2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

MGS-Track: Monocular 6DoF Pose Tracking via Masked 3D Prior and Online Gaussian Splatting

Zhiyuan Chen[1,2], Fan Lu[1†], Guo Yu[1], Sanqing Qu[1], Ya Wu[3], Yuan Huang[4], Alois Knoll[5], Guang Chen[1,2†]

Fig. 1: Given a monocular RGB sequence and object mask, **MGS-Track** utilizes a feedforward network to obtain geometric priors, enabling 6-DoF pose tracking and 3D reconstruction of the target. Results are directly visualized by our method.

***Abstract*— Tracking the 6DoF pose of previously unseen objects from monocular RGB videos is crucial for robotic manipulation, yet remains challenging due to depth ambiguity and limited object-centric visual context. Existing trackers often rely on accurate depth sensors, which constrains deployment in low-cost settings, while substituting monocular pseudo-depth frequently introduces geometric errors that reduce tracking robustness. To this end, We propose MGS-Track, an object-centric online tracking and reconstruction framework that combines learning-based geometric priors with differentiable 3D Gaussian Splatting (3DGS). Specifically, we first introduce a mask-augmented DUSt3R network (DUSt3R-M) to establish pairwise correspondences and predict point maps, which serve as geometric priors for initializing and guiding an online 3DGS representation. We then jointly optimize Gaussian parameters and 6DoF object poses in a coarse-to-fine manner, enabling robust tracking and high-fidelity reconstruction. To control model growth and maintain efficiency over time, we further introduce adaptive Gaussian management with capacity-aware selection and mask-consistent pruning. Experiments on YCBInEOAT and HO3D show that MGS-Track consistently outperforms competitive monocular baselines on both pose tracking and object reconstruction in challenging object-centric scenarios.**

I. INTRODUCTION

6DoF object pose tracking aims to continuously estimate the precise 3D position and orientation of target objects from consecutive video sequences captured in dynamic scenes. This provides consistent and accurate positional information for objects being manipulated, which is essential for applications such as robotic manipulation and control [1, 2].

Early 6DoF object pose estimation approaches typically require pre-defined 3D object models [3, 4] or category templates [5, 6], and estimate poses through geometric feature matching pipelines [7]. This reliance on known models or templates limits their ability to generalize to novel objects. To track unknown objects, recent works adapt online localization strategies from SLAM frameworks [8, 9]. By processing RGBD video sequences with synchronized RGB–depth fusion and cross-modal geometric verification, these methods can achieve 6DoF tracking, but they inherently depend on accurate depth to maintain structural constraints. In particular, many of these algorithms rely on SDF-based online reconstruction, which is sensitive to depth inaccuracies and therefore further amplifies the dependence on depth quality. As a result, deploying such methods on lightweight robotic platforms that rely on monocular vision remains challenging, where accurate depth perception is difficult to obtain [10].

Monocular models [11, 12] provide an alternative to depth sensors, but they remain insufficient for reliable 6DoF tracking. First, monocular predictions often suffer from scale ambiguity and limited multi-scale aggregation over time, which can introduce geometric errors and degrade tracking accuracy. To alleviate this issue, DUSt3R [13] and VGGT [14] adopt multi-frame joint optimization to better aggregate information across views. However, these methods reconstruct dynamic foreground objects mainly through background masking and do not explicitly estimate object poses; this setting discards useful context and can further hurt performance. As a result, these approaches are still brittle in

[1]School of Computer Science and Technology, Tongji University, Shanghai, China;[2] Shanghai Innovation Institute, Shanghai, China;[3]China National Nuclear Corporation No. 8 Institute, China; [4] Beijing Institute of Control Engineering, Beijing, China; [5]Technical University of Munich (TUM), Munich, Germany.
[†]Corresponding authors: Fan Lu (lufan@tongji.edu.cn) and Guang Chen (guangchen@tongji.edu.cn).

dynamic scenes.

To this end, we propose MGS-Track, an online framework for joint 6DoF pose tracking and 3D reconstruction enhanced by geometric priors. Building upon DUSt3R, we first introduce an end-to-end reconstruction module that focuses on the foreground object. In this module, uncalibrated image pairs are back-projected into 3D point maps and used as geometric priors. We then adopt 3DGS as the core representation and leverage differentiable rendering to jointly optimize the object pose and the 3D representation, enabling stable tracking even when the priors are imperfect. Finally, to improve efficiency, we introduce an adaptive Gaussian management strategy that controls the number of Gaussian primitives and accelerates optimization.

To evaluate the proposed method, extensive experiments are conducted on two monocular RGB object pose tracking datasets, *i.e.*, YCBInEOAT dataset [15] and HO3D dataset [16]. 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. (1) We propose a feedforward network for reconstructing foreground objects from monocular video streams. (2) We introduce an online 3DGS reconstruction approach that leverages geometric priors from monocular videos and uses differentiable rendering to jointly optimize the Gaussian object field and the object pose, enabling online pose updates. (3) We propose an adaptive Gaussian pruning method based on voxel partitioning to control the number of Gaussians and improve computational efficiency.

II. RELATED WORK

6DoF Object Pose Estimation and Tracking. Early approaches for 6DoF pose estimation rely on CAD models to establish precise geometric correspondences. Methods like [17–19] leverage offline training with object-specific CAD data but fail to generalize to unseen instances. Category-level templates [5, 6] alleviate this limitation by sharing shape priors within object categories, yet remain constrained by template accuracy. Estimating poses without CAD models is inherently ill-posed due to depth ambiguities in monocular RGB images. Recent works mitigate this by integrating 3D shape reconstruction into pose estimation: [20] combines Mask-RCNN with differentiable rendering to recover object geometry, while neural implicit representations like NeRF [21] and explicit 3D Gaussians [22, 23] enable multi-view consistent feature synthesis. These methods bypass CAD dependency but typically require pre-captured reference views, limiting applicability in dynamic scenarios. For 6DoF pose tracking, temporal information is leveraged to estimate object poses across video frames. Some studies construct 3D models from multi-view video frames to extend tracking to unknown objects [9, 24]. BundleSDF [9] shares our goal of joint pose tracking and reconstruction for unseen objects, but requires RGB-D data for implicit SDF optimization. Our 3DGS framework [25] uses differentiable Gaussian rendering to co-optimize pose and texture from RGB inputs. Unlike geometry-centric approaches, we incorporate priors from [13] to enable robust 6DoF estimation and high-fidelity surfaces in depth-deficient scenarios.

Simultaneous Localization and Mapping Algorithms. RGB-SLAM algorithms estimate camera pose and reconstruct scenes from monocular RGB sequences [26], posing challenges similar to ours. Although dynamic SLAM variants [27, 28] handle moving scenes, they typically mask dynamic objects, which limits reconstruction of those objects. Object-aware SLAM approaches [29, 30] incorporate semantic detection and quadric-based geometry, but still cannot model object–environment interactions or recover complete object geometry. Our method advances this line by using 3D Gaussian Splatting to continuously fuse new RGB observations into a consistent object-centric representation. Unlike prior work that separates pose estimation from reconstruction, our framework jointly optimizes 6DoF object poses and photorealistic geometry, enabling simultaneous tracking and reconstruction in interactive dynamic scenarios.

3D Reconstruction. Learning-based methods have long studied 3D reconstruction from 2D images [31–33]. Neural scene representations such as NeRF [34] and 3DGS [25] can reconstruct high-quality geometry and appearance from images, but they typically assume known camera poses. Pose-free methods [35–37] relax this requirement yet are mostly designed for static scenes and often break in dynamic-foreground settings. BundleSDF [9] reconstructs objects with SDFs using depth acquired from sensors, yet it has limited ability to model fine-grained appearance; in contrast, our RGB-supervised 3DGS preserves appearance details more faithfully. Although feed-forward networks can directly predict 3D structure, their limited robustness to domain shift, occlusions, and appearance variation makes them substantially less reliable than optimization-based methods in in-the-wild settings. Our method unifies both paradigms by using a feed-forward network to provide geometric priors and performing online optimization over a 3DGS representation.

III. METHOD

A. Preliminary

Problem Formulation. Given a monocular RGB video sequence $F = \{F_t\}_{t=0}^{n-1}$ ($F_t \in \mathbb{R}^{W\times H\times 3}$) depicting a dynamic, object-centric scene and the first-frame segmentation mask M_0, MGS-Track aims to perform online 6DoF pose tracking while reconstructing a photorealistic textured 3D model of the object.

Preliminary for 3DGS [25]. As mentioned before, we use 3D Gaussian Splatting (3DGS) as our basic object representation. 3DGS provides differentiable real-time rendering through explicit anisotropic 3D Gaussians, making it suitable as a basic 3D representation for object pose tracking. Specifically, 3DGS represents a scene as a set of anisotropic 3D Gaussian spheres. 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 projects all 3D Gaussian spheres into 2D Gaussian distributions through a differentiable Gaussian Splatting pipeline, and then blends

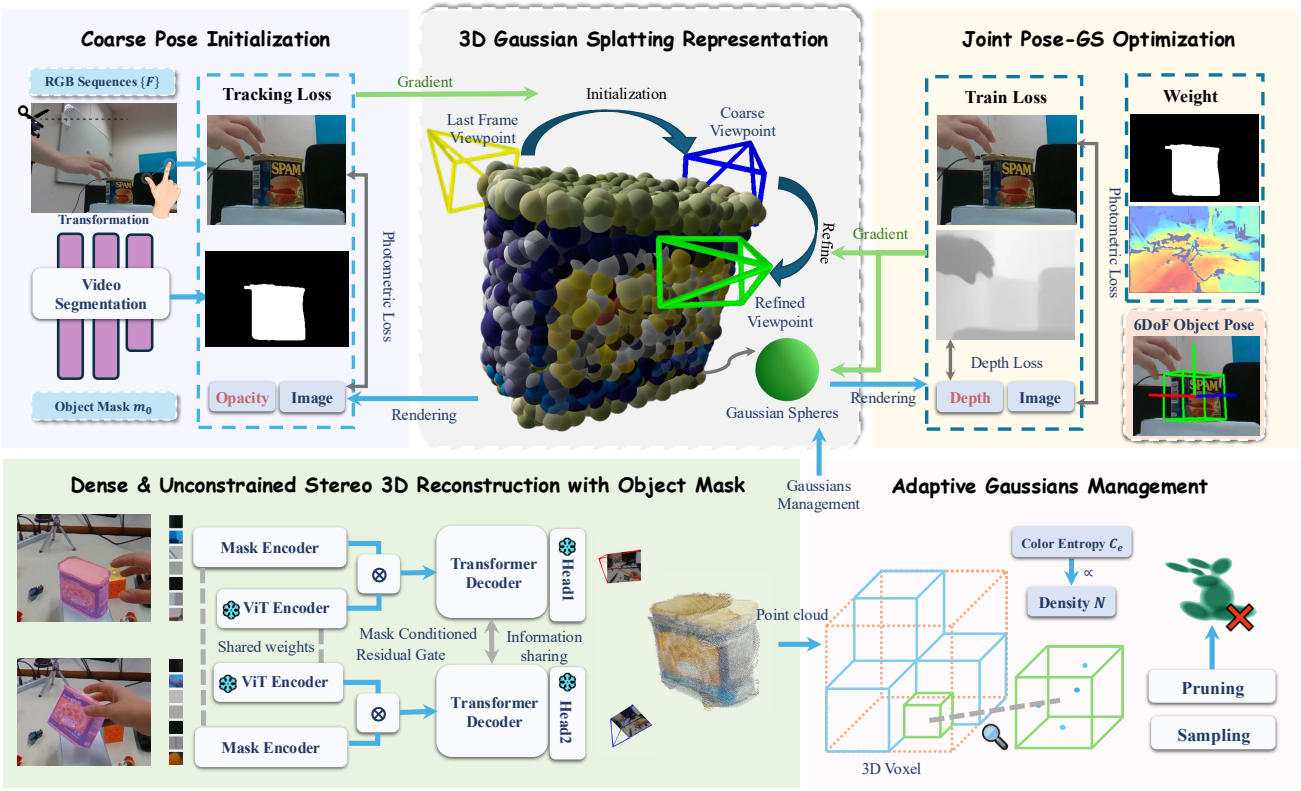

Fig. 2: **Overview of our proposed MGS-Track.** Given a monocular video, we segment the object, initialize a coarse pose, and use DUSt3R-M to generate object-centric pointmaps as geometric priors. We then perform joint pose and 3DGS optimization with photometric alignment and confidence-weighted depth consistency, which stabilizes tracking. An adaptive Gaussian management module with entropy-guided voxel budgets and mask-consistency pruning controls point growth and keeps the representation compact. The full pipeline runs online for robust 6DoF tracking and high-fidelity reconstruction.

the colors using fast alpha blending. The rendering process can be summarized as follows:

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

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

where π is the projection operation, $\boldsymbol{T}$ is the camera pose, $\boldsymbol{W}$ is the rotational part of $\boldsymbol{T}$, and $\boldsymbol{J}$ is the Jacobian of π.

B. MGS-Track Overview

We propose *MGS-Track*, an online framework for monocular 6DoF pose tracking and 3D reconstruction. As shown in Figure 2, MGS-Track combines object-centric feed-forward geometry with rendering-based optimization in 3DGS. For each incoming frame, DUSt3R-M (Section III-C) produces dense, object-aligned pointmaps that serve as geometric priors. These priors initialize and update an *online 3DGS representation* of the object (Section III-D). We then perform coarse-to-fine optimization that jointly refines the 3DGS representation and the object's 6DoF pose. To keep the representation compact, we introduce an adaptive Gaussian management strategy that prunes redundant Gaussians from the learned object field (Section III-E). Together, these components form a streaming loop that enables robust, drift-resistant 6DoF pose estimation from monocular video.

C. Object-Centric Feed-Forward Reconstruction

Simultaneous online reconstruction and 6DoF pose tracking from monocular video is a highly non-convex problem with many local minima. To stabilize optimization, we first use a feed-forward reconstruction module to regress dense pointmaps $\{\mathbf{x}_u \in \mathbb{R}^3\}$ from the input images and use them to initialize the Gaussian object field.

While DUSt3R[13] targets static scenes, we introduce DUSt3R-M to reconstruct dynamic foreground objects. Given an image pair I_t and $I_{t'}$ with corresponding masks M_t and $M_{t'}$, the network predicts two pointmaps $X^{t;\,tt'}$ and $X^{t';\,tt'} \in \mathbb{R}^{H\times W\times 3}$ along with confidence maps $C^{t;\,tt'}$ and $C^{t';\,tt'} \in [0,1]^{H\times W}$. The first superscript (t or t') denotes the temporal index of the reconstructed surface, while the concatenation tt' indicates that both frames are provided as input. The key difference from DUSt3R is that all pointmaps in DUSt3R-M are expressed in the object coordinate frame at time t, rather than in the camera frame.

Mask Fusion. To focus on masked objects, we augment DUSt3R's image encoder with a mask encoder. Using the same patch-embedding operator $\phi(\cdot)$ as the image branch, a binary mask $M \in \mathbb{R}^{H\times W\times C}$ is embedded into tokens $Z_m = \phi(M)$ that are index-aligned with the image tokens $Z_x = \phi(I)$. Leveraging this alignment, we inject mask cues into the image features via a multiplicative residual gate that

emphasizes foreground objects while maintaining stability:

$$\mathrm{Gate}(F, m; g, \alpha) = F \odot \big(1 + \sigma(g)\,\alpha\, m\big), \tag{3}$$

where g is a learnable gate parameter, $\alpha > 0$ controls the fusion strength, $\sigma(\cdot)$ denotes the sigmoid, and $\odot$ indicates element-wise multiplication. Across the multi-stage decoder, we insert this gate to progressively strengthen mask cues, sharpening spatial focus while, as much as possible, preserving the optimization geometry imparted by the pretrained model.

Loss. Following DUSt3R, our first training objective is 3D point regression. From the ground-truth depth map D and intrinsics K, we back-project valid pixels $i \in \mathcal{V}^v, v \in \{1, 2\}$ to obtain the ground-truth pointmap $\bar{X}_i = D(i)\,K^{-1}$ in 3D, and supervise the prediction $\mathbf{x}_u$ with a Euclidean loss.

$$\ell_{\text{regr}}(v, i) = \left\| X_i^{v,1} - \bar{X}_i^{v,1} \right\|, \tag{4}$$

To suppress erroneous registration and background predictions, we introduce a confidence-based loss with two components: within the object mask, we compute a confidence-weighted regression loss over all valid pixels; outside the mask, we penalize predicted confidence.

$$\begin{aligned}\mathcal{L}_{\text{conf}} = &\sum_{v\in\{1,2\}} \sum_{i\in\mathcal{V}^v} \big(C_i^{v,1}\,\ell_{\text{regr}}(v,i) - \alpha \log C_i^{v,1}\big) \\ &+ \sum_{i\in\tilde{\mathcal{V}}^v} \alpha \log C_i^{v,1},\end{aligned} \tag{5}$$

where $C_i^{v,1}$ denotes the confidence of pixel i in the pointmap of view v, and α is the weighting coefficient.

Optimization. We construct a pairwise graph $\mathcal{H}$ whose nodes $\mathcal{E}$ represent video frames and whose edges e correspond to image pairs. Unlike DUSt3R, we implement this graph in a streaming fashion: for each newly arriving frame, we form pairs only with historical keyframes and feed these pairs to DUSt3R-M to obtain pointmaps $X_i^{u,e}, X_i^{v,e}$ and confidence maps $C_i^{u,e}, C_i^{v,e}$ via feed-forward reconstruction.

Following 6DOPEGS[38], we cluster camera poses by quantizing viewing directions to the vertices of an icosahedron on the unit sphere. Around each vertex, we select the frame with the lowest tracking loss, which is the one most aligned with the current global reconstruction, as the historical keyframe. We pair each selected keyframe with the current frame to form image pairs, which avoids all-to-all computation while providing broad, well-distributed viewpoint coverage for reliable reconstruction.

Finally, we align all pairs to a common coordinate system by introducing a pairwise pose $P_e \in \mathbb{R}^{3\times4}$ and a scale factor σ_e for each image pair. We then jointly optimize the pairwise poses, scale parameters, and node pointmaps by minimizing Eq. (6), yielding a globally consistent reconstruction.

$$\arg\min_{X,P,\sigma} \sum_{e\in\mathcal{H}} \sum_{v\in\mathcal{E}_e} \sum_{i=1}^{hw} \boldsymbol{C}_i^{v,e}\, \|X_i^v - P_e \sigma_e X_i^{v,e}\| \tag{6}$$

where (h, w) is the image size and $\mathcal{E}_e$ the nodes of edge e.

D. Gaussian Splatting Guided Pose Optimization

Although DUSt3R-M can produce object poses, its feed-forward nature introduces non-negligible errors across scenes. Following BundleSDF [9], we maintain an online 3D representation and use differentiable 3DGS rendering to back-propagate reconstruction losses to the pose, yielding an optimization-based tracker. We implement a coarse-to-fine online reconstruction and pose tracking module, termed *Online 3DGS Representation*, comprising:

Coarse pose initialization. Feed-forward reconstructions at small baselines exhibit orientation bias. During tracking, we initialize the object pose by aligning the current image to the 3DGS. Concretely, we set $P_t \leftarrow P_{t-1}$, freeze all 3DGS parameters $\mathcal{G}_{t-1}$, and minimize the reconstruction loss in Eq. (7) while updating only the pose parameters, yielding a coarse estimate for frame t.

$$\mathcal{L}_{\text{tracking}} = \frac{1}{|\Omega|} \sum_{p\in\Omega} \|R(p) - I(p)\|^2, \tag{7}$$

where $R(\cdot)$ denotes the 3DGS-rendered RGB under pose P_t, $I(\cdot)$ is the observed RGB, and Ω is the set of object pixels.

Joint optimization of pose and 3DGS. Given the coarse pose from Section III-D and the geometric prior from Section III-C, we perform online joint optimization of the pose and 3DGS by minimizing a unified objective:

$$\mathcal{L}_{\text{joint}} = \lambda_{\text{phot}}\mathcal{L}_{\text{tracking}} + \lambda_{\text{depth}} \frac{1}{|\Omega|} \sum_{p\in\Omega} C_p^t \big|Z(p) - D(p)\big|. \tag{8}$$

where $Z(\cdot)$ is the 3DGS-rendered depth, $D(\cdot)$ is the prior depth from the geometric graph, and C_p^t denotes the per-pixel confidence at pixel p of frame t provided by the geometric prior module (Section III-C); it down-weights unreliable or occluded pixels in the depth term. This objective combines a photometric rendering loss for appearance alignment with a depth-consistency term that enforces agreement between the rendered depth and the reconstructed geometry, thereby enabling object pose tracking while updating the 3DGS representation online.

E. Adaptive Gaussians Management

For online pose tracking, we formulate object reconstruction as a continuous process that incorporates new views while maintaining a compact representation. To mitigate *point-count inflation* arising from repeated insertions and allocate limited computational and memory resources to information-dense regions, we propose an adaptive Gaussian management pipeline consisting of two coordinated components: (i) an entropy-driven *capacity-aware selection* during Gaussian insertion, and (ii) a lightweight *mask-consistency pruning* step that eliminates spurious Gaussians resulting from imperfect geometry priors.

Capacity-aware selection We discretize the space into voxels of edge length v_{size} and compute a voxel index for each 3D location. For each frame F_t, the image is partitioned once into a $G \times G$ grid (G=128), and a per-cell Shannon entropy $H^{(t)}(r, c)$ is computed from the grayscale histogram.

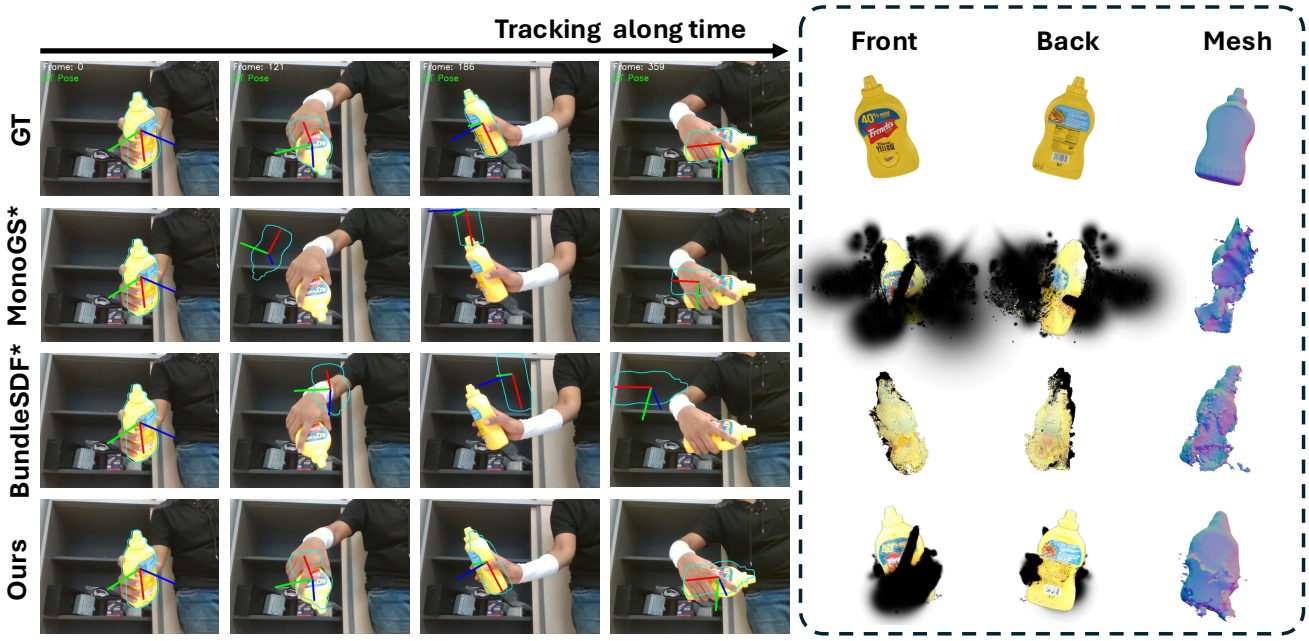

Fig. 3: **Qualitative Comparison of MGS-Track and Baselines on HO3D.** *Left:* tracking results over time on a hand–object sequence (rows: GT, MonoGS*, BundleSDF*, Ours). Poses are visualized by projecting the object frame as colored axes (R/G/B = x/y/z); the *contour of the ground-truth pose projection* is overlaid as a thin outline for reference. *Right:* reconstruction quality from two novel views (**Front/Back**) and the extracted **Mesh**.

Using the camera intrinsics K_t and extrinsics $\left[R_t \mid \mathbf{t}_t\right]$, each voxel center $\mathbf{c}_{i,j,k}$ is projected to pixel coordinates (u, v) and mapped to the grid cell$(r_t, c_t) = \left(\left\lfloor \frac{u}{W} G \right\rfloor, \left\lfloor \frac{v}{H} G \right\rfloor\right)$ for visible, in-bounds projections. We then define the entropy of voxel (i, j, k) at time t as $H^{(t)}(r_t, c_t)$.

We maintain a per-voxel entropy table by updating the previous values $E^{(t-1)}_{i,j,k}$ with a max-pool over the grid-cell entropies $H^{(t)}(r_t, c_t)$, yielding $E^{t}_{i,j,k}$, the allowed capacity for voxel (i, j, k) is

$$N_{\max}(i,j,k) = N_{\text{base}} + \lambda\, N_{\text{base}} \frac{E^{(t)}_{i,j,k}}{E^{(t)}_{\max}} \tag{9}$$

where $E^{(t)}_{\max} = \max_{i,j,k} E^{(t)}_{i,j,k}$. When the candidates in a voxel exceed this budget, we retain the top N_{max}Gaussians ranked by equation as follows,

$$S_i = \alpha\, \text{opacity}_i + \beta/(\|\mathbf{s}_i\|_2 + \delta), \tag{10}$$

the opacity term promotes components that contribute most to the rendered appearance, while the inverse-scale term penalizes overly large (over-blurring) Gaussians, keeping the model compact and concentrating capacity in high-information regions.

Mask-consistency pruning. To mitigate geometric inaccuracies in the initial pointmaps, we apply a lightweight mask-based pruning strategy inspired by [39]. After each training round on F_t, we sample a few reference frames from the past and remove any *newly added* Gaussians whose projections fall outside the object mask in any selected reference frame. This step filters out outliers introduced by noisy priors without affecting well-supported Gaussians.

IV. EXPERIMENTS

A. Experimental Setup

Training data. Following DUSt3R [13], we continue training our model on a mixture of three object-centric datasets, including *CO3Dv2*[40], *Wild6D*[41], and *WildRGBD*[42], to improve performance in this regime. We construct image pairs using the same sampling protocol as DUSt3R, with one change to background handling: instead of replacing the background with random colors, we provide the instance mask as an auxiliary input channel.

Datasets. We evaluate our approach on two public, object-centric datasets: HO3D [16] and YCBInEOAT [15]. HO3D contains hand–object interaction sequences with frequent and severe occlusions; following prior practice, we use RGB frames and adopt the object masks provided by BundleSDF [9]. YCBInEOAT provides monocular videos of everyday objects undergoing active manipulation; for this benchmark, we initialize the first frame with a Segment Anything (SAM) mask and obtain subsequent masks using the video segmentation network XMem [43].

Baselines. For a comprehensive evaluation, we compare our method against three families of baselines on the task of online 6DoF pose estimation for unseen objects from monocular video streams. **(i) Monocular SLAM-based methods:** DROID-SLAM [44] and MonoGS [39], augmented with monocular depth priors predicted by DepthAnythingV2 [12], which is executed independently as an external module. **(ii) Pose-tracking pipelines:** BundleSDF [9] combined with DepthAnythingV2 to enable depth-guided monocular tracking, under the same external-prior augmentation protocol

Method	ADD-S(%)[0-0.1]m ↑					ADD(%)[0-0.1]m ↑				
	AP	MPM	SB	SM	Avg	AP	MPM	SB	SM	Avg
DROID-SLAM [44]	9.17	6.65	0.28	9.32	6.36	3.14	2.38	17.28	4.39	6.80
StreamVGGT [45]	16.27	23.13	13.82	28.12	20.34	7.96	15.43	4.06	11.70	9.79
MonoGS* [39]	12.03	22.37	15.72	19.17	17.32	4.32	9.91	5.66	9.39	7.32
BundleSDF* [9]	40.72	**45.95**	38.52	66.15	47.84	5.19	13.61	27.51	20.95	16.82
MGS-Track	**56.07**	40.95	**49.67**	**70.03**	**54.18**	**31.07**	**23.85**	**29.41**	**50.42**	**33.69**

Method	PSNR ↑					SSIM ↑					Reconstruction CD (cm) ↓				
	AP	MPM	SB	SM	Avg	AP	MPM	SB	SM	Avg	AP	MPM	SB	SM	Avg
DROID-SLAM [44]	—	—	—	—	—	—	—	—	—	—	110.33	99.80	81.86	100.87	98.22
StreamVGGT [45]	—	—	—	—	—	—	—	—	—	—	27.13	14.44	43.29	19.82	26.17
MonoGS* [39]	18.57	20.13	17.89	20.50	19.27	0.84	0.86	0.83	0.86	0.85	25.14	16.49	14.23	10.40	16.57
BundleSDF* [9]	19.85	20.91	17.32	19.87	19.49	0.90	0.91	0.84	0.80	0.86	5.22	**4.95**	8.74	1.61	5.13
MGS-Track	**26.81**	**24.20**	**25.20**	**27.04**	**25.81**	**0.97**	**0.96**	**0.95**	**0.97**	**0.96**	**4.97**	6.28	**7.94**	**1.60**	**5.20**

TABLE I: **Quantitative comparison on the HO3D dataset.** We compare our method with baselines to evaluate reconstruction and tracking performance. MonoGS* and BundleSDF* denote the versions of MonoGS and BundleSDF augmented with geometric priors, respectively.

adopted for the SLAM-based baselines. **(iii) Feed-forward 3D predictors:** methods from the DUSt3R-family [13, 46] and VGGT-family [14, 45] lines of work, adapted to the 6DoF pose estimation setting. To align with the streaming-input setting, we adopt StreamVGGT [45] as the representative VGGT-family baseline. An asterisk (*) denotes variants augmented with external priors.

Metrics. We evaluate performance in terms of both pose tracking and reconstruction quality. For 6DoF object poses, we report the area under the ADD and ADD-S curves within a threshold range of 0–0.1 m, using ground-truth object geometries for metric computation [3, 47]. For reconstruction, we evaluate appearance fidelity using PSNR and SSIM [48], and geometric accuracy using the Chamfer Distance between reconstructed meshes and the corresponding first-frame ground-truth meshes [9].

B. Implementation Details

For training, we curate 0.8M image pairs from CO3D [40], Wild6D [41], and WildRGBD [42]. We first freeze all DUSt3R parameters and train only the Mask Encoder. After 10 epochs, we unfreeze the full network and continue training for an additional 15 epochs to obtain the final model. At inference time, for each video frame, we use object segmentation to rescale and crop the image, so as to focus computation on the target object. Following 3DGS [25], we implement time-critical rasterization and gradient computation in CUDA. The coarse optimization stage runs for 300 iterations, followed by 125 pose-refinement iterations per frame after pose estimation. Training is conducted on 4× NVIDIA H20 GPUs, and inference is performed on an NVIDIA GeForce RTX 4090.

C. Results on the HO3D Dataset

As summarized in Table I, our method outperforms all baselines on HO3D in both 6DoF pose tracking and 3D reconstruction. In particular, it achieves consistently stronger overall performance on pose metrics (ADD / ADD-S) while also improving reconstruction quality in terms of appearance fidelity and geometric accuracy. *DROID-SLAM* performs poorly under object-centric crops. *StreamVGGT* supports end-to-end reconstruction but yields limited performance in the same setting. *MonoGS* provides comparatively strong appearance reconstruction, yet its pose tracking remains weaker and degrades under large, fast object motions. *BundleSDF* shows stronger geometric tracking and recovers coarse object geometry with monocular depth priors, but its appearance reconstruction quality remains limited.

Figure 3 presents qualitative comparisons. Across challenging real-world scenarios, including severe hand occlusions, self-occlusions, texture-poor frames, and specular highlights, our method maintains stable 6DoF pose tracking and reconstructs high-fidelity object appearance. In addition, because our method directly models the observed object instance, the recovered textures and colors more closely match the in-scene object than the scan-based reference models in several cases.

D. Results on the YCBInEOAT Dataset

Method	ADD-S ↑	ADD↑	PSNR↑	SSIM↑
MonoGS [39]	5.28	1.62	20.19	0.90
DROID-SLAM [44]	0.94	0.23	—	—
BundleSDF* [9]	**41.59**	17.32	19.20	0.89
MGS-Track	40.36	**19.79**	**24.22**	**0.94**

TABLE II: **Quantitative comparison on the YCBInEOAT**

The quantitative results on YCBInEOAT (Table II) indicate that our method is well suited to robot-manipulation scenarios characterized by severe occlusions, large inter-frame motions, and small object footprints in the image. Under these conditions, StreamVGGT, which relies on black-background masking, performs poorly because background removal

Method	ADD-S%	ADD%	PSNR ↑	SSIM ↑
w/o DUSt3R-M	35.36	19.73	23.54	0.95
w/o Graph Optimization	56.31	35.16	22.21	0.96
w/o Coarse Pose	67.74	44.31	26.12	**0.97**
w/o Joint Optimization	64.23	42.19	24.59	0.95
Ours	**70.03**	**50.42**	**26.81**	**0.97**

Method	ADD-S%	ADD%	Point numbers	Time
w/o Adaptive Gaussians	66.26	45.55	268611	3h 8min
Ours	**70.03**	**50.42**	182655	**1h 48min**

TABLE III: **Ablation studies of the proposed method.**

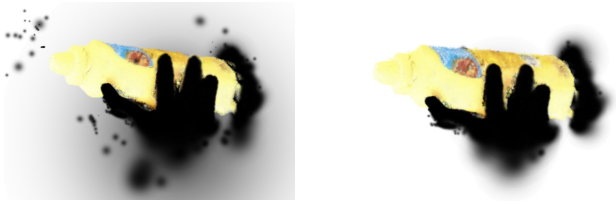

Fig. 4: Impact of our Gaussian pruning strategy on reconstruction quality. The proposed strategy improves geometric accuracy and effectively removes floaters.

leaves insufficient usable visual signal, while MonoGS fails to maintain reliable tracking under large motions. Compared with the prior-enhanced BundleSDF baseline, our method achieves higher ADD accuracy (with slightly lower ADD-S) and, importantly, substantially better reconstruction quality.

E. Ablation Study

We conduct ablation studies on the SB13 sequence of HO3D to evaluate the contribution of each component in our pipeline. Quantitative results are summarized in Table III, and qualitative effects of Gaussian pruning are shown in Figure 4, together providing a consistent view of accuracy, reconstruction quality, and efficiency.

Object-Centric Feed-Forward Reconstruction. Replacing DUSt3R-M with vanilla DUSt3R substantially degrades both pose tracking and reconstruction quality, with ADD-S dropping by about half and ADD showing a similarly large decline. Disabling graph-based post-processing further reduces performance, indicating stronger temporal error accumulation in sequential prediction.

Gaussian Splatting Guided Pose Optimization. Removing the coarse-pose initialization stage degrades pose tracking and reconstruction quality. Further removing joint optimization of pose and the Gaussian representation leads to an additional performance drop, highlighting the benefit of jointly optimizing geometry and camera pose during online reconstruction and refinement.

Adaptive Gaussian Management. Disabling adaptive Gaussian management increases the number of Gaussians and runtime substantially, while also reducing pose accuracy. As shown in Figure 4, the pruning strategy further improves reconstruction quality by suppressing floaters, yielding a better efficiency–quality trade-off and cleaner object-centric reconstructions.

V. CONCLUSION

We address online 6DoF pose tracking and photorealistic 3D reconstruction of previously unseen objects from monocular RGB video. Our framework, *MGS-Track*, unifies object-centric feed-forward geometric priors with rendering-based optimization in 3D Gaussian Splatting. A mask-augmented DUSt3R variant (DUSt3R-M) predicts object-aligned pointmaps and confidences that initialize and guide a streaming pairwise graph, while an online 3DGS representation jointly refines pose and appearance through coarse-to-fine optimization with photometric and depth-consistency objectives. To control model growth and allocate capacity to informative regions, we further introduce adaptive Gaussian management, constrained by voxel entropy, to regulate insertion and pruning using confidence-based ranking. Experiments on HO3D and YCBInEOAT show that MGS-Track consistently delivers robust, drift-resistant pose tracking and faithful appearance reconstruction under challenging object-centric conditions, including severe occlusions, rapid motions, and imperfect initial geometry.

VI. ACKNOWLEDGMENT

This work was supported by the National Key Research and Development Program of China (No. 2024YFE0211000), in part by the National Natural Science Foundation of China (No. 62372329, 62303048, 62506263, 62506264), in part by the Shanghai Scientific Innovation Foundation (No. 23DZ1203400), in part by the China Postdoctoral Science Foundation (No. BX20250383, GZB20250385, 2025M771530, 2025M771539), in part by the Open Found of the Engineering Research Center of Intelligent Swarm Systems, Ministry of Education (ZZU-CISS-2024001), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program.

REFERENCES

[1] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6401–6408.

[2] J.-P. Sleiman, F. Farshidian, M. V. Minniti, and M. Hutter, "A unified mpc framework for whole-body dynamic locomotion and manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4688–4695, 2021.

[3] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[4] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4561–4570.

[5] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. 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*, 2019, pp. 2642–2651.

[6] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Asian conference on computer vision*. Springer, 2012, pp. 548–562.

[7] G. Pitteri, S. Ilic, and V. 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*, 2019, pp. 0–0.

[8] L. Yang, Y. Wu, Y. Deng, R. Tian, X. Hu, and T. Ma, "Uniquadric: A slam backend for unknown rigid object 3-d tracking and light-weight modeling," *IEEE Transactions on Intelligent Transportation Systems*, 2025.

[9] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. 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*, 2023, pp. 606–617.

[10] A. Ganj, Y. Zhao, H. Su, and T. Guo, "Mobile ar depth estimation: Challenges & prospects," in *Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications*, 2024, pp. 21–26.

[11] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "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*, 2023, pp. 18 537–18 546.

[12] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 10 371–10 381.

[13] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20 697–20 709.

[14] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[15] B. Wen, C. Mitash, B. Ren, and K. E. Bekris, "se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 367–10 373.

[16] S. Hampali, M. Rad, M. Oberweger, and V. 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*, 2020, pp. 3196–3206.

[17] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1521–1529.

[18] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "Cosypose: Consistent multi-view multi-object 6d pose estimation," in *European conference on computer vision*. Springer, 2020, pp. 574–591.

[19] K. Park, T. Patten, and M. Vincze, "Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7668–7677.

[20] M. Cai and I. 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*, 2020, pp. 3153–3163.

[21] F. Li, S. R. Vutukur, H. Yu, I. Shugurov, B. Busam, S. Yang, and S. Ilic, "Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2123–2133.

[22] D. Cai, J. Heikkilä, and E. Rahtu, "Gs-pose: Generalizable segmentation-based 6d object pose estimation with 3d gaussian splatting," in *2025 International Conference on 3D Vision (3DV)*. IEEE, 2025, pp. 1001–1011.

[23] L. Luo, S. Sun, J. Yang, L. Zheng, J. Du, and J. Liu, "Object gaussian for monocular 6d pose estimation from sparse views," *arXiv preprint arXiv:2409.02581*, 2024.

[24] B. Wen and K. Bekris, "Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8067–8074.

[25] B. Kerbl, G. Kopanas, T. Leimkühler, G. Drettakis *et al.*, "3d gaussian splatting for real-time radiance field rendering." *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139–1, 2023.

[26] R. Mur-Artal and J. D. Tardós, "Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras," *IEEE transactions on robotics*, vol. 33, no. 5, pp. 1255–1262, 2017.

[27] B. Bescos, J. M. Fácil, J. Civera, and J. Neira, "Dynaslam: Tracking, mapping, and inpainting in dynamic scenes," *IEEE robotics and automation letters*, vol. 3, no. 4, pp. 4076–4083, 2018.

[28] B. Bescos, C. Campos, J. D. Tardós, and J. Neira, "Dynaslam ii: Tightly-coupled multi-object tracking and slam," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 5191–5198, 2021.

[29] S. Yang and S. Scherer, "Cubeslam: Monocular 3-d object slam," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 925–938, 2019.

[30] L. Nicholson, M. Milford, and N. Sünderhauf, "Quadricslam: Dual quadrics from object detections as landmarks in object-oriented slam," *IEEE Robotics and Automation Letters*, vol. 4, no. 1, pp. 1–8, 2018.

[31] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *European conference on computer vision*. Springer, 2016, pp. 628–644.

[32] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y.-G. Jiang, "Pixel2mesh: Generating 3d mesh models from single rgb images," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 52–67.

[33] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang, "Deepmvs: Learning multi-view stereopsis," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2821–2830.

[34] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[35] Y. Fu, S. Liu, A. Kulkarni, J. Kautz, A. A. Efros, and X. Wang, "Colmap-free 3d gaussian splatting," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 20 796–20 805.

[36] C.-H. Lin, W.-C. Ma, A. Torralba, and S. Lucey, "Barf: Bundle-adjusting neural radiance fields," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 5741–5751.

[37] B. Ye, S. Liu, H. Xu, X. Li, M. Pollefeys, M.-H. Yang, and S. Peng, "No pose, no problem: Surprisingly simple 3d gaussian splats from sparse unposed images," *arXiv preprint arXiv:2410.24207*, 2024.

[38] Y. Jin, V. Prasad, S. Jauhri, M. Franzius, and G. Chalvatzaki, "6dope-gs: Online 6d object pose estimation using gaussian splatting," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025, pp. 8032–8043.

[39] H. Matsuki, R. Murai, P. H. Kelly, and A. J. Davison, "Gaussian splatting slam," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024, pp. 18 039–18 048.

[40] J. Reizenstein, R. Shapovalov, P. Henzler, L. Sbordone, P. Labatut, and D. Novotny, "Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 901–10 911.

[41] Y. Fu and X. Wang, "Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27 469–27 483, 2022.

[42] H. Xia, Y. Fu, S. Liu, and X. Wang, "Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos," 2024. [Online]. Available: https://arxiv.org/abs/2401.12592

[43] H. K. Cheng, Y.-W. Tai, and C.-K. Tang, "Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5559–5568.

[44] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[45] D. Zhuo, W. Zheng, J. Guo, Y. Wu, J. Zhou, and J. Lu, "Streaming 4d visual geometry transformer," *arXiv preprint arXiv:2507.11539*, 2025.

[46] B. Smart, C. Zheng, I. Laina, and V. A. Prisacariu, "Splatt3r: Zero-shot gaussian splatting from uncalibrated image pairs," *arXiv preprint arXiv:2408.13912*, 2024.

[47] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen, "Fs6d: Few-shot 6d pose estimation of novel objects," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 6814–6824.

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.