

Any6D: Model-free 6D Pose Estimation of Novel Objects

Taeyeop Lee¹ Bowen Wen² Minjun Kang¹ Gyuree Kang¹ In So Kweon¹ Kuk-Jin Yoon¹

¹KAIST ²NVIDIA

Abstract

We introduce Any6D, a model-free framework for 6D object pose estimation that requires only a single RGB-D anchor image to estimate both the 6D pose and size of unknown objects in novel scenes. Unlike existing methods that rely on textured 3D models or multiple viewpoints, Any6D leverages a joint object alignment process to enhance 2D-3D alignment and metric scale estimation for improved pose accuracy. Our approach integrates a render-and-compare strategy to generate and refine pose hypotheses, enabling robust performance in scenarios with occlusions, non-overlapping views, diverse lighting conditions, and large cross-environment variations. We evaluate our method on five challenging datasets: REAL275, Toyota-Light, HO3D, YCBINEOAT, and LM-O, demonstrating its effectiveness in significantly outperforming state-of-the-art methods for novel object pose estimation. Project page: <https://taeyeop.com/any6d>

1. Introduction

Object 6D pose estimation is a crucial problem in computer vision and robotics, focusing on determining the rigid 6D transformation, comprising both 3D orientation and 3D translation, between reference and camera coordinates. This task has numerous practical applications, such as robotic manipulation [4, 13, 64, 69, 73, 78] and augmented reality [43, 44, 55]. In recent years, significant progress has been made in this field [3, 13, 20, 33], and research is ongoing.

Research in 6D object pose estimation can be broadly categorized into three approaches: instance-level [22, 51, 62, 68], category-level [26–28, 32, 61, 65], and category-agnostic methods [47, 48, 71, 77]. Instance-level methods provide high precision but come with significant limitations. They rely on exact RGB-textured CAD models for estimating object poses, which restricts their effectiveness to only those objects seen during training. A major drawback is their inability to handle new objects without additional fine-tuning. Category-level methods partially

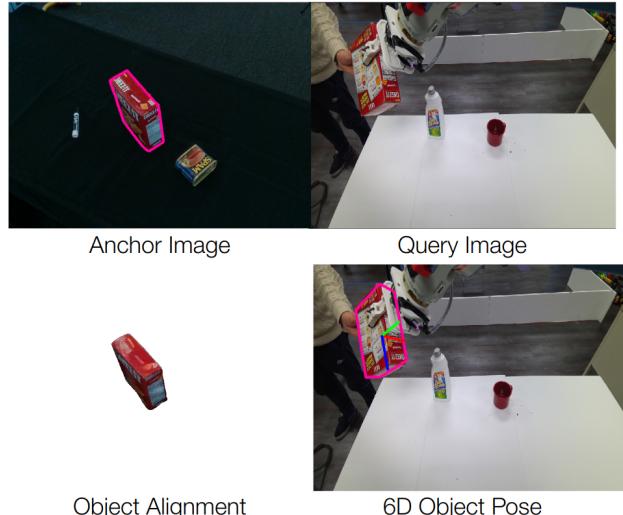


Figure 1. Our method accurately estimates 6D object pose for novel objects on drastically different scenes and viewpoints using only a single RGB anchor image. We achieve robust pose estimation without requiring precise CAD models or posed multi-view reference images.

mitigate these limitations by using category-specific prior knowledge, but they are still restricted to predefined object categories. Moreover, they face considerable challenges in acquiring comprehensive training datasets due to the complexity of aligning canonical poses. In contrast, category-agnostic methods aim to generalize across different objects without being limited to specific instances or categories, offering a more flexible approach to 6D pose estimation.

Recent research has shifted toward category-agnostic approaches [25, 30, 47, 47, 52, 58, 70, 71] to address the limitations of both category-level and instance-level pose estimation. These efforts can be broadly divided into two directions: model-based methods [30, 47, 52], which require textured RGB 3D CAD models at test time, and model-free methods [16, 39, 59, 77], which utilize multiview reference images or video sequences of the target object during inference. Although both approaches show promising results, they still face significant practical limitations when dealing with unseen objects that are not physically accessible. In robotic manipulation scenarios, for example, these methods

face difficulties when a robot encounters unexpected objects in a new environment without available 3D models or multiview images. This dependency on prior object data significantly limits their effectiveness in real-world applications.

To address these challenges, Oryon [11] introduces a novel model-free approach for estimating the 6D pose of objects from a single reference RGBD image. Unlike earlier methods that rely heavily on detailed object data at test time, Oryon uses language guidance to perform pose estimation with just a single RGBD reference. This method demonstrates impressive adaptability, functioning effectively even when reference and target images come from entirely different environments. Oryon establishes correspondences between images using language cues and estimates the pose by aligning visible point clouds from the reference. However, its performance degrades when occlusions or minimal overlap object regions between the reference and target objects, limiting the number of matching points. Consequently, Oryon struggles in object manipulation scenarios involving human or robotic arms, particularly when targets are occluded, appear in non-overlapping views, or lack sufficient texture [15, 67], as shown in our experiments.

To overcome these limitations, we present Any6DPose, a novel model-free approach for object pose estimation using a single anchor RGBD image. Inspired by recent advances in image-to-3D models [21, 36, 41, 72, 74, 76], our method estimates metric scale object shape for object pose estimation. Although existing 3D generation methods achieved promising photorealistic consistency between the input image and the generated 3D shape in normalized space, they often neglect critical aspects of 2D-3D alignment, especially the metric scale, which is crucial for accurate pose estimation and the subsequent downstream tasks. Therefore, we introduce a simple yet effective object alignment, jointly improving object size and pose estimation by alignment in 2D and 3D space. We generate multiple pose hypotheses and use a render and compare strategy to select the optimal one, building on previous work [71]. This enables our method to effectively handle disjoint views between query and anchor objects, as well as occlusions, diverse light conditions, and large cross-environment variations. We validate our approach across diverse scenarios on five public datasets: HO3D, YCBINEOAT, REAL275, Toyota-Light, and LM-O. Our Any6DPose significantly outperforms state-of-the-art methods in novel object pose estimation.

The main contributions of our work are as follows:

- We introduce Any6DPose, a novel framework that enables 6D pose and size estimation of novel objects in different scenes from only a single reference image.
- We propose a straightforward yet effective object alignment technique that addresses the challenges of existing 3D generation models, specifically improving 2D-3D

alignment and size estimation for accurate pose estimation.

- We validate our approach through extensive experiments, demonstrating superior performance compared to state-of-the-art methods across five benchmark datasets.

2. Related Works

2.1. CAD Model-based Object Pose Estimation

Instance-level pose estimation methods [17, 18, 24, 50, 64] rely on textured CAD models of specific objects, with training and testing conducted on the same instances. Category-level methods [8, 10, 31, 65, 79] generalize to novel instances within known categories but require expensive category annotations and remain constrained to predefined object classes. Category-agnostic methods [6, 9, 30] address these limitations by estimating the poses of arbitrary novel objects without category constraints. However, most still utilize ground-truth CAD models during inference, restricting their practical application in real-world scenarios where such models are unavailable. Some recent works [35, 54] try to retrieve CAD from existing databases.

2.2. Model-Free Object Pose Estimation

To overcome the limitations of CAD dependencies, model-free approaches have been developed that eliminate the need for explicit textured models by relying instead on a set of reference images of the target object. Gen6D [39], OnePose [60], and OnePose++ [16] generate 3D point clouds from videos or multiple views using structure-from-motion (SfM) [57] and utilize these point clouds for pose estimation through 2D-3D matching [23, 29]. FS6D [77] and FoundationPose [71] extend this approach to RGB-D images, achieving promising results. However, these methods still depend on multiview images, with most requiring camera poses to merge these views, which is often impractical in real-world scenarios.

Motivated by these challenges, recent work has focused on reducing these dependencies. For example, NOPE [46] estimates relative orientation using only a single anchor image. LoFTR [59] also contributes by using a transformer-based approach for feature matching, effectively enhancing the robustness of pose estimation from a single image. Similarly, Oryon [11] reduces the requirement to a single image by incorporating language guidance. However, these partial matching methods [11, 12, 34, 38, 53, 59] struggle in scenarios with non-overlap regions or significant occlusions in the target image. GigaPose [47], closely related to our approach, utilizes image-to-3D (Wonder3D [41]) to estimate object pose but still requires an initial novel object pose to determine object size. Concurrently, HIPPO [40], OmniManip [49], and SceneComplete [1] explore model-free object pose estimation from an image using image-to-3D methods.

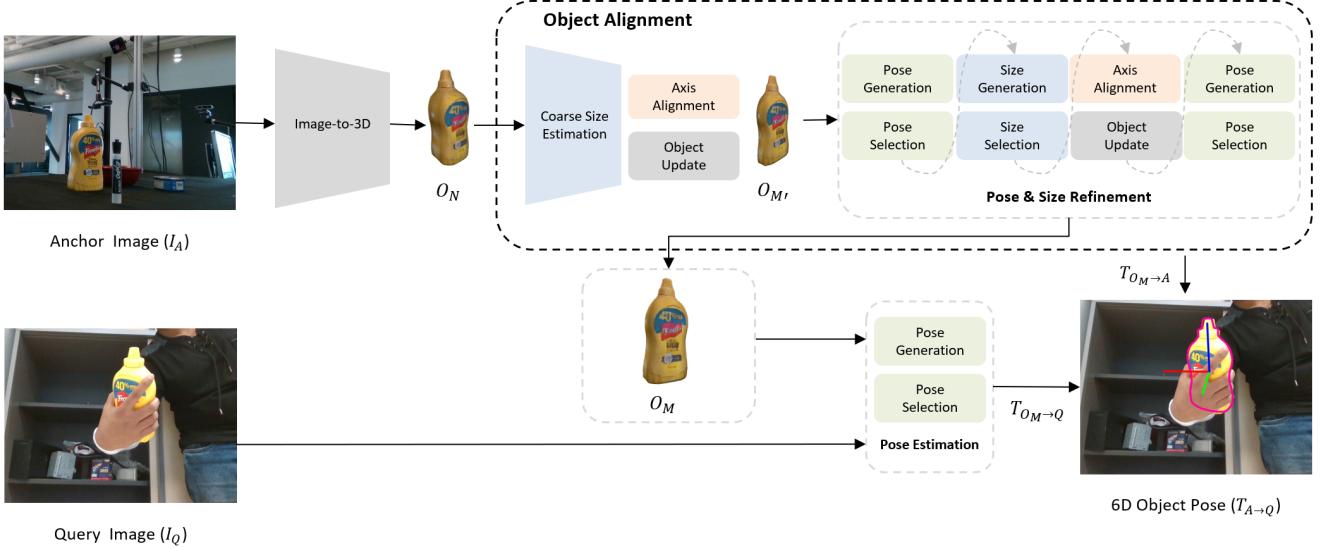


Figure 2. Overview of the Any6D framework for model-free object pose estimation. First, we reconstruct normalized object shape O_N from the image-to-3D model. Then, we estimate accurate object pose and size from anchor image I_A using the proposed object alignment (Sec. 3.1). Next, we use the query image I_Q to estimate the pose with the reconstructed metric-scale object shape O_M (Sec. 3.2).

3. Method

Given an RGB-D anchor image (I_A) and an RGB-D query image (I_Q), our task is to estimate the relative pose between them. The query images may capture the same object from drastically different viewpoints and scenes from the single anchor image. We formulate the problem as a relative pose estimation task [2, 11, 53, 59]. Our method aims to estimate the relative 6D pose $\mathbf{T}_{A \rightarrow Q} \in \text{SE}(3)$ between I_A and I_Q , where $\mathbf{T}_{A \rightarrow Q}$ is defined as the rigid transformation $[R \mid t]$, consisting of a rotation $R \in \text{SO}(3)$ and a translation $t \in \mathbb{R}^3$.

Previous approaches have attempted matching using either visible RGB images [56, 59] or point clouds [2, 53]. While effective with significant overlap, these methods struggle in scenarios with occlusions or large viewpoint variations, as partial-to-partial matching lacks sufficient shared features. To address these challenges, our method adopts a full-to-partial matching strategy by reconstructing a complete 3D object shape, ensuring robust alignment even under little overlap. Accurate pose estimation is achieved through a render-and-compare pipeline [71], effectively handling occluded or partially visible objects. The benefits of our approach are confirmed through extensive experiments (Sec. 4.3), demonstrating superior performance in low-overlap scenarios.

We introduce Any6D, a framework for estimating the relative pose $\mathbf{T}_{A \rightarrow Q}$ between an anchor image I_A and a query image I_Q . Our framework comprises two components: first, we reconstruct the normalized object shape O_N from the anchor image using an image-to-3D model, without considering real-world scale or pose, and then estimate the metric-scale object shape O_M by determining both the

actual object size $s \in \mathbb{R}^3$ and the pose $T_{O_M \rightarrow A}$, aligning it properly in 2D and 3D space (Sec. 3.1). Next, we use the reconstructed metric-scale object shape and the query image to estimate the pose, deriving the relative transformation $\mathbf{T}_{A \rightarrow Q}$ by combining $T_{O_M \rightarrow A}$ and $T_{O_M \rightarrow Q}$ (Sec. 3.2). The complete workflow of our Any6D framework is illustrated in Fig. 2.

3.1. Coarse Object Alignment

To our knowledge, there is no reliable existing solution for single-view metric-scale reconstruction from RGB-D that can handle diverse objects effectively. Given the recent advancements in RGB-based single-view reconstruction [21, 37, 76], we thus resort to InstantMesh [76] which has shown promising results across various objects. One key limitation, however, is the 3D object reconstruction only yields a shape with a normalized scale O_N , in the range [-1, 1] for each XYZ axis, meaning the resulting meshes are not properly scaled or positioned relative to the actual scene. This limitation prevents us from obtaining accurate pose alignment further, thus motivating our object alignment step, where we first estimate a coarse size of the object shape O_M and then refine this size by jointly solving accurate pose $T_{O_M \rightarrow A}$. Our approach involves estimating and aligning object shapes in both 3D and 2D spaces between I_A and O_N , including $\mathbf{T}_a \in \text{SE}(3)$, and size $s \in \mathbb{R}^3$.

Specifically, we estimate the object size s in a coarse-to-fine manner using I_A . We first initialize the coarse object size by comparing point clouds between I_A and O_N from their respective object centers. While the mean of points is a straightforward approach for center estimation, it becomes unreliable due to partial viewpoint and noisy outlier

points in the anchor image, as shown in Fig. 3-(a, b). Using a simple axis-aligned bounding box also leads to inaccurate center estimation due to the partial observability, as shown in Fig. 3-(c). Therefore, we propose using an oriented bounding box to determine the object center, as shown in Fig. 3-(d), which yields a more reliable coarse center estimate for I_A . For the axis alignment, we align the oriented bounding box with the XYZ axis. We then sample various rotation angles and calculate the IoU between the rotated bounding boxes of I_A and O_N across different angles. The combination of rotation and scale that leads to the highest IoU is used to transform O_N into a coarsely aligned object shape, updating it to an initial object shape $O_{M'}$, which is subsequently used for accurate pose and size estimation.

3.2. Fine Object Alignment

Our framework aims to determine the relative pose $\mathbf{T}_{A \rightarrow Q}$ between an anchor image I_A and a query image I_Q . This process involves multiple steps, starting with the reconstruction of a coarse object shape $O_{M'}$ and subsequently refining it to obtain the final metric-scale object shape O_M . Given the coarsely scaled initial shape $O_{M'}$, we jointly refine both the pose and the object size through our object and size refinement.

To refine the object shape and pose, we draw inspiration from FoundationPose [71] for its effective pose generation, refinement, and selection capabilities. However, it either requires ground-truth metric-scale object CAD for its model-based setup or multiple posed reference images in its model-free setup. This prevents its direct application in our considered setup, where only a single RGBD reference image is provided. We thus develop a joint module that injects the size estimation task into its pose refinement process. This allows us to estimate both the metric-scale size and pose reliably simultaneously.

Our refinement pipeline involves three primary modules: pose estimation, size estimation, and axis alignment—all of which work together in a unified process by alternating between the task of size refinement and pose refinement. We begin by estimating an initial pose using $O_{M'}$ and simultaneously refining the object size. In FoundationPose [71], sampling for pose hypothesis generation was only performed in $SO(3)$ without considering size variations. Instead, we additionally sample different sizes, together with $SO(3)$ sampling. In particular, the size samples are drawn in the range of $\Delta s \in [s_0, s_1]$ (we empirically set $s_0 = 0.6, s_1 = 1.4$) along each axis. We then refine the sampled pose hypothesis using the refinement module provided by [71] and render them to compare with the query image observation. The optimal pose hypothesis is selected based on the comparison score indicated by the pose selection module in [71]. Once the optimal size is determined, we scale the object shape and switch to the pose refinement

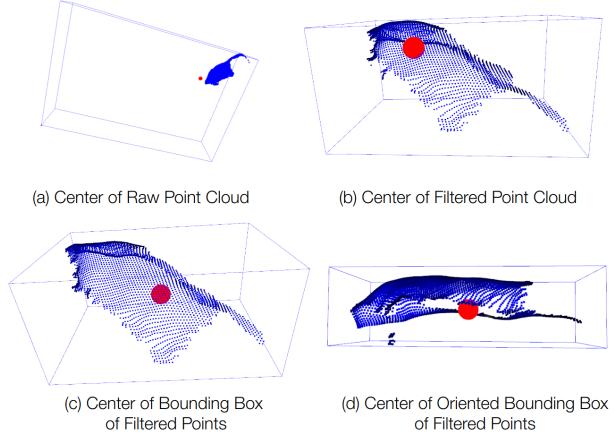


Figure 3. Visualization of each point clouds and center of mustard object.

stage, including an axis alignment step for more precise accuracy. This updated alignment leverages the joint estimation of size and pose, leading to greater accuracy than traditional IoU-based methods.

After refining the object parameters, we determine the final object pose $\mathbf{T}_{O_M \rightarrow A}$, which provides an accurate alignment of the reconstructed object shape. With the anchor image I_A , query image I_Q , and metric-scale object shape O_M , we estimate the relative pose $\mathbf{T}_{A \rightarrow Q}$ by composing two transformations: from the object to the anchor ($\mathbf{T}_{O_M \rightarrow A}$) and from the object to the query ($\mathbf{T}_{O_M \rightarrow Q}$). The relative pose can be expressed as follows:

$$\mathbf{T}_{A \rightarrow Q} = (\mathbf{T}_{O_M \rightarrow A})^{-1} \cdot \mathbf{T}_{O_M \rightarrow Q} \quad (1)$$

For pose selection, we employ a two-level render-and-compare strategy. Initially, a pose ranking network evaluates each hypothesis by comparing its rendered view to the cropped observation, producing an embedding to quantify alignment quality. Then, we apply self-attention to the concatenated embeddings of all hypotheses, incorporating global context to generate final scores for selecting the optimal pose.

4. Experiments

4.1. Datasets

We evaluate our method on five diverse real-world datasets: HO3D [15], YCBInEOTA [67], Toyota-Light [19], REAL275 [65], and LM-O [5]. Each dataset presents unique challenges in object pose estimation under various interaction scenarios and environmental conditions.

HO3D [15] captures close-range RGB-D images of human-hand object interactions. Using the latest HO3D-V3 version, we evaluate our method on 4 objects across 13 video sequences, totaling 2K images. The dataset presents large variations in viewpoints and significant hand-based occlusions. We sample pair images from the DexYCB [7] dataset.

Table 1. Model-free pose estimation results measured by AUC of ADD, and ADD-S, AR on HO3D dataset.

| Modality Metrics | Oryon [11] | | | LoFTR [59] | | | Gedi [53] | | | Ours | | |
|---------------------|------------------|-----|-----|------------|------|------|-----------|------|------|-------------|-------------|-------------|
| | RGB-D & Language | | | RGB-D | | | Depth | | | RGB-D | | |
| | ADD-S | ADD | AR | ADD-S | ADD | AR | ADD-S | ADD | AR | ADD-S | ADD | AR |
| AP10 | 23.8 | 0.0 | 0.4 | 22.5 | 0.0 | 1.2 | 94.4 | 1.9 | 3.5 | 100.0 | 16.2 | 22.2 |
| AP11 | 25.6 | 0.0 | 1.3 | 59.4 | 15.6 | 14.8 | 100.0 | 55.0 | 32.3 | 100.0 | 73.8 | 59.0 |
| AP12 | 21.2 | 0.0 | 1.4 | 12.5 | 1.2 | 2.1 | 99.4 | 30.6 | 20.3 | 100.0 | 48.8 | 28.3 |
| AP13 | 26.2 | 0.0 | 0.6 | 31.9 | 1.9 | 1.9 | 100.0 | 13.1 | 8.8 | 100.0 | 74.4 | 45.0 |
| AP14 | 8.1 | 0.0 | 0.0 | 25.0 | 0.0 | 0.0 | 76.2 | 0.0 | 0.6 | 100.0 | 35.6 | 29.7 |
| SM1 | 24.7 | 0.0 | 1.1 | 52.8 | 3.4 | 1.9 | 82.0 | 0.0 | 1.6 | 86.5 | 34.8 | 27.8 |
| SB11 | 46.1 | 0.0 | 2.4 | 75.4 | 8.4 | 15.6 | 96.4 | 13.8 | 12.1 | 100.0 | 86.8 | 68.9 |
| SB13 | 29.9 | 0.0 | 4.2 | 33.5 | 0.0 | 1.9 | 98.2 | 11.4 | 9.4 | 99.4 | 64.1 | 54.6 |
| MPM10 | 8.3 | 0.0 | 0.2 | 13.4 | 0.0 | 0.3 | 29.9 | 0.0 | 3.1 | 98.7 | 26.8 | 31.3 |
| MPM11 | 33.8 | 0.0 | 0.1 | 26.1 | 0.0 | 0.0 | 35.0 | 0.0 | 0.6 | 100.0 | 3.2 | 32.4 |
| MPM12 | 17.2 | 0.0 | 0.1 | 5.1 | 0.0 | 0.0 | 42.0 | 0.0 | 0.4 | 100.0 | 1.3 | 23.5 |
| MPM13 | 24.2 | 0.0 | 0.4 | 10.2 | 0.0 | 0.3 | 45.2 | 0.0 | 1.3 | 100.0 | 15.9 | 30.9 |
| MPM14 | 9.6 | 0.0 | 1.4 | 15.9 | 0.0 | 2.0 | 35.7 | 0.0 | 1.5 | 98.7 | 43.9 | 44.0 |
| MEAN | 23.0 | 0.0 | 1.0 | 29.5 | 2.3 | 3.2 | 71.9 | 9.7 | 7.4 | 98.7 | 40.4 | 38.3 |

YCBInEOTA [67] features mid-range RGBD images of dual-arm robot manipulations. The dataset includes three types of interactions: single-arm pick-and-place, within-hand manipulation, and pick-and-place with handoff between arms. We evaluate our method on 5 objects across 9 videos, totaling 749 frames, with pair images sampled from DexYCB [7]. The dataset presents diverse viewpoints and robot arm occlusions.

Toyota-Light [19] focuses on single-object pose estimation under challenging lighting conditions. It features significant lighting variations between scenes, which present crucial challenges. Following Oryon[11], we evaluate our method on 2K image pairs.

REAL275 [65] consists of RGBD images captured across different scenes, featuring six object categories with three instances per category. The dataset exhibits limited viewpoint variations and includes scenarios with minor occlusions. Following Oryon[11], we evaluate our method on 2K image pairs selected from the original real-world test set.

LM-O [5] consists of RGBD images captured in cluttered scenes, featuring 12 distinct textureless objects. The dataset has viewpoint variations and realistic occlusion scenarios. We follow GigaPose[47] of input image, segmentation, and image-to-3d methods for comparison.

4.2. Metrics

Our primary evaluation metrics are based on the Average Distance of Model Points (ADD) [64, 75], which computes the mean distance between corresponding 3D model points under-predicted and ground truth poses. For asymmetric objects, we use ADD directly, while for symmetric objects, we employ ADD-S, which calculates the average distance to the closest model point. We report the Area Under the Curve (AUC) and the recall rate at a threshold of 0.1 times the object diameter [16, 77]. Additionally, we evaluate using the metrics established in the BOP challenge [20], including the Average Recall (AR) of Visual Surface Discrepancy (VSD), Maximum Symmetry-aware Surface Distance (MSSD), and Maximum Symmetry-aware Projection Distance (MSPD). These metrics provide complementary perspectives on pose accuracy by evaluating recalls over multiple thresholds.

YCBInEOTA [67] features mid-range RGBD images of dual-arm robot manipulations. The dataset includes three types of interactions: single-arm pick-and-place, within-hand manipulation, and pick-and-place with handoff between arms. We evaluate our method on 5 objects across 9 videos, totaling 749 frames, with pair images sampled from DexYCB [7]. The dataset presents diverse viewpoints and robot arm occlusions.

4.3. Comparison with State-of-the-art

The evaluation is conducted on 5 challenging datasets: HO3D, YCBInEOTA, Toyota-Light, REAL275, and LM-O, each presenting unique challenges for pose estimation. For evaluation, we align the relative pose $T_{A \rightarrow Q}$ with the object pose $T_{O \rightarrow Q}$ by multiplying it with $T_{O \rightarrow A}$, following the approach used in Oryon [11]. We use ground-truth segmentation masks to evaluate the HO3D, YCBInEOTA, Toyota-Light, and REAL275 datasets.

We compare our approach against several recent state-of-the-art methods, including Oryon [11], a multi-modal object pose estimation method, and single-image matching baselines such as LoFTR [59], Gedi [53], and ObjectMatch [14]. For LoFTR, we use matched point clouds combined with pose optimization [63] to estimate accurate object poses. ObjectMatch leverages SuperGlue [56] for match estimation, as outlined in Oryon [11]. These baselines cover a range of approaches to pose estimation, from traditional feature matching to recent learning-based methods.

Table 1 summarized the results on HO3D dataset. The experimental results on the HO3D dataset demonstrate that our proposed method significantly outperforms state-of-the-art approaches across all evaluation metrics. Specifically, our method achieves mean scores of 98.7%, 40.4%, and 38.3% for ADD-S, ADD, and AR metrics, respectively, substantially surpassing previous methods, including Oryon (4.1%, 0%, 0.2%), LoFTR (29.5%, 2.3%, 3.2%), and Gedi (71.9%, 9.7%, 7.4%). These results are particularly notable given our method’s consistent performance across object instances, even in challenging scenarios involving human-hand interactions and occlusions. The exceptional improve-

Table 2. Model-free pose estimation results measured by AUC of ADD, ADD-S, and AR on YCBINEOAT dataset.

| Modality Metrics | Oryon [11] | | | LoFTR [59] | | | Gedi [53] | | | Ours | | |
|-----------------------------|------------------|-----|-----|------------|------|------|-----------|------|------|-------------|-------------|-------------|
| | RGB-D & Language | | | RGB-D | | | Depth | | | RGB-D | | |
| | ADD-S | ADD | AR | ADD-S | ADD | AR | ADD-S | ADD | AR | ADD-S | ADD | AR |
| sugar_box1 | 44.0 | 0.0 | 1.1 | 47.3 | 0.0 | 0.1 | 95.6 | 0.0 | 1.7 | 96.7 | 14.3 | 11.3 |
| sugar_box_yalehand0 | 34.7 | 3.0 | 5.2 | 41.6 | 0.0 | 0.1 | 82.2 | 6.9 | 21.5 | 89.1 | 75.2 | 44.4 |
| mustard0 | 48.6 | 0.0 | 3.5 | 47.3 | 20.3 | 15.2 | 100 | 0.0 | 19.1 | 100 | 23 | 32.4 |
| mustard_easy_00_02 | 36.2 | 0.0 | 0.3 | 23.2 | 0.0 | 1.9 | 78.3 | 0.0 | 20.2 | 78.3 | 53.6 | 39.2 |
| bleach0 | 10.4 | 0.0 | 1.5 | 55.2 | 0.0 | 0.9 | 74.6 | 0.0 | 7.7 | 98.5 | 68.7 | 56 |
| bleach_hard_00_03_chaitanya | 24.4 | 6.7 | 6.1 | 60 | 15.6 | 18.7 | 66.7 | 62.2 | 35.5 | 73.3 | 51.1 | 37.7 |
| tomato_soup_can_yalehand0 | 32.8 | 0.0 | 4.5 | 10.7 | 0.0 | 6.8 | 60.3 | 0.0 | 7.8 | 70.2 | 0 | 14.1 |
| cracker_box_reorient | 13.2 | 0.0 | 0 | 26.3 | 0.0 | 0 | 97.4 | 0.0 | 1.8 | 100 | 60.5 | 44.2 |
| cracker_box_yalehand0 | 15.0 | 0.0 | 1.2 | 22.6 | 0.0 | 0.2 | 89.5 | 0.0 | 10.4 | 97.7 | 63.9 | 58.2 |
| MEAN | 28.8 | 1.1 | 2.6 | 37.1 | 4 | 4.9 | 82.7 | 7.7 | 14.0 | 89.3 | 45.6 | 37.5 |

Table 3. Model-free pose estimation results measured by AUC of ADD(-S), AR, MSSD, MSPD, and VSD on the Toyota-Light (TOYL) dataset.

| Method | ADD(-S) | AR | MSSD | MSPD | VSD |
|----------------|-------------|-------------|-------------|-------------|-------------|
| SIFT [42] | 14.1 | 30.3 | 39.6 | 44.1 | 7.3 |
| Obj. Mat. [14] | 5.4 | 9.8 | 13.0 | 14.0 | 2.4 |
| Oryon [11] | 22.9 | 34.1 | 42.9 | 45.5 | 13.9 |
| Ours | 32.2 | 43.3 | 55.8 | 58.4 | 15.8 |

Table 4. Model-free pose estimation results measured by AUC of ADD(-S), AR, MSSD, MSPD, and VSD on the REAL275 dataset.

| Method | ADD(-S) | AR | MSSD | MSPD | VSD |
|----------------|-------------|-------------|-------------|-------------|-------------|
| SIFT [42] | 16.4 | 34.1 | 37.9 | 48.0 | 16.5 |
| Obj. Mat. [14] | 13.4 | 26.0 | 31.7 | 30.8 | 15.5 |
| Oryon [11] | 34.9 | 46.5 | 50.9 | 56.7 | 32.1 |
| Ours | 53.5 | 51.0 | 56.5 | 65.3 | 31.1 |

ment in ADD-S metrics, approaching 100% in most cases, validates our method’s robustness and accurate pose estimation under dynamic conditions.

For the YCBInEOAT dataset (Table 2), our approach achieves superior performance with mean scores of 89.3, 45.6, and 37.5 for ADD-S, ADD, and AR, significantly surpassing the corresponding scores of Gedi, which are 82.7, 7.7, and 14.0. The substantial improvement in ADD metrics particularly highlights our method’s capability in precise pose estimation. These results validate our method’s effectiveness in challenging scenarios with occlusions and non-overlapping viewpoints, which are critical for real-world robotic applications.

On the Toyota-Light dataset (Table 3), our method demonstrates substantial improvements across all metrics. We achieve 32.2% in ADD(-S), 43.3% in AR, 55.8% in MSSD, and 58.4% in MSPD, consistently outperforming Oryon by significant margins (9.3%, 9.2%, 12.9%, and 12.9% respectively). Our approach shows particular robustness under varying lighting conditions, which often pose significant challenges for existing methods. The consistent performance across all metrics demonstrates the stability of our approach in handling different lighting scenarios.

On the REAL275 dataset (Table 4), our method demonstrates remarkable performance, achieving 53.5% in ADD(-S), 51.0% in AR, 56.5% in MSSD, and 65.3% in MSPD. These results significantly surpass previous methods, par-

Table 5. Model-free pose estimation results measured by AUC of AR, MSSD, MSPD, and VSD on the Linemod Occlusion (LM-O) dataset.

| Method | Segmentation | Image-to-3D | Metrics | | | |
|---------------|--------------|------------------|-------------|-------------|-------------|-------------|
| | | | AR | MSPD | MSSD | VSD |
| GigaPose [47] | CNOS [45] | Wonder3D [41] | 17.5 | 35.8 | 9.0 | 7.6 |
| Ours | CNOS [45] | Wonder3D [41] | 28.6 | 36.1 | 32.0 | 17.6 |
| Ours | CNOS [45] | InstantMesh [76] | 25.2 | 29.5 | 27.4 | 18.7 |

ticularly showing substantial improvements over Oryon in most metrics (improvements of 18.6%, 4.5%, 27.1%, and 33.8%, respectively). While maintaining competitive performance in VSD (31.1% versus Oryon’s 32.1%), our method excels in all other metrics, demonstrating superior generalization across various object categories and poses. This comprehensive evaluation validates the robustness and versatility of our approach in handling diverse real-world scenarios.

Finally, on the Linemod Occlusion (LM-O) dataset (Table 5), we compare the estimation of the pose with a 3D model predicted from a single image. We compare the same input images, segmentation [45], and image-to-3D [41] followed by GigaPose [47] for fair comparison. Our method outperforms GigaPose under the same settings. Note that Gigapose is an RGB-based method but requires the initial pose to determine object size, while our method estimates object size from a single RGB-D image.

These extensive experiments across multiple datasets demonstrate that our method consistently outperforms existing approaches, often by significant margins. The strong performance across different metrics and various challenging scenarios validates the effectiveness and reliability of our approach for real-world applications.

4.4. Qualitative Results

Fig. 4 shows qualitative results of our method compared to Oryon [11], LoFTR [59], and Gedi [53] on the HO3D dataset. Given the anchor image (left), we overlay the object contours in pink, along with the rendered object and rotation axes, to visualize the pose estimation results. The query images are selected to have no overlapping parts with the anchor images, creating a challenging scenario for pose estimation. While Oryon, LoFTR, and Gedi struggle under



Figure 4. Qualitative comparison of state-of-the-art methods on the HO3D Dataset. In this challenging scenario, the left anchor image shows only partially visible objects, while the query images are not visible due to occlusion or different viewing angles. This represents the most challenging case for matching. Gedi, being a depth-based method, shows ambiguity when dealing with RGB-based non-symmetric objects.

these conditions due to their reliance on partial anchor information, our method effectively reconstructs complete object shapes, allowing for robust pose estimation even when parts of the object are occluded or not visible in the anchor view.

The first example features a white cleanser object where only a portion is visible in the anchor image, with a significantly different viewpoint in the query image. Our method accurately estimates the pose, while others fail to handle the limited visibility of key features. In the second example with a SPAM can, the anchor image shows only the back view, while the query image captures the logo side partially occluded by a hand. Despite these challenging conditions, our method successfully estimates the correct pose, while competing methods fail to align due to their reliance on limited anchor information. The final example shows a blue pitcher where Gedi roughly estimates the center position but fails to capture the correct orientation. In contrast, our method successfully aligns rotational and translational components, closely matching the ground truth pose.

Fig. 5 demonstrates our pose estimation results in robotic manipulation scenarios on the YCBInEOAT dataset. We compare our method with Oryon, LoFTR, and Gedi across various examples of robot-object interaction. In the first example with a red cracker box, our method accurately aligns both position and orientation, while other methods struggle with pose estimation. Notably, the anchor image only shows the front cheese logo, yet our method successfully handles the back view in the query image. The second example involves a yellow mustard bottle manipulated by two dual robot arms. Our method maintains accurate pose estimation throughout the interaction, proving robust to occlusions from the gripper. In the final example with a yel-

low sugar box, our method achieves precise pose estimation despite minimal visible information in the anchor image, while competing methods fail to handle the challenging viewpoint variations. These qualitative results highlight our method handles challenging scenarios involving significant occlusions and viewpoint changes, consistently outperforming existing approaches in pose estimation accuracy.

4.5. Ablation Studies

In this section, we conduct ablation studies to evaluate our object shape and alignment estimation.

Object Shape. We evaluate different object reconstruction approaches for pose estimation, including a comparison with a partial view-based baseline. For this baseline, given a camera-to-object pose $T_{C \rightarrow O}$ and a reference RGB-D image I_r , we train an object-centric NeRF [70] model and use marching cubes [66] to extract a textured mesh. This baseline relies on ground-truth poses to train NeRF with aligned object coordinates, ensuring a fair comparison with the same pose estimation setup. It follows the single image version of FoundationPose [71]. As shown in Fig. 6, our reconstructed shapes differ significantly from the baseline, which exhibits missing regions in rear and diagonal views, leading to ambiguity in rendering and pose estimation. Table 6 presents the pose estimation results and mesh quality evaluation using Chamfer Distance (CD). The CD between our reconstructions and ground-truth meshes indicates that lower values correspond to better accuracy. Additionally, we found that improper axis alignment leads to distortions in the X, Y, and Z ratios, highlighting the importance of precise axis alignment.

Object Alignment. We evaluate our simple but effective

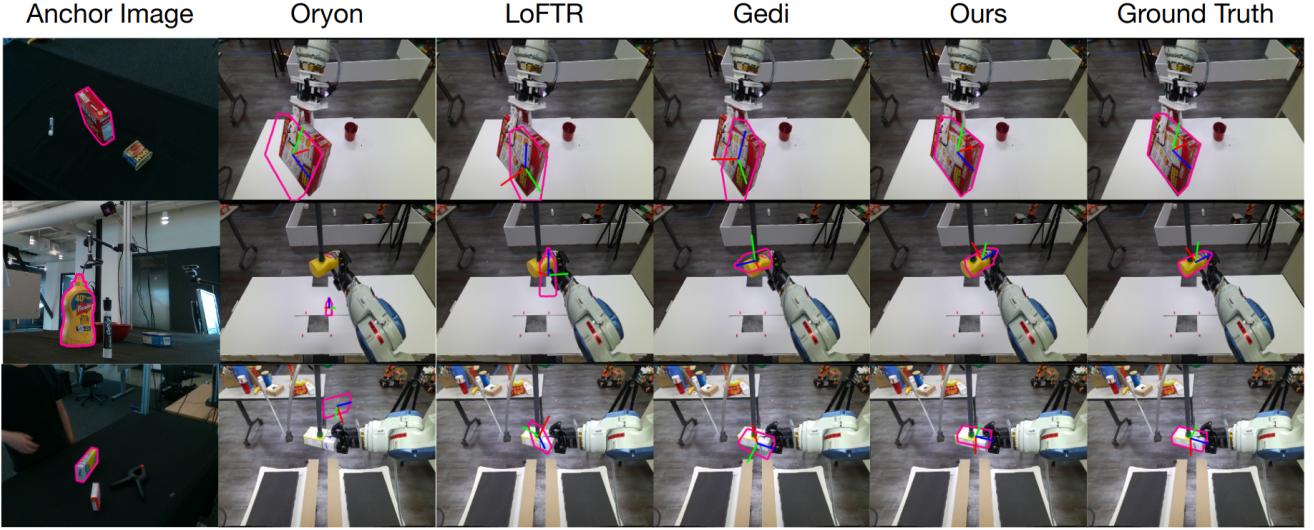


Figure 5. Qualitative comparison of state-of-the-art methods on the YCBInEOAT Dataset. In this challenging scenario, the left anchor image shows only partially visible objects, while the query images are not visible due to occlusion or different viewing angles. This represents the most challenging case for matching. Gedi, being a depth-based method, shows ambiguity when dealing with RGB-based non-symmetric objects.



Figure 6. Comparison of shape quality between baseline method and ours.

Table 6. Ablation Studies of Size Estimation on the HO3D dataset.

| Method | Object Alignment | | | Metrics | | | |
|----------|------------------------------------|------------------------------------|------------------------------------|----------------------|--------------------|-------------------|---------------------|
| | Coarse Size | Refinement | Axis Align | ADD-S (\uparrow) | ADD (\uparrow) | AR (\uparrow) | CD (\downarrow) |
| Baseline | X | X | X | 28.6 | 0.00 | 0.20 | 1.02 |
| (1) | X | X | X | 0.0 | 0.0 | 0.0 | 1.47 |
| (2) | X | ✓ | ✓ | 98.0 | 25.5 | 26.8 | 0.53 |
| (3) | ✓ | X | ✓ | 83.7 | 26.6 | 22.5 | 0.92 |
| (4) | ✓ | ✓ | X | 92.3 | 23.6 | 24.9 | 0.66 |
| Ours | ✓ | ✓ | ✓ | 98.7 | 40.4 | 38.3 | 0.49 |

object alignment module: coarse size estimation, size and pose refinement and axis alignment. The results of our ablation study on these components are presented in Table 6. First, as shown in Table 6-(1), using raw object shape estimation without any alignment steps leads to failure in accurately estimating the object’s rotation and translation, with all metrics showing subpar performance. This underscores the importance of size alignment as a foundational step in pose estimation. In Table 6-(2) and in our full method, incorporating coarse size estimation substantially improves the ADD metric. This shows that even a basic size esti-

mation allows the model to approximate the object’s pose better. Next, Table 6-(3) demonstrates that incorporating axis alignment further enhances performance, particularly on the ADD-S and AR metrics. This process not only improves object shape estimation but also yields significant gains in pose accuracy by aligning the object’s axes to avoid distortion in its proportions along the x, y, and z directions. Finally, our full method, incorporating coarse size estimation, refinement, and axis alignment, achieves the best results across all metrics. Specifically, as shown in Table 6, our method reaches an ADD of 40.4 and an AR of 38.3, outperforming the other configurations. These results validate the effectiveness of our alignment approach in enhancing object pose estimation.

5. Conclusion

We introduce Any6D, a novel framework for model-free object pose estimation that reduces dependence on CAD models and multi-view images, particularly in challenging object manipulation scenarios. Our method proposes an efficient object alignment method for precise pose and size estimation. Extensive experiments demonstrate that Any6D significantly outperforms state-of-the-art methods for occlusions and varying viewpoints. While our method shows promising results on pose and size estimation through image-to-3D alignment, it has limitations when the initial 3D shape is inaccurate, as our approach does not incorporate shape updating. A future direction would be to refine shape to enhance robustness and applicability in scenarios with inaccurate initial shapes.

Acknowledgment

This work was partly supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (NRF2022R1A2B5B03002636), the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068, and KAIST Cross-Generation Collaborative Lab Project.

References

- [1] Aditya Agarwal, Gaurav Singh, Bipasha Sen, Tomás Lozano-Pérez, and Leslie Pack Kaelbling. Scenecomplete: Open-world 3d scene completion in complex real world environments for robot manipulation. *arXiv preprint arXiv:2410.23643*, 2024. 2
- [2] Xuyang Bai, Zixin Luo, Lei Zhou, Hongkai Chen, Lei Li, Zeyu Hu, Hongbo Fu, and Chiew-Lan Tai. Pointdsc: Robust point cloud registration using deep spatial consistency. In *CVPR*, 2021. 3
- [3] Prithviraj Banerjee, Sindi Shkodrani, Pierre Moulon, Shreyas Hampali, Fan Zhang, Jade Fountain, Edward Miller, Selen Basol, Richard Newcombe, Robert Wang, et al. Introducing hot3d: An egocentric dataset for 3d hand and object tracking. *arXiv preprint arXiv:2406.09598*, 2024. 1
- [4] Valts Blukis, Taeyeop Lee, Jonathan Tremblay, Bowen Wen, In So Kweon, Kuk-Jin Yoon, Dieter Fox, and Stan Birchfield. One-shot neural fields for 3D object understanding. In *CVPRW*, 2023. 1
- [5] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014. 4, 5
- [6] Andrea Caraffa, Davide Boscaini, Amir Hamza, and Fabio Poiesi. Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In *ECCV*, 2024. 2
- [7] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 4, 5
- [8] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6D object pose and size estimation. In *CVPR*, 2020. 2
- [9] Jianqiu Chen, Zikun Zhou, Mingshan Sun, Rui Zhao, Liwei Wu, Tianpeng Bao, and Zhenyu He. Zeropose: Cad-prompted zero-shot object 6d pose estimation in cluttered scenes. *IEEE TCSVT*, 2024. 2
- [10] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *CVPR*, 2024. 2
- [11] Jaime Corsetti, Davide Boscaini, Changjae Oh, Andrea Cavallaro, and Fabio Poiesi. Open-vocabulary object 6d pose estimation. In *CVPR*, 2024. 2, 3, 5, 6
- [12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabovich. Superpoint: Self-supervised interest point detection and description. In *CVPRW*, 2018. 2
- [13] Guoguang Du, Kai Wang, Shiguo Lian, and Kaiyong Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 2021. 1
- [14] Can Gümeli, Angela Dai, and Matthias Nießner. Object-match: Robust registration using canonical object correspondences. In *CVPR*, 2023. 5, 6
- [15] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnote: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2, 4
- [16] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. OnePose++: Keypoint-free one-shot object pose estimation without CAD models. *NuerIPS*, 2022. 1, 2, 5
- [17] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation. In *CVPR*, 2020. 2
- [18] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. FFB6D: A full flow bidirectional fusion network for 6D pose estimation. In *CVPR*, 2021. 2
- [19] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018. 4, 5
- [20] Tomas Hodan, Martin Sundermeyer, Yann Labbe, Van Nguyen Nguyen, Gu Wang, Eric Brachmann, Bertram Drost, Vincent Lepetit, Carsten Rother, and Jiri Matas. Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In *CVPR*, 2024. 1, 5
- [21] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*, 2023. 2, 3
- [22] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In *ICCV*, 2017. 1
- [23] Laurent Kneip, Davide Scaramuzza, and Roland Siegwart. A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In *CVPR*, 2011. 2
- [24] Yann Labb , Justin Carpenter, Mathieu Aubry, and Josef Sivic. CosyPose: Consistent multi-view multi-object 6D pose estimation. In *ECCV*, 2020. 2
- [25] Yann Labb , Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpenter, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 1

- [26] Taeyeop Lee, Byeong-Uk Lee, Myungchul Kim, and In So Kweon. Category-level metric scale object shape and pose estimation. *IEEE RA-L*, 2021. 1
- [27] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. UDA-COPE: Unsupervised domain adaptation for category-level object pose estimation. In *CVPR*, 2022.
- [28] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. Tta-cope: Test-time adaptation for category-level object pose estimation. In *CVPR*, 2023. 1
- [29] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o (n) solution to the pnp problem. *IJCV*, 2009. 2
- [30] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *CVPR*, 2024. 1, 2
- [31] Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *CVPR*, 2024. 2
- [32] Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A. Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an rgb image. In *ICRA*, 2022. 1
- [33] Jian Liu, Wei Sun, Hui Yang, Zhiwen Zeng, Chongpei Liu, Jin Zheng, Xingyu Liu, Hossein Rahmani, Nicu Sebe, and Ajmal Mian. Deep learning-based object pose estimation: A comprehensive survey. *arXiv preprint arXiv:2405.07801*, 2024. 1
- [34] Jian Liu, Wei Sun, Kai Zeng, Jin Zheng, Hui Yang, Lin Wang, Hossein Rahmani, and Ajmal Mian. Novel object 6d pose estimation with a single reference view. *arXiv preprint arXiv:2503.05578*, 2025. 2
- [35] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *NuerIPS*, 2023. 2
- [36] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *NuerIPS*, 2024. 2
- [37] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *ICCV*, 2023. 3
- [38] Xingyu Liu, Gu Wang, Ruida Zhang, Chenyangguang Zhang, Federico Tombari, and Xiangyang Ji. Unopose: Unseen object pose estimation with an unposed rgb-d reference image. *arXiv preprint arXiv:2411.16106*, 2024. 2
- [39] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *ECCV*, 2022. 1, 2
- [40] Yibo Liu, Zhaodong Jiang, Binbin Xu, Guile Wu, Yuan Ren, Tongtong Cao, Bingbing Liu, Rui Heng Yang, Amir Rasouli, and Jinjun Shan. Hippo: Harnessing image-to-3d priors for model-free zero-shot 6d pose estimation. *arXiv preprint arXiv:2502.10606*, 2025. 2
- [41] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *CVPR*, 2024. 2, 6
- [42] David G Lowe. Object recognition from local scale-invariant features. In *CVPR*, 1999. 6
- [43] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: A hands-on survey. *IEEE TVCG*, 2015. 1
- [44] Eitan Marder-Eppstein. Project Tango. In *ACM SIGGRAPH Real-Time Live!*, 2016. 1
- [45] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponomatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *CVPR*, 2023. 6
- [46] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponomatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. Nope: Novel object pose estimation from a single image. In *CVPR*, 2024. 2
- [47] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *CVPR*, 2024. 1, 2, 5, 6
- [48] Evin Pinar Örnek, Yann Labb  , Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *ECCV*, 2024. 1
- [49] Mingjie Pan, Jiyao Zhang, Tianshu Wu, Yinghao Zhao, Wenzhong Gao, and Hao Dong. Omnimaniip: Towards general robotic manipulation via object-centric interaction primitives as spatial constraints. *arXiv preprint arXiv:2501.03841*, 2025. 2
- [50] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation. In *ICCV*, 2019. 2
- [51] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Junjun Bao. PVNet: Pixel-wise voting network for 6DoF pose estimation. In *CVPR*, 2019. 1
- [52] Evin Pinar Örnek, Yann Labb  , Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. *arXiv e-prints*, 2023. 1
- [53] Fabio Poiesi and Davide Boscaini. Learning general and distinctive 3d local deep descriptors for point cloud registration. *IEEE TPAMI*, 2022. 2, 3, 5, 6
- [54] Georgy Ponomatkin, Martin C  fka, Tom   Sou  ek, M  d  ric Fourmy, Yann Labb  , Vladimir Petrik, and Josef Sivic. 6D Object Pose Tracking in Internet Videos for Robotic Manipulation. In *ICLR*, 2025. 2
- [55] Martin Runz, Maud Buffier, and Lourdes Agapito. MaskFusion: Real-time recognition, tracking and reconstruction of multiple moving objects. In *ISMAR*, 2018. 1
- [56] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 3, 5
- [57] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 2

- [58] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *CVPR*, 2022. 1
- [59] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 1, 2, 3, 5, 6
- [60] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. In *CVPR*, 2022. 2
- [61] Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6D object pose and size estimation. In *ECCV*, 2020. 1
- [62] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. In *CoRL*, 2018. 1
- [63] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE TPAMI*, 1991. 5
- [64] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. DenseFusion: 6D object pose estimation by iterative dense fusion. In *CVPR*, 2019. 1, 2, 5
- [65] 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 *CVPR*, 2019. 1, 2, 4, 5
- [66] LORENSEN WE. Marching cubes: A high resolution 3d surface construction algorithm. *Computer graphics*, 1987. 7
- [67] Bowen Wen, Chaitanya Mitash, Baozhang Ren, and Kostas E Bekris. se (3)-tracknet: Data-driven 6d pose tracking by calibrating image residuals in synthetic domains. In *IROS*, 2020. 2, 4, 5
- [68] Bowen Wen, Chaitanya Mitash, Sruthi Soorian, Andrew Kimmel, Avishai Sintov, and Kostas E Bekris. Robust, occlusion-aware pose estimation for objects grasped by adaptive hands. In *ICRA*, 2020. 1
- [69] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. *RSS*, 2022. 1
- [70] 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 *CVPR*, 2023. 1, 7
- [71] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *CVPR*, 2024. 1, 2, 3, 4, 7
- [72] Hao Wen, Zehuan Huang, Yaohui Wang, Xinyuan Chen, Yu Qiao, and Lu Sheng. Ouroboros3d: Image-to-3d generation via 3d-aware recursive diffusion. *arXiv preprint arXiv:2406.03184*, 2024. 2
- [73] Jay M Wong, Vincent Kee, Tiffany Le, Syler Wagner, Gian-Luca Mariottini, Abraham Schneider, Lei Hamilton, Rahul Chipalkatty, Mitchell Hebert, David MS Johnson, et al. Segicp: Integrated deep semantic segmentation and pose estimation. In *IROS*, 2017. 1
- [74] Jianfeng Xiang, Zelong Lv, Sicheng Xu, Yu Deng, Ruicheng Wang, Bowen Zhang, Dong Chen, Xin Tong, and Jiaolong Yang. Structured 3d latents for scalable and versatile 3d generation. *arXiv preprint arXiv:2412.01506*, 2024. 2
- [75] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes. *RSS*, 2018. 5
- [76] Jiale Xu, Weihao Cheng, Yiming Gao, Xiantao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2, 3, 6
- [77] He Yisheng, Wang Yao, Fan Haoqiang, Chen Qifeng, and Sun Jian. Fs6d: Few-shot 6d pose estimation of novel objects. In *CVPR*, 2022. 1, 2, 5
- [78] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge. In *ICRA*, 2017. 1
- [79] Jiayao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *ECCV*, 2024. 2