IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

HIPPo: Harnessing Image-to-3D Priors for Model-Free Zero-Shot 6D Pose Estimation

Yibo Liu, *Member, IEEE*, Zhaodong Jiang, Binbin Xu, Guile Wu, Yuan Ren, Tongtong Cao, Bingbing Liu, Rui Heng Yang, Amir Rasouli, and Jinjun Shan, *Senior Member, IEEE*

***Abstract*—This work focuses on the problem of 6D pose estimation for novel objects when a reference 3D model or posed reference images are not available. While existing methods can estimate the precise 6D pose of objects, they heavily rely on curated CAD models or reference images, the preparation of which is a time-consuming and labor-intensive process. Moreover, in real-world scenarios, 3D models or reference images may not be available in advance and instant robot reaction is desired. In this work, we propose a novel framework named HIPPo, which eliminates the need for curated CAD models and reference images by harnessing image-to-3D priors from Diffusion Models, enabling model-free zero-shot 6D pose estimation. Specifically, we construct HIPPo Dreamer, a rapid image-to-mesh model built on a multiview Diffusion Model and a 3D reconstruction foundation model. Our HIPPo Dreamer can generate a 3D mesh of any unseen objects from a single glance in just a few seconds. Then, as more observations are acquired, we propose to continuously refine the diffusion prior mesh model by joint optimization of object geometry and appearance. This is achieved by a measurement-guided scheme that gradually replaces the plausible diffusion priors with more reliable online observations. Consequently, HIPPo can instantly estimate and track the 6D pose of a novel object and maintain a complete mesh for immediate robotic applications. Thorough experiments on various benchmarks show that HIPPo outperforms state-of-the-art methods in 6D object pose estimation when prior reference images are limited.**

***Index Terms*—Computer vision, diffusion models, generative AI, pose estimation.**

I. Introduction

6D POSE estimation [1], [2], [3] is crucial for robotic applications such as grasping, navigation, exploration, and collision avoidance. Although many 6D pose estimation methods [1], [2], [3], [6], [7], [8] exist, as shown in Fig. 1, they often demand the textured CAD model of the object in advance. Crafting a curated CAD model is time-consuming and labor-intensive. Thus, some research [9], [10] focuses on using reference images or a video of the object as input instead of a 3D model. Nevertheless, many of them [1], [9], [10] still need to perform object reconstruction to transform the reference images into a textured 3D model or require posed images of the object as a reference [2]. Unfortunately, in real-world robotic applications, 3D models or reference images may not be available a priori, limiting the system's deployment in open-world scenarios where object models are not always accessible. Recently, image-to-3D methods [11], [12], [13] have shown robust zero-shot prediction capabilities. Specifically, Diffusion Models [11], [12] trained on the large-scale dataset [14] can render novel views of arbitrary unseen objects. Inspired by this, we aim to harness the learned image-to-3D priors from Diffusion Models [11], [12] to

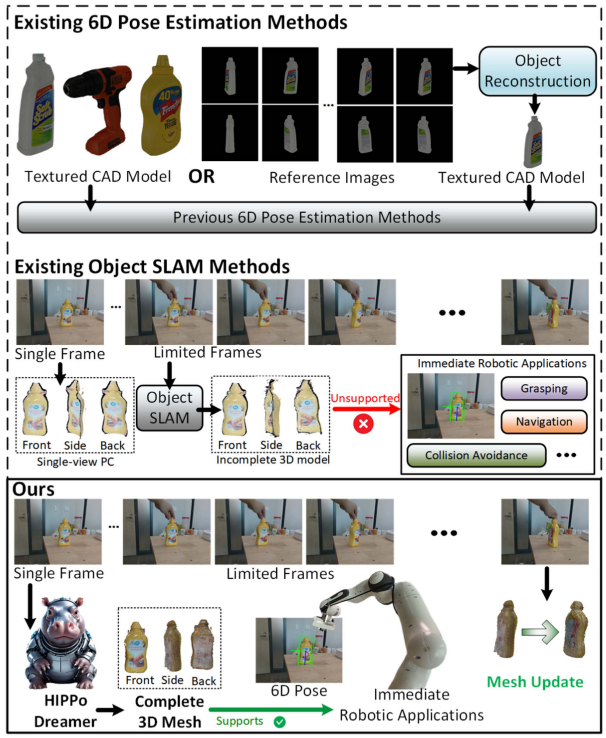

Fig. 1. Compared to existing SOTA 6D pose estimation methods [1], [2], [3], HIPPo eliminates the need for a textured 3D model or reference images in advance, while also optimizing the reference 3D model online. Compared to existing object SLAM methods [4], [5], HIPPo sustains a complete 3D model from the first glance of the object, enabling immediate robotic applications.

Received 14 February 2025; accepted 16 June 2025. Date of publication 2 July 2025; date of current version 9 July 2025. This article was recommended for publication by Associate Editor Y. Xiang and Editor M. Vincze upon evaluation of the reviewers' comments. *(Yibo Liu and Zhaodong Jiang are co-first authors.) (Corresponding author: Tongtong Cao.)*

Yibo Liu was with Huawei Noah's Ark Lab, Markham, ON L3R 5A4, Canada. He is now with York University, Toronto, ON M3J 1P3, Canada (e-mail: buaayorklau@gmail.com).

Zhaodong Jiang was with Huawei Noah's Ark Lab, Markham, ON L3R 5A4, Canada. He is now with the University of Toronto, Toronto, ON M5S 1A1, Canada (e-mail: zhaodong.jiang@mail.utoronto.ca).

Binbin Xu, Guile Wu, Yuan Ren, Tongtong Cao, Bingbing Liu, Rui Heng Yang, and Amir Rasouli are with Huawei Noah's Ark Lab, Markham, ON L3R 5A4, Canada (e-mail: binbin.xu@huawei.com; guile.wu@huawei.com; yuan.ren3@huawei.com; caotongtong@huawei.com; liu.bingbing@huawei.com; rui.heng.yang2@huawei.com; amir.rasouli@huawei.com).

Jinjun Shan is with York University, Toronto, ON M3J 1P3, Canada (e-mail: jjshan@yorku.ca).

Project page: https://hippope.github.io/.

Digital Object Identifier 10.1109/LRA.2025.3585384

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

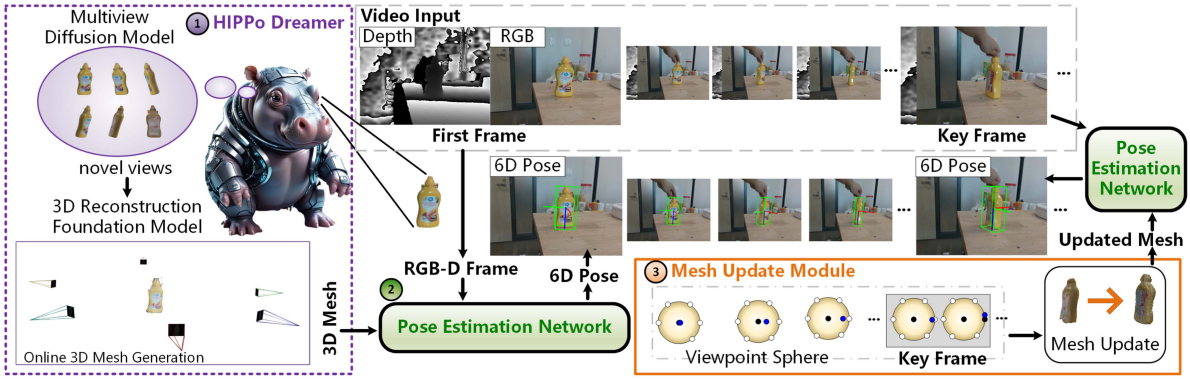

Fig. 2. **Overview of HIPPo.** The HIPPo framework consists of three components. ① HIPPo Dreamer (Section III-A) is responsible for initializing HIPPo by instantly generating a 3D mesh from the detected first frame. ② Pose estimation network (Section III-B) estimates the 6D pose based on the reference mesh. ③ Mesh update module (Section III-C) refines the mesh by updating the diffusion prior with more reliable online measurements.

boost 6D pose estimation without relying on CAD models or reference images.

Yet this task is challenging due to the two limitations of existing image-to-3D methods [11], [12], [13]. First, 6D pose estimation [1] requires the reference 3D model to have the same scale as the real-world object, whereas image-to-3D methods do not account for the scale of the generated model. Second, the image-to-3D problem is inherently ill-posed [13]: given a conditioned image, the uncaptured views can exhibit many plausible appearances and geometries, causing discrepancies between the generated 3D model and the actual views. Unfortunately, existing image-to-3D methods focus solely on generating a complete 3D model, rather than adapting to new observations. In addition, 6D pose estimation methods [1], [9], [10] treat object reconstruction as a preprocessing step for pose estimation and do not consider the optimization of the reference model during pose estimation. Although the object SLAM methods [4], [5] can conduct simultaneous 6D pose estimation and model refinement, as shown in Fig. 1, they often produce incomplete models from limited views, making their models less suitable for immediate robotic applications.

In this work, we propose HIPPo, a novel framework that leverages image-to-3D priors for model-free zero-shot 6D pose estimation. As shown in Fig. 1, compared to existing 6D pose estimation methods [1], [2], [3], HIPPo eliminates the need for preparing a textured CAD model in advance. It can be initialized from any first glance of the object and simultaneously estimates the 6D pose while optimizing the 3D model of the object online. For the initialization of HIPPo, we design HIPPo Dreamer, a rapid image-to-mesh strategy based on a multiview Diffusion Model [11] and a 3D reconstruction foundation model [15]. It generates a 3D mesh of an unseen object from a single reference image in just a few seconds and can recover the mesh's physical scale. Recognizing that the diffusion-based mesh has limited reliability in uncaptured views due to the ill-posed nature [13], we further develop a measurement-guided algorithm to continuously optimize the mesh. Specifically, a viewpoint sphere tracks the relative pose changes between frames. Mesh optimization is triggered when a keyframe is recognized to replace the diffusion prior with more reliable online measurements of appearance and geometry. We conduct extensive experiments on various challenging benchmarks. Experimental results demonstrate that the proposed method significantly outperforms state-of-the-art (SOTA) 6D pose estimation methods [1], [2], [3] when prior reference images are limited.

The **contributions** of this work are:

- We develop a novel 6D pose estimation framework (see Fig. 2), named HIPPo, that is able to estimate the 6D pose of an unseen object even without a textured CAD model or a dense set of posed reference images.
- We propose an instant image-to-mesh strategy called HIPPo Dreamer. HIPPo Dreamer acts as an alternative method to BundleSDF [5], the most widely used method for collecting 3D models for pose estimation. Compared to BundleSDF, which requires several minutes and dense observations to reconstruct a complete object, HIPPo Dreamer generates a 3D model in a few seconds and consistently maintains a complete 3D mesh from the first glance.
- We design a measurement-guided method to optimize the mesh online, enhancing mesh fidelity with more observations and surpassing the diffusion prior mesh.

II. Related Work

Instance-level Image-to-3D Methods: Instance-level image-to-3D approaches [11], [12], [13] aim to generate 3D representations from a single image. Specifically, Diffusion Models [11], [12] have demonstrated strong zero-shot prediction abilities, benefiting from training on large-scale datasets such as Objaverse [14]. Since reconstruction quality is usually prioritized over efficiency, they take several minutes and even tens of minutes to reconstruct 3D models, which limits their applicability in real-time scenarios. While some methods [12], [13] can achieve faster image-to-3D generation, they generally do not incorporate incremental optimization of the 3D model.

6D Pose Estimation Methods: The implementation[1] of most existing 6D pose estimation methods, such as FoundationPose [1], GigaPose [3], SAM6D [2], FoundPose [6], ZS6D [7], and MegaPose [8], requires a textured CAD model in advance,

[1]Their pose estimation networks require dense posed reference images, which are sampled from a textured CAD model in their pipelines.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

which requires intensive time and labor to craft [16]. For example, when handling an unseen object, FoundationPose [1] requires running BundleSDF [5] to generate the reference 3D model, which takes tens of minutes to complete. Similarly, OnePose [10] and OnePose++ [9] propose recording a video scan of the object and utilizing Structure-from-Motion (SfM) to reconstruct the object. However, the prior operation involving the reference 3D model or reference image sampling may be impractical in real-world settings, especially when instant robotic action is required at first sight of the object. Recently, Zero123-6D [17] proposes leveraging the Diffusion Model for 6D pose estimation. However, the category-level pose estimation strategy of Zero123-6D does not fully utilize the instance-level object generation ability of Diffusion Models [11] and, like previous methods [1], [2], [3], [9], [10], does not consider further model optimization during pose estimation. In contrast, object SLAM methods [4], [5] reconstruct objects in real-time without prior knowledge, tracking and optimizing the geometry [18] and appearance [19] from scratch. However, they struggle to provide complete models when observations are scarce.

III. Methodology

A. HIPPo Dreamer

Despite the existence of many image-to-3D methods [11], [12], some technology gaps remain in utilizing them to instantly generate 3D reference models for the 6D pose estimation application. In this section, we first analyze the existing technology gaps, which then inspire the design of corresponding modules in HIPPo Dreamer to address them. Our analysis and detailed solutions to these gaps constitute the key contributions of HIPPo Dreamer.

1) Scale Recovery Problem Analysis: The correct scale of the reference 3D model is the prerequisite for 6D pose estimation. In this work, scale recovery refers to finding the constant scale, s, represented by:

$$s = g_{\max}/r_{\max} \tag{1}$$

where $g_{\max}$ denotes the maximum side length of the oriented bounding box (OBB) [20] of the generated model, while $r_{\max}$ represents that of the real-world object. Suppose that the camera intrinsics are known, $r_{\max}$ can be obtained at the first frame through depth measurement after object segmentation. But it should be noted that $r_{\max}$ corresponds to the OBB that encloses the partial point cloud of the real-world object captured from the view of the first frame, denoted by the original view. Therefore, to recover the scale, we need to compute $g_{\max}$ by leveraging the estimated depth of the original view. For this reason, InstantMesh [12], though addressing the instant image-to-mesh problem, is not applicable here as it generates fixed views independent of the original view and does not provide direct depth estimation for it. Consequently, it is necessary to develop a new image-to-mesh framework that supports scale recovery.

2) Framework Design: **(1) Object Segmentation.** Given a frame of an unseen object, we use a guiding prompt and Grounding DINO [21] to segment the object first from the RGB image and then use the same mask to segment the object from the depth image. **(2) Instant Image-to-multiview-to-mesh Generation.** Inspired by InstantMesh [12], we apply an image-to-multiview-to-mesh strategy. In particular, we first employ the multiview Diffusion Model [2] from Wonder3D [11] for image-to-multiview generation. This is because the first view among the multiple views generated by Wonder3D always matches the original view. If the depth of the first view is estimated, scale recovery follows from (1). Thereafter, we adopt a modified MASt3R [15], a 3D reconstruction foundation model, for simultaneous depth estimation for scale recovery and multiview-to-mesh conversion. We modify MASt3R since it considers the background in optimization and relies on a hyperparameter, *the minimum confidence threshold*, to remove the background from the 3D reconstruction result. Unfortunately, as shown in Fig. 3(a), (b), and (c), this hyperparameter requires careful fine-tuning for each object. To address this problem, we modify the background point masking process in the vanilla MASt3R by integrating SAM [22] into the process to generate accurate object masks, instead of simply masking out all points with matching confidence below the threshold. Then, to further remove the artifacts, we apply the Statistical Outlier Removal (SOR) filter, defined as

Fig. 3. Comparison of the vanilla MASt3R (a), (b), and (c) and our modified MASt3R (d). (a): A low threshold preserves too many background points. (b): A high threshold results in an incomplete model by masking out some foreground object points. (c): Even with careful fine-tuning, artifacts may remain around the object, affecting the judgment of its scale. (d): Our modified MASt3R generates artifact-free 3D models without requiring fine-tuning of the hyperparameter.

$$\|\mathbf{p} - \mu\| > k\sigma \tag{2}$$

where $\mathbf{p}$ is a point in the cloud, μ is the mean position of its N nearest neighbors, σ is the standard deviation of distances, and k denotes threshold ($k = 300$ in practice). Accordingly, as shown in Fig. 3(d), our modified MASt3R can robustly generate 3D point clouds free of background points, without requiring hyperparameter fine-tuning. **(3) Scale Recovery.** An estimated partial point cloud of the object can be obtained through the estimation of depth and intrinsics of the original view by MASt3R. By computing the OBB of it, $g_{\max}$, as shown in (1), is acquired and the scale is determined through through (1). However, we experimentally found that for both measured and estimated depth images, even when employing SAM [22] to provide object masks, the partial point clouds of the object remain noisy, particularly around the edges, affecting the values of $r_{\max}$ and $g_{\max}$. Therefore, we apply the SOR filter, as shown in (2), to denoise both the measured and estimated point clouds before computing the scale. Although straightforward, this SOR denoising step is effective and necessary for scale recovery.

[2] Although Wonder3D [11] provides a complete solution for image-to-mesh, we only adopt its Diffusion Model for two reasons. First, its multiview-to-mesh step takes several minutes, which is not instant. Second, it does not directly provide depth estimation of the first (original) view.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

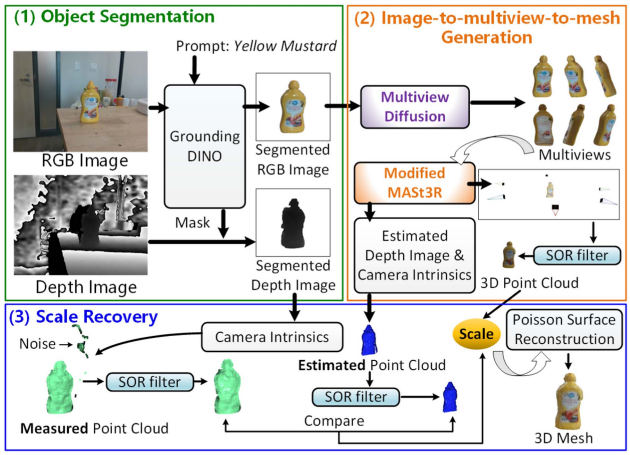

Fig. 4. An illustration of the HIPPo Dreamer pipeline. **Step 1**: The object is segmented using Grounding DINO [21], and the same mask is used to segment the object's depth image. **Step 2**: The Multiview Diffusion Model [11] and modified MASt3R [15] transform the object's RGB image into a 3D point cloud instantly, and the depth image of the original view is estimated. **Step 3**: Using the camera intrinsics, the measured and estimated depth images are converted into 3D point clouds. After denoising with the SOR filter, the scale is determined by comparing the point clouds, and Poisson surface reconstruction generates the 3D mesh.

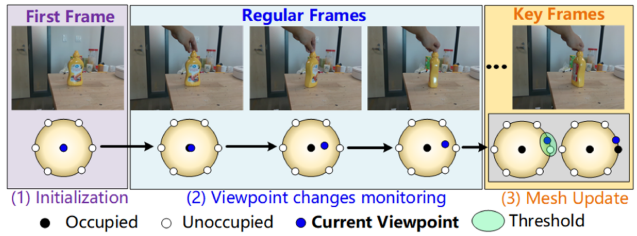

Fig. 5. An illustration of the viewpoint sphere. **(1)** A viewpoint sphere is initialized at the first frame, and an arbitrary viewpoint on it is occupied. **(2)** Relative viewpoint changes are monitored. If the current viewpoint does not align with any unoccupied viewpoints, it is considered a regular frame, indicating that no mesh update is required. **(3)** If the alignment between the current frame and an unoccupied viewpoint exceeds a predefined threshold, it is recognized as a key frame, and a mesh update is triggered.

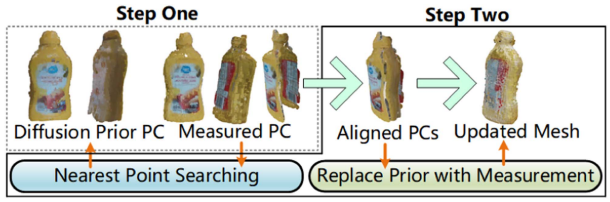

Fig. 6. An illustration of the proposed mesh update method. It consists of two steps. In step one, the measured point cloud and diffusion prior point cloud are aligned. In step two, the diffusion prior is replaced with measurements based on nearest point searching.

Finally, we convert the scaled 3D point cloud into a 3D mesh using Poisson surface reconstruction.

In summary, Fig. 4 shows the major steps of the proposed HIPPo Dreamer.

B. 6D Pose Estimation

Once a textured 3D mesh is generated by the proposed HIPPo Dreamer, we employ the Pose Refinement network from FoundationPose [1] for 6D pose estimation due to its SOTA performance. The network takes two inputs: one is a rendering of the generated object conditioned on the most recent pose estimate, and the other is a cropped observation from the camera. The Siamese Network comprises two feature embedding networks with shared weights that extract feature maps from the two RGB-D input branches. These feature maps are then concatenated and fed into additional CNN blocks, where they are tokenized by dividing them into patches with positional embeddings. A transformer then predicts a pose update, iteratively refining the pose estimation over a few iterations.

C. Mesh Update Module

Due to the ill-posed nature of the image-to-3D problem, HIPPo Dreamer may generate varying predictions of unseen views with low-fidelity rendering and geometry. This is unfavorable for 6D pose estimation, as the inconsistency between the reference model and the real-world object disrupts the matching between the rendered and measured views (see Section III-B). Thus, as more observations become available, it is desirable to update the reference mesh. To enable rapid reference mesh updates, a problem not considered in previous pose estimation and image-to-3D research [1], [2], [3], [11], [12], [13], we propose using viewpoint variation as a trigger for mesh updates.

We first design a viewpoint sphere to select new perspectives with significant viewpoint shifts and filter out redundant poses. As shown in Fig. 5, there are N viewpoints uniformly distributed on the viewpoint sphere. Then, frames are classified as keyframes for updates when they align with unoccupied viewpoints. We set a tolerance threshold to account for slight misalignments, ensuring effective yet controlled updates. The pose of the first frame is aligned with an arbitrary viewpoint on the sphere and we only monitor the relative rotation between frames. The corresponding viewpoint on the sphere is marked as occupied after conducting mesh update and will not trigger further mesh updates.

Next, we propose a mesh update method based on the modality of the 3D colored point cloud, as shown in Fig. 6.

(1) Step One: We register all the measured points into the first frame and apply the SOR filter to denoise the point cloud. Then, we transform the measured points into the object frame using the pose estimation result of the first frame. In this way, as seen in step one of Fig. 6, both the 3D models provided by HIPPo Dreamer and the online measurements are registered in the object frame, with their scales matched in Section III-A2.

(2) Step Two: We further build a KDTree on the diffusion prior 3D point cloud. For each point in the measured point cloud, we search for the nearest point in the diffusion prior mesh and replace the 3D position and color of the prior point with those of the measured point. Thereafter, the updated 3D colored point cloud is transformed into a mesh using Poisson surface reconstruction. In practice, we apply Farthest Point

TABLE I
QUANTITATIVE COMPARISON WITH SOTA METHODS ON YCB-V [23], LM-O [24], T-LESS [25], AND TUD-L [26]. IMG. REFERS TO THE NUMBER OF REFERENCE IMAGES

Dataset	Method	GigaPose [3]			SAM-6D [2]			FoundationPose [1]			**Ours**
	Img.	1	8	16	1	8	16	1	8	16	1
YCB-V	ADD	16.42	24.70	61.71	23.85	38.67	90.06	35.32	51.59	96.56	**89.07**
	ADD-S	32.85	43.73	82.23	40.62	71.28	98.54	79.71	90.24	99.47	**97.00**
LM-O	ADD	15.87	30.62	66.53	21.16	37.00	87.53	34.64	42.64	92.20	**88.49**
	ADD-S	31.89	54.45	87.45	39.70	70.35	93.00	68.16	82.23	97.17	**92.92**
T-LESS	ADD	28.77	31.27	63.22	33.62	43.20	84.33	37.37	45.82	90.10	**85.27**
	ADD-S	44.29	54.45	86.4	70.11	75.15	95.00	82.16	88.33	98.31	**94.70**
TUD-L	ADD	32.27	42.22	76.33	41.67	56.61	89.33	56.21	62.21	95.30	**90.23**
	ADD-S	52.21	72.65	92.25	76.90	83.31	98.22	72.33	80.51	97.20	**95.22**

Sampling [18] to downsample the measured point cloud if the number exceeds 30,000 to ensure that the mesh update can be completed within a few seconds.

IV. EXPERIMENTS

A. Comparison With SOTA 6D Pose Estimation Methods

Benchmarks and Experimental Setup: We evaluate the proposed HIPPo on four popular datasets: YCB-V [23], LM-O [24], T-LESS [25], and TUD-L [26]. In particular, for each object, we have three levels of reference images: 1, 8, and 16. One reference image represents a single glance of the object. For a fair comparison, the reference image provided to all methods is taken from the object in the first detected frame. If the object is severely occluded in the first frame, we render a reference image using the ground-truth relative pose and object model. Next, using the ground truth object pose from the first frame and the ground truth object model, we render 8 and 16 images by rotating the virtual RGB-D camera around the z-axis of the object frame in Blender. Then, we apply BundleSDF [5] to reconstruct the textured CAD models from the reference images. Following [1], 16 reference images are sufficient to construct a complete model, while the 3D shape is relatively incomplete when reconstructed from 8 images and extremely incomplete when using just 1 image. We set 36 key points on the viewpoint sphere (see Fig. 5).

Competitors and Metrics: We compare HIPPo with three SOTA 6D pose estimation methods: FoundationPose [1], SAM-6D [2], and GigaPose [3]. The implementations of the competitors require textured CAD models. Following [1], we use the Area Under the Curve (AUC) of ADD and ADD-S [23] as metrics to evaluate 6D pose estimation performance.

Experimental Results and Analysis: Qualitative and quantitative comparisons of our approach against SOTA methods [1], [2], [3] are presented in Fig. 7 and Table I, respectively. As shown across the four datasets, although the SOTA methods [1], [2], [3] achieve very promising results when 16 reference images are available, which indicates access to a complete reference mesh, their performance degrades significantly as the number of reference images decreases. In contrast, our method, which always uses only one reference image, is slightly inferior to FoundationPose [1] and SAM-6D [2], but outperforms GigaPose [3] when they are provided with 16 reference images. However, when reference images are insufficient, which is expected to occur in immediate robotic applications, the proposed method demonstrates significantly superior performance over the SOTA competitors. Despite the decent results, we also found that severe object occlusion can hinder the performance of our approach. A discussion is provided in Section IV-E.

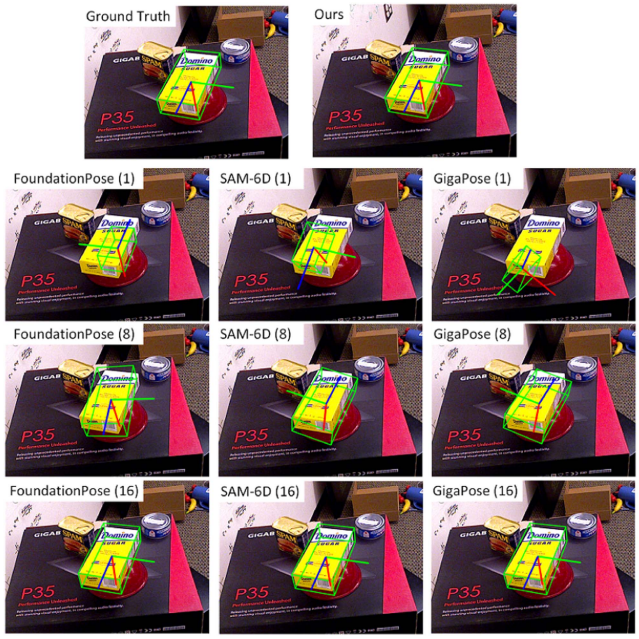

Fig. 7. Qualitative comparison of the proposed method against SOTA methods on the YCB-V dataset [23]. The number in brackets refers to the number of reference images known in advance.

B. Comparison With Object 3D Reconstruction Methods

Benchmark and Experimental Setup: In this section, we test HIPPo on a custom dataset to evaluate its object reconstruction quality and efficiency. In particular, this dataset contains RGB-D frames of four objects: three rendered using Blender and one collected in the real world. The first three objects, YCB_3 (sugar box), YCB_5 (mustard bottle), and YCB_{15} (power drill), belong to the YCB object set [27]. They are rendered with a virtual RGB-D camera at a resolution of 512×512, following a loop trajectory to capture 16 evenly spaced frames per object. The ground truth shapes are their CAD models from the YCB object set [27]. The experimental setup to scan the

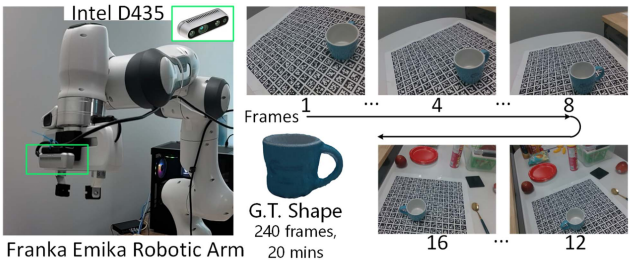

Fig. 8. An illustration of the experimental setup. The frames are captured by an Intel D435 RGB-D camera (1280 × 720) mounted on a 7-DoF Franka Emika robotic arm.

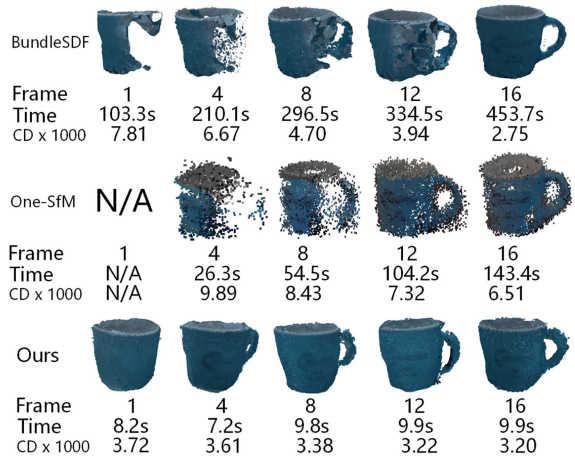

Fig. 9. Qualitative comparison with BundleSDF [5] and One-SfM [9] under different numbers of frames.

real-world object is shown in Fig. 8. The ground truth shape of the mug is constructed using BundleSDF [5] with dense observations consisting of 240 RGB-D frames, taking approximately 20 minutes to complete. In particular, the robotic arm follows multiple pre-designed ring-view trajectories to scan the mug at different relative altitudes, completing the first 360-degree loop at the 16th frame. Moreover, the rendered depth images are noise-free, while the real-world sampled depth images are noisy. All experiments are conducted on the same platform.

Competitor and Metrics: Although many object 3D reconstruction methods exist, we focus only on those designed for collecting 3D models for 6D pose estimation. Thus, the competitors are BundleSDF [5] and the 3D object reconstruction approach proposed in OnePose++ [9], denoted as One-SfM. For the competitors, ground-truth object poses are perfectly known for rendered frames, while relative poses for real-world sampled frames are obtained using SfM. In contrast, HIPPo always uses its own pipeline to estimate relative poses. We use the depth information to recover the scale of the 3D point cloud reconstructed by One-SfM. We apply the Chamfer Distance (CD) to evaluate the quality of the reconstructed object. The objects are normalized into a unit sphere before computing the CD, and the CD values are multiplied by 10^3 for display.

Experimental Results and Analysis: The qualitative and quantitative comparisons of HIPPo against BundleSDF [5] and One-SfM [9] are presented in Fig. 9 and Table II, respectively. Specifically, Fig. 9 shows that our method always maintains

TABLE II
QUANTITATIVE COMPARISON WITH BUNDLESDF AND ONE-SFM

Object	Method	Metric	Number of Frames				
			1	4	8	12	16
YCB_{15}	BundleSDF	CD $\times 10^3$	6.19	5.21	4.27	3.60	**2.52**
		Time (s)	110.6	114.2	123.6	135.8	142.3
Power	One-SfM	CD $\times 10^3$	–	9.01	8.34	6.95	6.31
		Time (s)	–	22.4	49.7	96.4	127.1
Drill	Ours	CD $\times 10^3$	**3.53**	**3.44**	**3.19**	**3.08**	2.97
		Time (s)	8.1	3.1	5.8	8.7	9.6
YCB_5	BundleSDF	CD $\times 10^3$	5.62	4.69	3.88	3.13	**2.18**
		Time (s)	105.5	109.3	117.2	121.4	135.5
Mustard	One-SfM	CD $\times 10^3$	–	9.17	8.48	7.09	6.35
		Time (s)	–	21.5	48.6	92.3	123.1
Bottle	Ours	CD $\times 10^3$	**3.12**	**3.01**	**2.89**	**2.86**	2.77
		Time (s)	8.1	1.8	4.6	7.3	9.3
YCB_3	BundleSDF	CD $\times 10^3$	4.48	3.52	2.75	2.27	**2.01**
		Time (s)	102.8	109.9	123.2	133.6	136.9
Sugar	One-SfM	CD $\times 10^3$	–	8.94	7.15	6.88	6.11
		Time (s)	–	21.2	44.7	91.6	113.8
Box	Ours	CD $\times 10^3$	**3.10**	**3.05**	**2.68**	**2.15**	2.63
		Time (s)	8.1	1.4	3.6	6.1	9.2
Real	BundleSDF	CD $\times 10^3$	7.81	6.67	4.70	3.94	**2.75**
		Time (s)	103.3	210.1	296.5	334.5	453.7
World	One-SfM	CD $\times 10^3$	–	9.89	8.43	7.32	6.51
		Time (s)	–	26.3	54.5	104.2	143.4
Mug	Ours	CD $\times 10^3$	**3.72**	**3.61**	**3.38**	**3.22**	3.20
		Time (s)	8.2	7.2	9.8	9.9	9.9

TABLE III
COMPUTATIONAL TIME ANALYSIS OF HIPPO

Process	Object Segmentation	Image-to--Multiview	Multiview-to -Mesh	6D Pose Estimation	Mesh Update
Time (s)	0.04	2.01	6.05	0.03	9.87

a complete 3D model from the first frame compared to competitors, with significantly better efficiency. Although One-SfM shows better efficiency than BundleSDF, it is not applicable when only a single frame is available and produces a relatively noisy 3D model compared to both BundleSDF and our method. Table II demonstrates that the 3D model fidelity of our method is superior to that of the competitors when the number [3] of frames is fewer than 16. The inference time of our method on the first frame is always 8 s, as the proposed HIPPo Dreamer processes a normalized input through a fixed pipeline. Our inference time on other frames is fewer than 10 s since we downsample the accumulated measured point cloud to 30,000 points if it exceeds this number. As introduced in Section III-C, the only time-consuming process in mesh update is the KDTree-based nearest point searching. An ablation study regarding this is provided in Section IV-C. Moreover, the continuous improvement in our model's fidelity demonstrates the benefit of the proposed mesh update module. BundleSDF [5] outperforms our method when the number of frames is 16, but its efficiency is notably lower than ours. An analysis is presented in Sec IV-E. One-SfM [9] is inferior to our method in both quality and efficiency across all settings.

C. Computational Time Analysis and Ablation Studies

Computational Time Analysis: We first report the computational time of the major steps of our method in Table III, where the results correspond to the tests shown in Table I. As seen, HIPPo can estimate the 6D pose of a novel object at around 15 FPS (object segmentation and 6D pose estimation) after the initialization of HIPPo Dreamer (image-to-multivew-to-mesh),

[3]Note that the number of frames is a different concept from the number of reference images in Section IV-A. Frames are the consecutive RGB-D frames provided to the methods. Reference images are known images used to reconstruct the model prior to 6D pose estimation.

TABLE IV
ABLATION STUDY ON THE EFFECT OF MESH UPDATE FREQUENCY

Num. of Viewpoint	1	25	36	64
ADD	86.57	88.71	89.07	90.31
ADD-S	94.62	96.55	97.00	97.21

TABLE V
ABLATION STUDY ON THE EFFECT OF THE NUMBER OF PRESERVED POINTS AFTER DOWNSAMPLING FOR MESH UPDATE

Num. of Points	5000	10,000	20,000	30,000	40,000
Time (s)	1.73	3.52	7.26	9.87	14.34
CD $\times 10^3$	3.55	3.42	3.30	3.17	3.13

TABLE VI
ABLATION STUDY REGARDING SCALE RECOVERY AND THE SOR FILTER

Method	w/o Scale Recovery w/o SOR Filter	w/o Scale Recovery w/ SOR Filter	w/ Scale Recovery w/o SOR Filter	w/ Scale Recovery w/ SOR Filter
ADD	15.77	41.20	80.87	89.07
ADD-S	31.43	75.64	92.35	97.00

which takes around 8 seconds. The most time-consuming process is the mesh update, and a more detailed analysis of this process is provided in the following ablation studies.

Effect of Mesh Update Frequency: We first study the effect of mesh update frequency. In particular, the frequency is controlled by the number of viewpoints on the sphere shown in Fig. 5. The result on YCB-V [23] is presented in Table IV. As seen, increasing the mesh update frequency can slightly improve the 6D pose estimation performance. However, considering that mesh update also consumes time, it is not practical to update the mesh with a high frequency in real applications. We choose 36 viewpoints for the balance of accuracy and efficiency. Moreover, one viewpoint on the sphere indicates that no mesh update is needed, and HIPPo can consistently run at 15 FPS after initialization. This is a feasible solution for scenarios where no time is allowed for mesh updates or collecting the object's asset is not necessary. However, Fig. 9 demonstrates the necessity of mesh updates for collecting high-fidelity 3D assets.

Effect of the Number of Points on Mesh Update: We then study the impact of the number of preserved points after downsampling. In Table V, we report the effect of the number of preserved measured points after downsampling, as the efficiency of the mesh update is determined by it. We use the mug scenario shown in Fig. 9 as an example and focus on the mesh optimization case with 16 frames. As shown, preserving more points, though beneficial for improving reconstruction quality, leads to a longer inference time. Thus, we opt to preserve 30,000 points for the quality-efficiency balance.

Effects of Scale Recovery and the SOR Filter: We study the effects of scale recovery, introduced in Section III-A2, and the SOR filter, which is used for denoising point clouds. The result on YCB-V [23] is presented in Table VI. As seen, since the correct scale of the reference mesh is a prerequisite for 6D pose estimation, removing both the scale recovery and the SOR filter leads to severe performance degradation. In particular, scale recovery impacts the performance more, as the SOR filter only removes the effect of noise on scale, while scale recovery directly determines the scale. The results demonstrate that both of them are necessary components in HIPPo Dreamer.

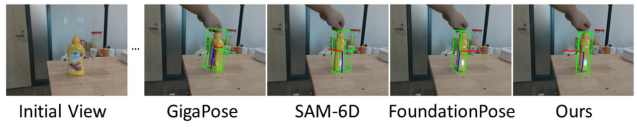

Fig. 10. A qualitative comparison with the SOTA methods [1], [2], [3].

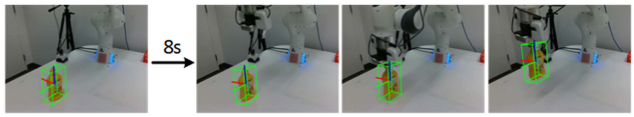

Fig. 11. An illustration of HIPPo's immediate robotic application.

D. Demonstration of Robotic Application

First, we present a qualitative comparison with the SOTA competitors [1], [2], [3] in the following scenario: the robot needs to instantly estimate the 6D pose of a novel object, while movement around the object to observe it is not allowed. Therefore, the only available information is the first RGB-D frame of the object. The SOTA competitors have to utilize the incomplete model of the object built from this single frame. In contrast, HIPPo can instantly construct the complete 3D model by harnessing image-to-3D priors and carry out this task. The qualitative comparison is shown in Fig. 10. Then, we present an application demo in Fig. 11. The task is to grasp a novel object while also obtaining its 3D oriented bounding box for further planning. The frames are captured by a static calibrated RGB-D camera. If we employ FoundationPose [1] for this task, we have to reconstruct [5] the object first, which could take several minutes. In comparison, by leveraging HIPPo, the robot only needs to wait a few seconds to execute the task. Moreover, thanks to the robust zero-shot prediction ability of the proposed HIPPo Dreamer, as shown in Fig. 12, our method can estimate the 6D pose of novel real-world objects.

E. Limitations

Most image-to-3D methods struggle with severe object occlusion because they only learn in an image-to-image manner [13]. As seen in Fig. 13, occlusion hinders the generation of a 3D mesh with correct scale and geometry, and thus affects the pose estimation result. A possible solution to this problem is to apply a vision-language model as in [13]. Efficiency is prioritized in HIPPo, so 3D colored point clouds are used for mesh updates. Consequently, when observations are sufficient and time consumption is not considered, its reconstruction performance is inferior to the method [5] focusing on rendering quality. Our work currently focuses only on initializing from a single image, rather than from few-shot unposed images, for which few-shot-to-3D generation methods [28] could be a solution.

V. CONCLUSION

We propose a new HIPPo framework that harnesses image-to-3D priors for model-free zero-shot 6D pose estimation. We design a novel instant image-to-mesh strategy, called HIPPo Dreamer, to generate a 3D mesh from the first glance of the object

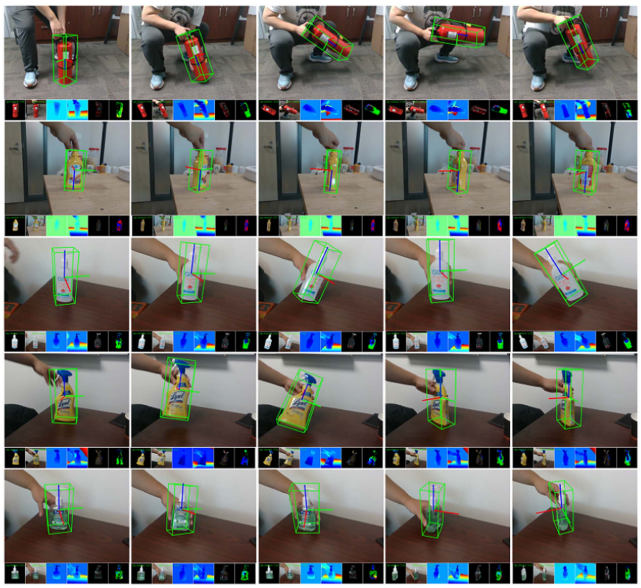

Fig. 12. Demonstration of HIPPo's zero-shot pose estimation ability. From top to bottom: Fire Extinguisher, Great Value Mustard, Germs Be Gone Sanitizer, Lysol Wipes, and Purell Sanitizer. At the bottom of each image, six sub-images are displayed: rendered RGB, measured RGB, rendered depth, measured depth, RGB residual, and depth residual.

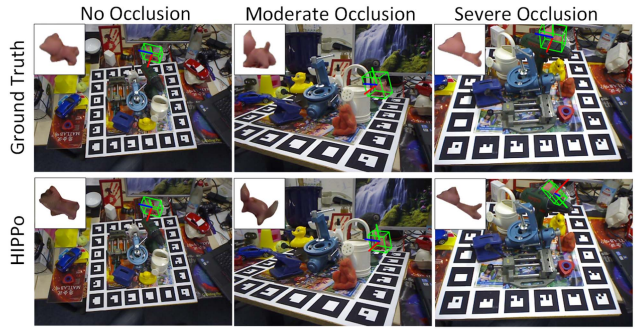

Fig. 13. An illustration of the effect of object occlusion on 3D mesh generation and 6D pose estimation. The sub-images in the top-left show object observations and the corresponding generated meshes in the top and bottom rows, respectively.

in mere seconds. In addition, we develop a measurement-guided formulation that gradually updates the diffusion prior with more reliable online measurements of geometry and appearance. HIPPo needs no 3D textured model or reference images in advance and maintains a complete mesh from the first glance. Qualitative and quantitative evaluations on various benchmarks show that the proposed approach outperforms the state-of-the-art 6D pose estimation methods when prior reference images are scarce.

REFERENCES

[1] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6 D pose estimation and tracking of novel objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17868–17879.

[2] J. Lin, L. Liu, D. Lu, and K. Jia, "SAM-6 D: Segment anything model meets zero-shot 6 D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27906–27916.

[3] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "GigaPose: Fast and robust novel object pose estimation via one correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9903–9913.

[4] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "MID-fusion: Octree-based object-level multi-instance dynamic SLAM," in *Proc. 2019 Int. Conf. Robot. Automat.*, 2019, pp. 5231–5237.

[5] B. Wen et al., "BundleSDF: Neural 6-DoF tracking and 3 D reconstruction of unknown objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 606–617.

[6] E. P. Örnek et al., "FoundPose: Unseen object pose estimation with foundation features," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 163–182.

[7] P. Ausserlechner, D. Haberger, S. Thalhammer, J.-B. Weibel, and M. Vincze, "ZS6D: Zero-shot 6 D object pose estimation using vision transformers," in *Proc. 2024 IEEE Int. Conf. Robot. Automat.*, 2024, pp. 463–469.

[8] Y. Labbé et al., "MegaPose: 6 D pose estimation of novel objects via render & compare," in *Proc. 6th Conf. Robot Learn.*, 2022, pp. 715–725.

[9] X. He, J. Sun, Y. Wang, D. Huang, H. Bao, and X. Zhou, "OnePose++: Keypoint-free one-shot object pose estimation without cad models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022,vol. 35, pp. 35103–35115.

[10] J. Sun et al., "OnePose: One-shot object pose estimation without cad models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6825–6834.

[11] X. Long et al., "Wonder3D: Single image to 3 D using cross-domain diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9970–9980.

[12] J. Xu, W. Cheng, Y. Gao, X. Wang, S. Gao, and Y. Shan, "InstantMesh: Efficient 3 D mesh generation from a single image with sparse-view large reconstruction models," 2024, *arXiv:2404.07191*.

[13] Y. Liu et al., "VQA-Diff: Exploiting VQA and diffusion for zero-shot image-to-3 D vehicle asset generation in autonomous driving," in *Proc. Eur. Conf. Comput. Vis.*, Springer, Nov. 2024, pp. 323–340.

[14] M. Deitke et al., "Objaverse: A universe of annotated 3 D objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 13142–13153.

[15] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3 D with MAST3R," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 71–91.

[16] Y. Liu, J. Shan, A. Haridevan, and S. Zhang, "L-PR: Exploiting LiDAR fiducial marker for unordered low overlap multiview point cloud registration," *IEEE Trans. Instrum. Meas.*, vol. 74, 2025, Art. no. 5011114.

[17] F. Di Felice et al., "Zero123-6 D: Zero-shot novel view synthesis for RGB category-level 6 D pose estimation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2024, pp. 14204–14211.

[18] Y. Liu et al., "MV-DeepSDF: Implicit modeling with multi-sweep point clouds for 3 D vehicle reconstruction in autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 8306–8316.

[19] Z. Yang et al., "Learning effective NeRFs and SDFs representations with 3 D GANs for object generation," in *Proc. NeurIPS Workshop Symmetry Geometry Neural Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=3NH6UMAqRH

[20] Y. Liu, J. Shan, and H. Schofield, "Improvements to thin-sheet 3 D LiDAR fiducial tag localization," *IEEE Access*, vol. 12, pp. 124907–124914, 2024.

[21] S. Liu et al., "Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection," in *Proc. Eur. Conf. Comput. Vis.*, Springer, 2024, pp. 38–55.

[22] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[23] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robot.: Sci. Syst.*, 2018, doi: 10.15607/RSS.2018.XIV.019.

[24] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D object pose estimation using 3 D object coordinates," in *Proc. 13th Eur. Conf., Comput. Vis.*, Springer, Zurich, Switzerland, Sep. 2014, pp. 536–551.

[25] T. Hodan, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. Zabulis, "T-LESS: An RGB-D dataset for 6 D pose estimation of texture-less objects," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2017, pp. 880–888.

[26] T. Hodan et al., "BOP: Benchmark for 6 D object pose estimation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 19–34.

[27] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *Proc. Int. Conf. Adv. Robot.*, 2015, pp. 510–517.

[28] H. Jiang, Z. Jiang, Y. Zhao, and Q. Huang, "LEAP: Liberate sparse-view 3 D modeling from camera poses," in *Proc. Int. Conf. Learn. Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=KPmajBxEaF