouette Loss	56.31	42.16	24.21	**0.97**
w/o Geometric Graph	25.20	15.93	22.77	0.95
w/o Image Pruning	50.99	39.44	23.59	**0.97**
w/o Geometric Optimization	51.08	32.96	24.30	0.96
Ours	**62.51**	**51.83**	**27.04**	**0.97**

when background information is excluded. However, tracking stability at finer details fluctuates, yielding weaker RPE metrics compared to DROID-SLAM. We also compare with the SfM+3DGS method, which reconstructs poses from full-scene, unsegmented views using 3DGS. Due to scene complexity, poses derived from SfM frequently show deviations, causing misalignment between reconstructed and ground-truth views during object reconstruction and reducing overall quality. This further underscores GSGTrack's advantages in object-centered reconstruction.

4.5. Ablation Study

We conduct extensive ablation studies on the HO3D dataset to validate the effectiveness of our proposed strategies, with results presented in Tab. 3 and Fig. 4.
Pose Tracking. As mentioned previously, we track poses across frames to initialize the object pose. To assess the impact of this strategy, we remove pose tracking during initialization (*w/o* Tracking) and instead use a PnP algorithm to estimate poses. The results demonstrate that tracking quickly drifts due to inaccurate initialization.
Silhouette Loss. We introduce a differentiable silhouette loss to mitigate errors caused by the simple photometric loss. To evaluate its effectiveness, we exclude this component from the experiments (*w/o* Silhouette Loss). The results show a clear performance decline, with ADD-S@0.3d and ADD@0.3d decreasing by 10% and 19% respectively.
Graph-based Geometric Optimization. To validate the importance of graph-based optimization, we remove it and rely only on the latest frame for updates (*w/o* Geometric Graph), resulting in error accumulation over time. We also examine the impact of removing the image pair pruning strategy (*w/o* Image Pruning), which leads to degraded performance due to failed edges. Finally, omitting gradient de-
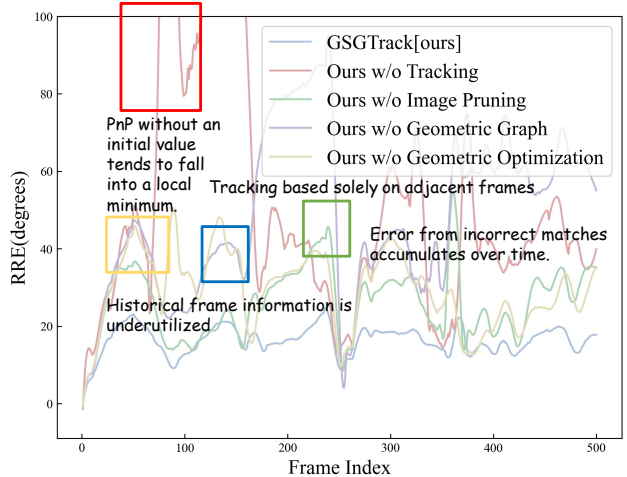

Figure 4. We visualize the Relative Rotation Error (RRE) of different settings of our method.

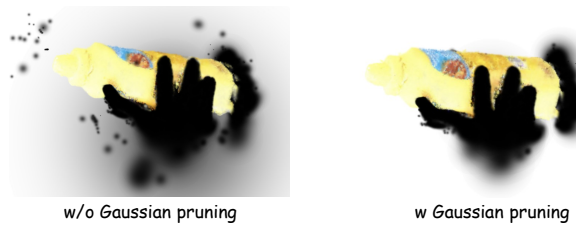w/o Gaussian pruning w Gaussian pruning

Figure 5. Impact of our Gaussian pruning strategy on reconstruction quality. Our strategy significantly enhances geometric accuracy and effectively eliminates floaters.

scent optimization on the geometric graph shows that geometric optimization is essential for improving performance.
Gaussian Pruning. To address geometric inaccuracies, we propose a Gaussian pruning strategy. To validate its effect, we remove it from the pipeline and provide qualitative comparisons in Fig. 5. The results reveal that our strategy improves geometric accuracy and effectively prunes floaters.

5. Conclusion

This paper addresses a challenging problem: 6D pose tracking of unknown objects from RGB videos without accurate depth information. To address this, we introduce GSGTrack, a novel method that leverages Gaussian splatting to enhance pose tracking. Our approach employs a joint optimization framework to simultaneously refine object poses and their 3D representation. To manage continuously incoming video frames during tracking, we develop an online 3DGS representation, enabling incremental object reconstruction. Furthermore, we propose a graph-based geometric optimization framework that integrates an image pair pruning strategy and a confidence-aware optimization strategy to improve accuracy and robustness. Extensive experiments across multiple datasets show that our framework achieves robust pose tracking and accurate 3D reconstruction, even with inaccurate initial geometric information.

References

[1] Berta Bescos, José M. Fácil, Javier Civera, and José Neira. DynaSLAM: Tracking, mapping, and inpainting in dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1–9. IEEE, 2018. 2

[2] Berta Bescos, José M. Fácil, Javier Civera, and José Neira. Dynaslam2: Real-time dense monocular slam with dynamic object removal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–9, 2020. 2

[3] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Gs-pose: Generalizable segmentation-based 6d object pose estimation with 3d gaussian splatting. *arXiv preprint arXiv:2403.10683*, 2024. 2

[4] Ming Cai and Ian Reid. Reconstruct locally, localize globally: A model free method for object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3153–3163, 2020. 2

[5] Christopher B Choy, Yinda Xu, Junhyuk Gwak, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3352–3361, 2016. 2

[6] Xinhan Di and Pengqian Yu. 3d reconstruction of simple objects from a single view silhouette image. *arXiv preprint arXiv:1701.04752*, 2017. 5

[7] Will Eastcott. Supersplat: 3d gaussian splat editor, 2024. Accessed: 2024-11-19. 1

[8] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Sparse-view sfm-free gaussian splatting in seconds, 2024. 2, 4

[9] Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. Colmap-free 3d gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20796–20805, 2024. 2

[10] Ashkan Ganj, Yiqin Zhao, Hang Su, and Tian Guo. Mobile ar depth estimation: Challenges & prospects – extended version. *arXiv preprint arXiv:2310.14437*, 2023. 1

[11] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3196–3206, 2020. 2, 6

[12] Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, and Qifeng Chen. Fs6d: Few-shot 6d pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6814–6824, 2022. 7

[13] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian Conference on Computer Vision (ACCV)*, pages 548–562, 2012. 1, 2

[14] Daniel Kappler, Franziska Meier, Jan Issac, Jim Mainprice, Cristina Garcia Cifuentes, Manuel Wüthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg. Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters*, 3(3):1864–1871, 2018. 1

[15] Wadim Kehl, Fabian Manhardt, Federico Tombari, Nassir Navab, and Slobodan Ilic. Sd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1521–1529, 2017. 2

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2, 3, 7, 8

[17] Georg Klein and David Murray. PTAM: Real-time tracking and mapping for augmented reality. In *IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 225–234. IEEE, 2007. 2

[18] Yann Labbé, Mathieu Caron, Mathieu Aubry, Josef Sivic, and Ivan Laptev. CosyPose: Consistent multi-view multi-object 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 574–591, 2020. 2

[19] Fu Li, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic. Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. *arXiv preprint arXiv:2203.04802*, 2022. 2

[20] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6548–6557, 2021. 2

[21] Jingwen Lin, Ziang Wang, Xinyu Yu, Siyu Zhu, Hujun Bao, Xiaowei Zhou, and Guofeng Wang. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6827–6836, 2022. 2, 6

[22] Luqing Luo, Shichu Sun, Jiangang Yang, Linfang Zheng, Jinwei Du, and Jian Liu. Object gaussian for monocular 6d pose estimation from sparse views. *arXiv preprint arXiv:2409.02581*, 2024. 2

[23] Hidenobu Matsuki, Riku Murai, Paul H. J. Kelly, and Andrew J. Davison. Gaussian splatting slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1234–1243, 2024. 4, 6, 7, 8, 2

[24] 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 *European Conference on Computer Vision (ECCV)*, pages 676–691, 2020. 2

[25] Raul Mur-Artal and J. D. Tardós. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. *IEEE Transactions on Robotics*, 33(5):1255–1262, 2017. 2

[26] Lachlan Nicholson, Michael Milford, and Niko Sünderhauf. Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics and Automation Letters*, 3(4):3540–3547, 2018. 2

[27] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 7668–7677, 2019. 2

[28] Sida Peng, Yuan Liu, Qixing Huang, Xibin Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4561–4570, 2019. 1

[29] Giorgia Pitteri, Slobodan Ilic, and Vincent Lepetit. Cornet: Generic 3d corners for 6d pose estimation of new objects without retraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pages 0–0, 2019. 1

[30] Jean-Pierre Sleiman, Farbod Farshidian, Maria Vittoria Minniti, and Marco Hutter. A unified mpc framework for whole-body dynamic locomotion and manipulation. *IEEE Robotics and Automation Letters*, 6(2):4688–4695, 2021. 1

[31] Jürgen Sturm, Nils Engelhard, Frank Endres, Wolfram Burgard, and Daniel Cremers. Evaluation for odometry. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4206–4212, 2012. 7

[32] Zachary Teed and Jis Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8348–8357, 2021. 6, 7, 8

[33] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2642–2651, 2019. 1, 2

[34] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20697–20709, 2024. 2, 4, 5, 6, 7, 8

[35] Xiu Li Wang, Zhi Zhang, Zhen Zhang, Yi Yang, Yichen Yu, and Lei Zhang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision*, pages 54–70, 2018. 2

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[37] Philippe Weinzaepfel, Vincent Leroy, Thomas Lucas, Romain Brégier, Yohann Cabon, Vaibhav Arora, Leonid Antsfeld, Boris Chidlovskii, Gabriela Csurka, and Jérôme Revaud. Splat3r: Zero-shot gaussian splatting from uncalibrated image pairs, 2024. 2

[38] Bowen Wen and Kostas E. Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021. 1, 2

[39] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6401–6408. IEEE, 2022. 1

[40] Bowen Wen, Jonathan Tremblay, Valts Blukis, Stephen Tyree, Thomas Müller, Alex Evans, Dieter Fox, Jan Kautz, and Stan Birchfield. Bundlesdf: Neural 6-dof tracking and 3d reconstruction of unknown objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 606–617, 2023. 1, 2, 7

[41] Cheng Wu. Visualsfm: A visual structure from motion system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2011. 7, 8

[42] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Proceedings of Robotics: Science and Systems (RSS)*, 2018. 1, 7

[43] Haolin Xiong, Sairisheek Muttukuru, Rishi Upadhyay, Pradyumna Chari, and Achuta Kadambi. Sparsegs: Real-time 360° sparse view synthesis using gaussian splatting. *arXiv preprint arXiv:2312.00206*, 2023. 4

[44] Linghao Yang, Yanmin Wu, Yu Deng, Rui Tian, Xinggang Hu, and Tiefeng Ma. Uniquadric: A slam backend for unknown rigid object 3d tracking and light-weight modeling. *arXiv preprint arXiv:2309.17036*, 2023. 1

[45] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1

[46] Shichao Yang and Sebastian Scherer. Cubeslam: Monocular 3-d object slam. *IEEE Transactions on Robotics*, 35(4):925–938, 2019. 2

[47] Yao Yao, Zhi Li, Yifan Yang, Ming-Hsuan Yang, and Zhen Xu. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2829, 2018. 2

[48] Botao Ye, Sifei Liu, Haofei Xu, Xueting Li, Marc Pollefeys, Ming-Hsuan Yang, and Songyou Peng. Noposplat: No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images. In *arXiv preprint arXiv:2410.24207*, 2024. 2

[49] Hao Zhang and Shuaijie Zhang. Shape-iou: More accurate metric considering bounding box shape and scale. *arXiv preprint arXiv:2312.17663*, 2023. 5

[50] Zhi Zhang, Sifei Liu, Yi Yang, Zhen Zhang, Yichen Yu, and Lei Zhang. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15510–15519, 2021. 2

[51] Zhi Zhang, Sifei Liu, Yi Yang, Zhen Zhang, Yichen Yu, and Lei Zhang. Drivefeedforward: Unleashing generalization of end-to-end autonomous driving with controllable long video generation. In *arXiv preprint arXiv:2406.01349*, 2023. 2

[52] Y. Zhao et al. Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9378–9388, 2023. 1

GSGTrack: Gaussian Splatting-Guided Object Pose Tracking from RGB Videos

Supplementary Material

In this supplementary material, we provide the implementation details of the experiments, along with the algorithm evaluation metrics and comprehensive information about the datasets. Furthermore, we provide qualitative results for challenging scenarios in the datasets, analyze the limitations of the proposed algorithm, and discuss its broader impact.

6. Implementation details

6.1. Data Preprocessing

During the data preprocessing stage, for the segmented video image F_t, we first enlarge the image to reduce the relative proportion of background noise. Subsequently, using the segmented mask as a reference, we crop the image so that the projection center of the object in the 2D image aligns as closely as possible with the center of the image, thereby reducing the difficulty of geometric estimation. During the image processing, we simultaneously adjust the camera's intrinsic parameters to maintain the validity of the solved camera extrinsic parameters. Specifically, the transformation matrix for the camera's intrinsic parameters M is as follows:

$$M = \begin{bmatrix} K & 0 & -Kx_0 + \frac{w'}{2} \\ 0 & K & -Ky_0 + \frac{h'}{2} \\ 0 & 0 & 1 \end{bmatrix}, \qquad (9)$$

where (h', w') represents the dimensions of the target image, K is the scaling factor, and (x_0, y_0) denotes the center of the object in the 2D image.

6.2. System Details and Hyperparameters

During the pose initialization process (Sec 3.4 in the main manuscript), if the pose of the previous frame is not available as a direct reference(*e.g.* missing detection by the segmentation or object reappearing after complete occlusion), the algorithm first performs generalized stereo matching between the current frame and historical keyframes in the pose graph to estimate the pose of the current frame. This estimated pose is then used as the initial pose for the tracking process. If the tracking state of the previous frame is valid, the initial pose of the current frame's tracking process is directly set to the pose of the previous frame. In experiments, 100 iterations of tracking are executed, and if the magnitude of the pose update falls below 10^{-4}, the iterations are terminated early.

During the geometric graph optimization process, we limit the optimization to two layers: optimizing the pose and depth map of the current frame simultaneously, as well as the depth maps of the reference frames. Pixels with a confidence score lower than 2 are excluded from the computation of L_{pg}. The geometric graph is optimized for approximately 300 iterations in total.

During the training process of the online 3D Gaussian Splatting algorithm, considering that the geometry optimization algorithm has already provided good initial values for the 3D Gaussian points, the learning rate for the Gaussian point positions is reduced to 0.000032. In object-level scenes, the Gaussian point size threshold is set to 3. For the initial frame view of each object, the algorithm trains for 325 iterations to initialize the 3D Gaussian Splatting algorithm. Subsequently, Gaussian optimization is conducted for 125 iterations per video frame, incrementally reconstructing the 3D Gaussian representation of the object online while optimizing the object's 6-DoF pose. Every 25 iterations, the algorithm executes a density control strategy consistent with the classical 3D Gaussian Splatting algorithm. During the final iteration of Gaussian optimization, the algorithm applies a Gaussian pruning strategy, as described in (Sec 3.3), to remove geometrically inaccurate 3D Gaussian points.

6.3. Keyframing

To ensure the efficiency of algorithm execution, it is impractical to perform the same optimization process for every frame. Therefore, in our implementation, we calculate the rotational geodesic distance of each frame relative to the nodes in the geometric pose graph. This approach ensures that the keyframes added to the geometric graph provide novel information, including texture details, viewpoint diversity, and scale variations of the object. Non-keyframes, on the other hand, are only initialized with a pose estimate (Sec 3.4) and do not participate in subsequent geometric graph optimization or Gaussian model refinement processes.

6.4. Visualization

For the reconstructed 3D Gaussian Splatting model, we visualize it using supersplat [7]. For the 3D pointmaps, we select points with confidence greater than 2, merge them in the world coordinate system to form a unified object point cloud, and apply voxel-based uniform downsampling to obtain a consistent point cloud representation of the object. The mesh model used for visualization is generated from the object point cloud by reconstructing the object surface using the Poisson reconstruction algorithm.

6.5. Gradient Derivation

For efficiency, 3DGS [16] employs CUDA-based rasterization, requiring explicit computation of parameter derivatives. Consequently, the chain rule is applied to differentiate Eq.(19), yielding partial derivatives as follows:

$$\frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{T}} = \frac{\partial \boldsymbol{\mu}'}{\partial \boldsymbol{\mu}} \frac{\mathcal{D}\boldsymbol{\mu}}{\mathcal{D}\boldsymbol{T}}, \tag{10}$$

$$\frac{\partial \boldsymbol{\Sigma}'}{\partial \boldsymbol{T}} = \frac{\partial \boldsymbol{\Sigma}'}{\partial \mathbf{J}} \frac{\partial \mathbf{J}}{\partial \boldsymbol{\mu}} \frac{\mathcal{D}\boldsymbol{\mu}}{\mathcal{D}\boldsymbol{T}} + \frac{\partial \boldsymbol{\Sigma}'}{\partial \mathbf{W}} \frac{\mathcal{D}\mathbf{W}}{\mathcal{D}\boldsymbol{T}}. \tag{11}$$

Following gaussian splatting SLAM [23], we derived the minimal Jacobian matrix on the manifold using Lie algebra and explicitly computed the derivatives of the camera pose.

$$\frac{\mathcal{D}\boldsymbol{\mu}}{\mathcal{D}\boldsymbol{T}} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{\mu}^{\times} \end{bmatrix}, \frac{\mathcal{D}\mathbf{W}}{\mathcal{D}\boldsymbol{T}} = \begin{bmatrix} \mathbf{0} & -\mathbf{W}_{i,1}^{\times} \\ \mathbf{0} & -\mathbf{W}_{i,2}^{\times} \\ \mathbf{0} & -\mathbf{W}_{i,3}^{\times} \end{bmatrix}. \tag{12}$$

7. Metrics

To evaluate the results of the algorithm, we assess both 6DoF pose tracking and object reconstruction. For the 6DoF pose tracking results, we compute the Area Under the Curve (AUC) percentages for the ADD and ADD-S metrics,

$$\text{ADD} = \frac{1}{|\mathcal{M}|} \sum_{x \in \mathcal{M}} \|(Rx + t) - (\tilde{R}x + \tilde{t})\|_2, \tag{13}$$

$$\text{ADD-S} = \frac{1}{|\mathcal{M}|} \sum_{x_1 \in \mathcal{M}} \min_{x_2 \in \mathcal{M}} \left\|(Rx_1 + t) - \left(\tilde{R}x_2 + \tilde{t}\right)\right\|_2, \tag{14}$$

where $\mathcal{M}$ is the object model. Since the CAD model of the novel, unknown object is unavailable for defining a coordinate system, we utilize the ground-truth pose from the first frame to establish the canonical coordinate frame for each video, enabling pose evaluation.

For 3D reconstruction evaluation, the object's 3D model is projected onto 2D images and rendered from corresponding viewpoints. The projections are then compared with ground-truth images using PSNR and SSIM metrics. To evaluate the accuracy of 3D reconstruction shapes, we followed BundleSDF [40]. Specifically, we assessed the Chamfer Distance between the final geometrically optimized point cloud and the downsampled point cloud from the ground-truth mesh. The symmetric formula used is as follows:

$$d_{CD} = \frac{1}{2|\mathcal{M}_1|} \sum_{x_1 \in \mathcal{M}_1} \min_{x_2 \in \mathcal{M}_2} \|x_1 - x_2\|_2 + \frac{1}{2|\mathcal{M}_2|} \sum_{x_2 \in \mathcal{M}_2} \min_{x_1 \in \mathcal{M}_1} \|x_1 - x_2\|_2. \tag{15}$$

In the shape evaluation process, we downsampled the point cloud to a uniform resolution of 5 mm.

Table 4. **Scene sequences of HO3D and OnePose**.

HO3D		OnePose	
Pitcher Base	AP11	0500-Chocfranzzi-Box	Choc-01
	AP14	0501-Matchafranzzi-Box	Mat-01
Potted Meat Can	MPM14	0518-Jasmine-Box	Jas-01
Bleach Cleanser	SB11	0535-Odbmilk-Box	Odb-01
	SB13	0543-Brownhouses-Others	Brown-01
Mustard Bottle	SM1		

Pitcher Base   Potted Meat Can   Bleach Cleanser   Mustard Bottle

Figure 6. Visualization for the objects of in HO3D dataset

Chocfranzzi Box   Matchafranzzi Box   Jasmine Box

Odbmilk Box   Brownhouse

Figure 7. Visualization for the objects of in OnePose dataset

8. Datasets

As shown in the Fig. 6, we selected 6 representative video sequences from the HO3D dataset, which include 4 dynamic objects. Each scene contains approximately 1,000 frames of data, featuring dynamic objects and hands interacting with them. The scale information for the first frame was calculated using the ground-truth depth values provided by the dataset. Based on this, we conducted experiments on the dataset. Fig. 9 presents the ADD-S and ADD recall curves obtained from these experiments, while Fig. 8 and Fig. 10 show the qualitative results for this dataset. It can be observed that our method outperforms all other methods in both qualitative and quantitative metrics on the HO3D dataset. For the OnePose dataset, we selected 5 video sequences as illustrated in Fig. 7, containing five static ob-

jects. Each scene comprises approximately 500 frames, while Fig. 11 show the qualitative results for this dataset. . The significant variations across scenes serve as an accurate indicator of the reconstruction capabilities of current methods. The indexing of the scene sequences is provided in Tab. 4.

9. Limitation

Although our method demonstrates greater robustness than the baseline algorithm in handling low-textured objects and occlusions (as shown in Fig. 8 and Fig. 10), it performs poorly when dealing with uniformly colored objects that lack geometric, color, or texture features. The method relies on the first frame of the video to initialize the object's local coordinate system, making it sensitive to the quality of initial matching. Such matching can be compromised by segmentation errors, lighting issues, or insufficient texture in the initial viewpoint. For instance, in the AP10 sequence of the HO3D dataset, the absence of geometric cues in the first frame significantly degrades performance. Additionally, the method assumes that each 2D image point corresponds to a 3D world point, limiting its applicability to transparent objects.

10. Broader Impact

The GSGTrack framework introduces a significant leap forward in the field of 6-DoF pose tracking and 3D object reconstruction, particularly for applications relying solely on monocular RGB video data. By eliminating the reliance on accurate depth information, the proposed approach offers broader accessibility and applicability in scenarios such as robotic manipulation, augmented reality, and autonomous systems, where lightweight and cost-effective sensing solutions are required. The novel 3D Gaussian Splatting representation and integrated graph-based geometric optimization framework enable robust pose tracking and high-fidelity object reconstruction, advancing theory and offering practical tools for interdisciplinary applications.
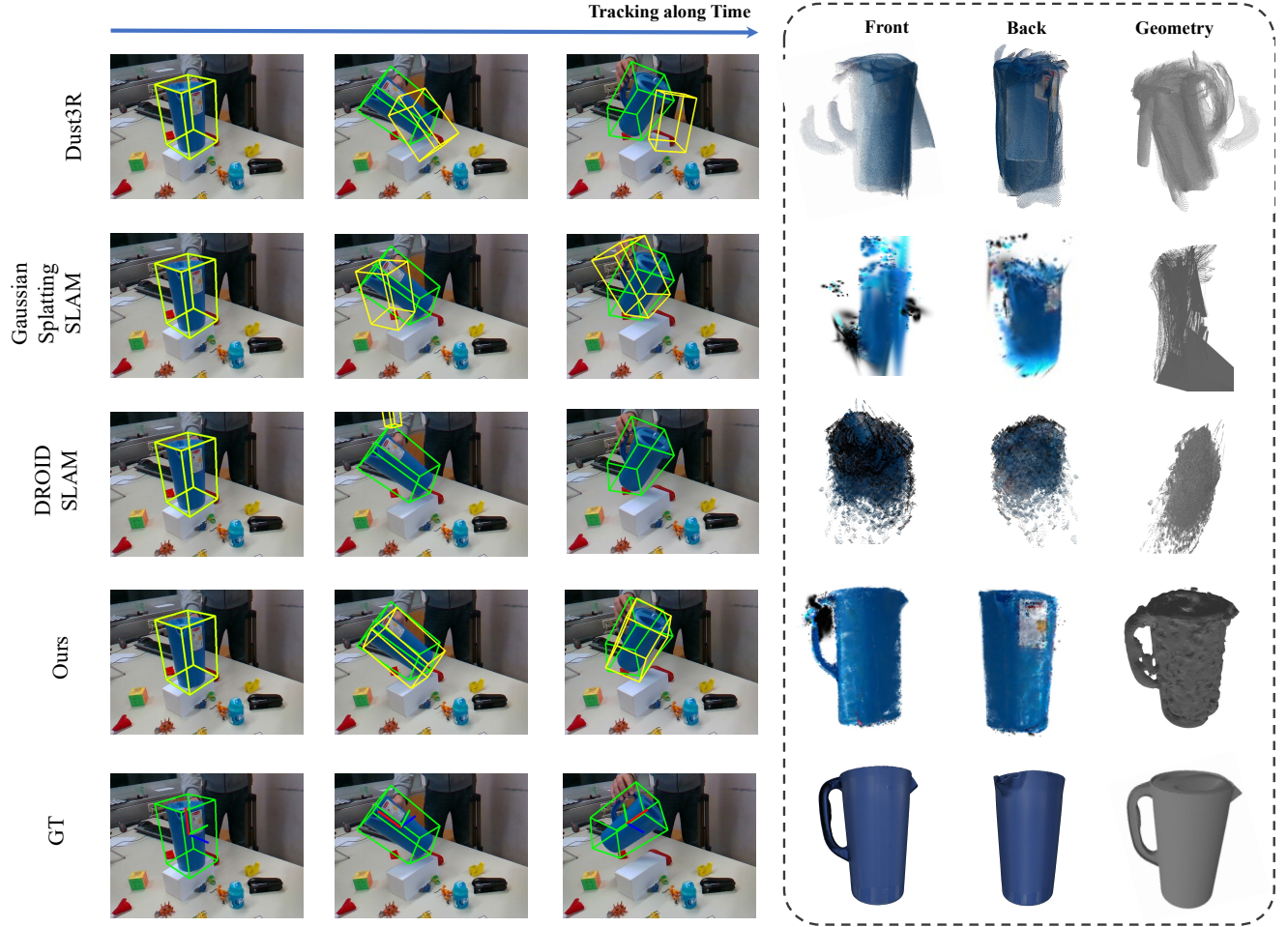
Figure 8. **Qualitative Comparison of GSGTrack and Baseline on HO3D(Seq-AP11).** Left: 6-DOF pose tracking with green and yellow boxes showing ground truth and estimated poses, respectively. Right: front and back views of reconstruction results, highlighting the object's geometric structure. A blue pitcher with low texture is presented. The qualitative results demonstrate that, compared to the baseline algorithm, our method exhibits significantly enhanced robustness in handling low-texture objects.

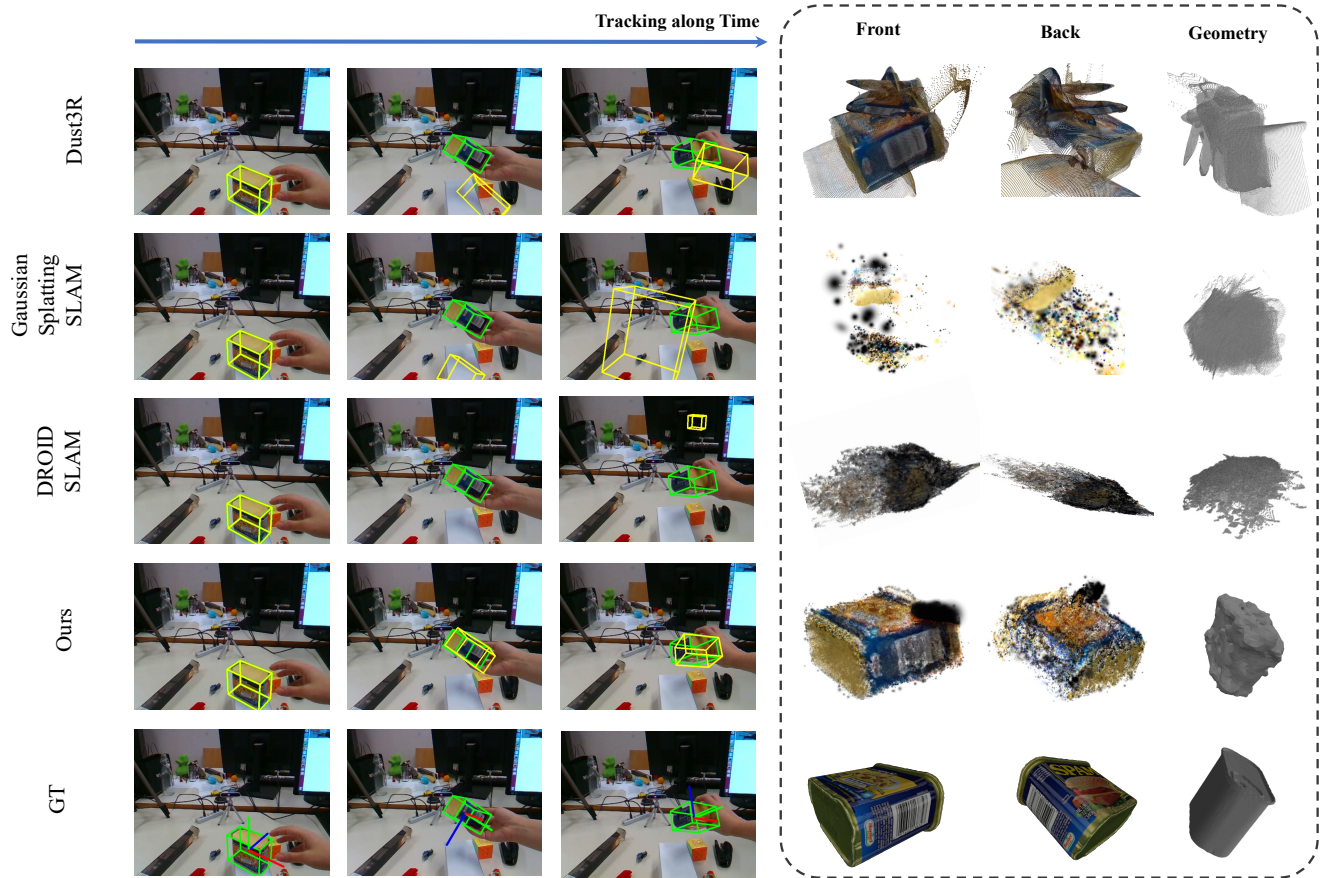Figure 9. Recall curve of ADD-S (left) and ADD (right) metric on HO3D Dataset..

Figure 10. **Qualitative Comparison of GSGTrack and Baseline on HO3D(Seq-MPM14).** Left: 6-DOF pose tracking with green and yellow boxes showing ground truth and estimated poses, respectively. Right: front and back views of reconstruction results, highlighting the object's geometric structure. A Potted Meat Can object partially occluded by a hand is presented. The qualitative results demonstrate that, compared to the baseline algorithm, our approach exhibits significantly enhanced robustness in handling scenarios with occlusion.
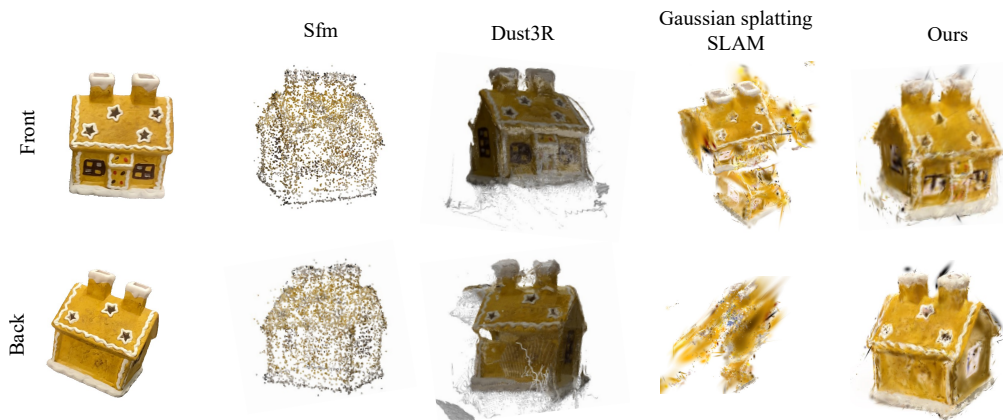

Figure 11. **Qualitative Comparison of GSGTrack and Baseline on OnePose.** Our method demonstrates superior object reconstruction quality.