Pos3R: 6D Pose Estimation for Unseen Objects Made Easy

Weijian Deng[1] Dylan Campbell[1] Chunyi Sun[1] Jiahao Zhang[1]
Shubham Kanitkar[2] Matthew E. Shaffer[2] Stephen Gould[1]
[1]Australian National University [2]RIOS Intelligent Machines
[1]{firstname.lastname}@anu.edu.au [2]{firstname.lastname}@rios.ai

Abstract

Foundation models have significantly reduced the need for task-specific training, while also enhancing generalizability. However, state-of-the-art 6D pose estimators either require further training with pose supervision or neglect advances obtainable from 3D foundation models. The latter is a missed opportunity, since these models are better equipped to predict 3D-consistent features, which are of significant utility for the pose estimation task. To address this gap, we propose Pos3R, a method for estimating the 6D pose of any object from a single RGB image, making extensive use of a 3D reconstruction foundation model and requiring no additional training. We identify template selection as a particular bottleneck for existing methods that is significantly alleviated by the use of a 3D model, which can more easily distinguish between template poses than a 2D model. Despite its simplicity, Pos3R achieves competitive performance on the Benchmark for 6D Object Pose Estimation (BOP), matching or surpassing existing refinement-free methods. Additionally, Pos3R integrates seamlessly with render-and-compare refinement techniques, demonstrating adaptability for high-precision applications.

1. Introduction

Six-dimensional (6D) object pose estimation—the task of determining the exact position and orientation of objects relative to a camera—is essential for applications in robotics, augmented reality, and autonomous systems. Reliable pose estimation enables critical tasks like object manipulation, grasping, and assembly, allowing these systems to interact effectively in complex and dynamic environments [6, 18, 29, 30, 49, 54]. Traditional approaches often rely on learning-based methods tailored to specific objects or categories, achieving high accuracy but struggling to generalize to new categories or unseen objects [4, 5, 7, 26, 33, 46, 47, 53–55, 59, 61]. This limitation is particularly problematic in dynamic, data-scarce environments where adaptability and flexibility are crucial.

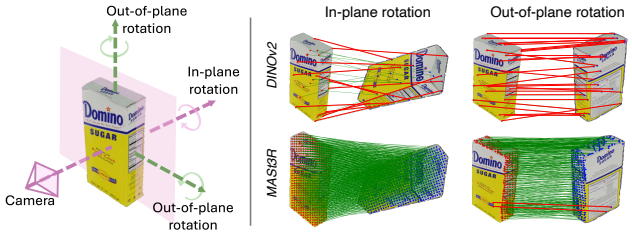

Figure 1. **Illustration of In-Plane and Out-of-Plane Rotations with Correspondence Quality between DINOv2 and MASt3R.** The left diagram illustrates in-plane rotations (within the image plane) and out-of-plane rotations (3D orientation changes) from a camera view. On the right, DINOv2 [40] (top row) and MASt3R [24] (bottom row) are compared in their handling of these rotations. DINOv2 exhibits sparse and inconsistent correspondences (shown by red lines), particularly under out-of-plane rotations, due to its 2D feature limitations. In contrast, MASt3R provides dense and stable correspondences across both types of rotations, reflecting its ability to produce 3D-consistent features.

To overcome these limitations, there has been a recent shift toward model-based approaches that aim to generalize pose estimation to unseen objects without object-specific training [2, 17, 41, 56]. These typically rely on a two-stage pipeline: detecting and localizing objects within a scene, followed by a render-and-compare process that matches detected object regions to a set of template models. Building on this concept, recent works have sought to improve generalization through large-scale model training with significant attention to object diversity. For example, FoundationPose [56] uses synthetic training data combined with a large language model and contrastive learning, enhancing feature alignment across varied domains for robust generalization.

The recent emergence of training-free methods offers a promising alternative, enabling 6D pose estimation for unseen objects without object- or task-specific training. Foundation models like DINOv2 [40] have demonstrated strong zero-shot capabilities, capturing spatial and semantic details through learned features [1, 58]. FoundPose [41], for instance, leverages DINOv2 descriptors to bridge synthetic

and real data, achieving competitive results on the BOP challenge [17]. These advancements underscore the potential of training-free frameworks in 6D pose estimation.

Building on this trend, we explore the strengths of 3D foundation models. Accurate pose estimation requires handling both in-plane and out-of-plane rotations, which are demonstrated in Fig. 1. In-plane rotations occur within the image plane, as when an object rotates around the camera's line of sight, while out-of-plane rotations involve changes in 3D orientation, such as tilting or turning, which significantly alters appearance due to perspective shifts. Addressing both types of rotation is essential for robust pose estimation, as objects often appear from varied angles. Although 2D foundation models like DINOv2 handle in-plane rotations well through training augmentations, they lack robustness to out-of-plane rotations due to their 2D limitations (shown in Fig. 1). In contrast, 3D foundation models, such as MASt3R [24], are specifically designed to produce 3D-consistent features across different viewpoints, enabling reliable feature alignment even with out-of-plane rotations.

Motivated by this, we propose Pos3R, a training-free method for 6D pose estimation of unseen objects using only RGB inputs. Pos3R leverages the 3D foundation model MASt3R [24] for pose estimation without requiring additional training. The core of Pos3R lies in the image matching process between test crops and rendered templates from a CAD model. While common matchers like LoFTR [50] often struggle with synthetic-to-real matching due to domain gaps, MASt3R produces high-quality 2D correspondences in this case. These reliable matches enable Pos3R to establish the 2D-3D correspondences required for pose estimation using the P*n*P-RANSAC algorithm.

By harnessing MASt3R's ability to produce dense correspondences robust to viewpoint and illumination changes, we simplify template generation and achieve computational efficiency. Unlike existing methods that render hundreds of templates, we place camera positions to cover both in-plane and out-of-plane rotations. Specifically, we use eight base templates per object, positioned at the vertices of a cube centered on the CAD model, to capture out-of-plane rotations. For each base template, we generate five in-plane rotational templates along the object's principal axis, resulting in forty templates per object. To ensure efficient and accurate pose estimation, Pos3R dynamically selects the optimal rotation based on matching quality. This simple yet effective strategy maintains computational efficiency while achieving reliable pose estimation. Tested on seven diverse datasets in the BOP challenge, Pos3R demonstrates strong performance as a robust, scalable option for training-free 6D pose estimation of unseen objects, adaptable to render-and-compare refinement techniques.

In summary, our contributions are:

- While existing methods rely on 2D foundation models for unseen object pose estimation, we present Pos3R, a training-free method that leverages the 3D foundation model MASt3R [24] to improve robustness.
- Leveraging MASt3R's robust dense correspondences, Pos3R uses only forty strategically placed templates per object to capture in-plane and out-of-plane rotations. A simple selection technique based on correspondence quality ensures accurate matching, achieving strong performance on the BOP challenge.

2. Related Work

Seen Object Pose Estimation. Instance-level object pose estimation refers to the task of estimating the poses of specific objects previously encountered during model training [5, 7, 26, 33, 46, 54, 61]. Among common approaches are correspondence-based methods, which learn to identify precise alignments between input data and CAD models of the objects [7, 27, 46]. Other methods use template-based strategies, where the model selects the most similar pose-labeled template from a predefined set of examples [25, 35, 52]. Additionally, regression-based approaches directly predict object poses from the learned object-specific features [12, 19, 26, 42].

These instance-specific techniques offer high precision but typically require retraining for new object instances, limiting their generalizability. This limitation has led to the exploration of category-level methods, which generalize within specific object categories and allow estimation of unseen objects in known categories [4, 47, 53, 55, 59]. For example, SecondPose [4] enhances category-level 6D pose estimation by integrating object-specific geometric features with DINOv2 SE(3)-consistent semantic priors, effectively addressing intra-class shape variation. However, these methods struggle outside their target categories, while our approach enables pose estimation without relying on such constraints, allowing broader applications.

Unseen Object Pose Estimation. To enhance pose estimation flexibility, many methods aim to generalize across new object instances without object-specific training [22, 31, 37, 39, 43]. Approaches for this task can be broadly divided into manual reference view-based and CAD model-based methods. (i) Reference view-based methods use multi-view images with known poses as reference data for pose estimation [14, 31, 43, 51]. For example, OnePose [14, 51] reconstructs 3D object point clouds from posed RGB images, establishing 2D-3D correspondences to solve 6D poses. In a different approach, Nope [38], adopting the perspective of generating new views, trains models to predict discriminative embeddings for novel object perspectives. (ii) CAD model-based approaches leverage 3D object model to facilitate pose estimation. Some methods match features between the CAD model and the query image [20, 28, 60],

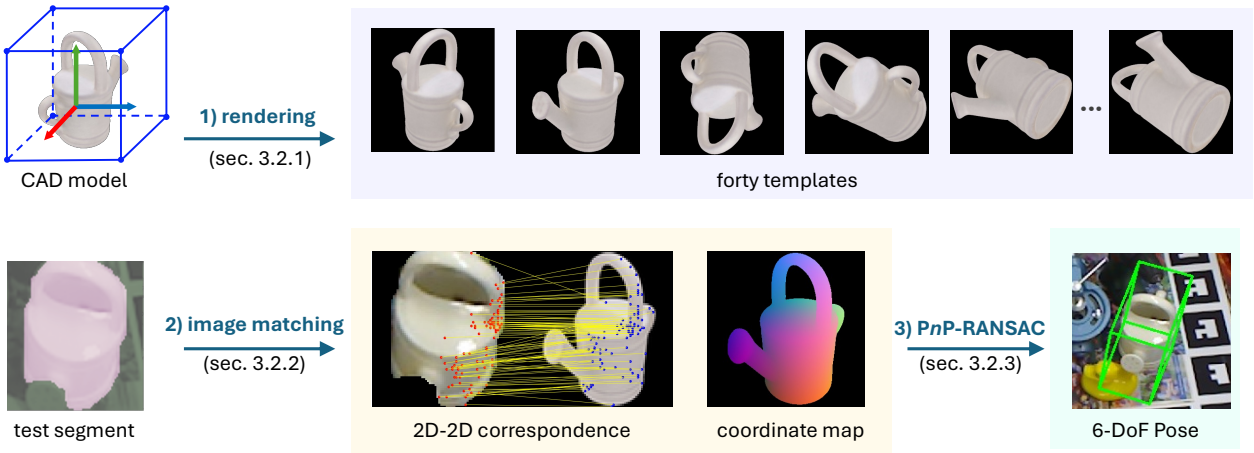

Figure 2. **Overview of the 6D Pose Estimation Process in Pos3R.** (1) *Template Rendering* (Sec. 3.2.1): A CAD model is used to generate a set of forty templates, each representing a unique orientation that covers both in-plane and out-of-plane rotations. This is achieved by positioning the camera at the vertices of a cube around the object. (2) *Image Matching* (Sec. 3.2.2): Given a test segment from CNOS [37], we establish 2D–2D correspondences between the test image and each rendered template. We leverage the 3D-consistent features generated by MASt3R [24] for accurate matching and select the best template based on matching quality. Each template also includes a 3D object coordinate map that records the corresponding 3D points. (3) *Pose Fitting* (Sec. 3.2.3): Using the selected correspondences, we apply the P*n*P-RANSAC algorithm [23] to obtain a final pose that aligns the object with its observed position in the scene.

while others employ template matching to find the closest initial pose, refining it further using specialized refiners for higher accuracy [22, 34, 39]. For example, GigaPose [39] introduces an efficient two-network system: one network for retrieving template views (out-of-plane rotation) and another for estimating the remaining degrees of freedom (in-plane rotation and 3D translation). This separation reduces computational costs, contrasting with methods like MegaPose [22], which applies a single network to every possible test crop-template pair. Our work builds on the CAD model-based approach, aligning with recent BOP challenge protocols [17]. We focus on a training-free pipeline for unseen object pose estimation, a direction distinct from methods that require extensive training of task-specific networks.

Training-Free Object Pose Estimation. Traditionally, 6D object pose estimation is achieved by establishing 3D-to-2D correspondences followed by a P*n*P algorithm [11, 13, 44, 61]. Recently, leveraging features from foundation models has gained traction, especially with models like DINO, which offer spatial detail and semantic consistency for pose estimation [1, 58]. FoundPose [41] uses DINOv2 descriptors to bridge synthetic and real data domains, providing strong performance for symmetric objects with RGB-only inputs. FreeZe integrates 3D point descriptors from GeDi [45] with image features from DINOv2 [40] for RGB-D pose estimation. By leveraging the 3D consistent feature of MASt3R to produce dense correspondences, our work explores training-free, unseen object pose estimation with RGB-only inputs, enabling effective pose estimation across diverse object categories without task-specific training.

3. 6D Pose Estimation with Pos3R

3.1. Task Definition

Given a 3D model of a query object $\mathbf{Q}$ and an RGB image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$ containing $\mathbf{Q}$, the task is to estimate the 6D pose of $\mathbf{Q}$ relative to the camera's reference frame, with known intrinsics $\mathbf{K}$. Specifically, we aim to determine the 6DoF transformation $\mathbf{T} = (\mathbf{R}, \mathbf{t})$ in 3D space, where $\mathbf{R} \in \mathbb{R}^{3\times3}$ represents the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ denotes the translation vector. The segmented object region $\mathbf{I}_m = \mathbf{M} \odot \mathbf{I}$ is created by element-wise multiplication of the binary segmentation mask $\mathbf{M}$ and $\mathbf{I}$, isolating the visible part of $\mathbf{Q}$. We note that $\mathbf{Q}$ may be partially occluded.

This work leverages the 3D foundation model MASt3R [24] to tackle model-based 6D pose estimation for unseen objects. We introduce Pos3R, a training-free, RGB-only method that effectively estimates 6D poses without object- or task-specific training.

3.2. Training-Free Pipeline

Overview. Following the standard model-based pipeline for 6D pose estimation of unseen objects [2, 39, 41], Pos3R comprises two components: object detection and pose estimation. We keep each component frozen, avoiding any object- or task-specific training.

For *object detection*, we use CNOS [37] to generate segmentation masks and object identities for each target instance, enabling localization of the target segment $\mathbf{I}_m$ within the RGB image $\mathbf{I}$. As a default method in the BOP challenge for segmenting unseen objects, CNOS requires

only 3D models for onboarding and does not depend on additional data or object-specific training.

For *6D pose estimation*, we illustrate three steps in Fig. 2. Specifically, given a set of templates rendered from a textured CAD model, Pos3R utilizes the 3D foundation model MASt3R [24] to extract features from both the target segment $\mathbf{I}_m$ and each template, enabling estimation of the 6D transformation $\mathbf{T} = (\mathbf{R}, \mathbf{t})$ through 2D–3D correspondences using the P*n*P-RANSAC algorithm.

3.2.1. Template Rendering

Using a textured 3D CAD model, we render templates of the object from various orientations. The rendering process follows standard rasterization methods as described in [39], with a black background and fixed lighting. The rendering camera is configured with the same intrinsic parameters, $\mathbf{K}$, as the test camera, and the rendered templates match the size of the test image $\mathbf{I}$. The object remains centered in each template. Additionally, the 3D locations in the coordinate space of the 3D CAD model corresponding to each pixel in the rendered template are recorded, enabling the establishment of 2D–3D correspondences.

A critical component of Pos3R is establishing correspondences between pixels in the target segment and the template. Accurate pose estimation requires addressing both in-plane and out-of-plane rotations, as objects are frequently viewed from diverse orientations. **In-plane rotations** occur within the image plane, such as when an object rotates around the camera's line of sight. **Out-of-plane rotations**, on the other hand, involve 3D orientation changes—such as tilting or turning—that significantly alter an object's appearance due to perspective shifts. Existing methods tackle this challenge through extensive template libraries and specialized selection mechanisms. For example, MegaPose [22] uses hundreds of templates to capture a broad range of object poses, but maintaining such a large template library necessitates a dedicated selection network.

Template Configuration. To reduce reliance on extensive template libraries, Pos3R leverages the 3D foundation model MASt3R, which generates 3D-consistent features across viewpoints. This enables Pos3R to handle out-of-plane rotations effectively without requiring hundreds of templates or complex selection mechanisms. We use a set of eight base templates, denoted by $\{\mathbf{I}_i\}_{i=1}^{8}$, covering essential orientations. They are positioned at the vertices of a cube centered around the CAD model, effectively capturing out-of-plane rotations.

To further address ambiguities from in-plane (axial) rotations, which can impact correspondence quality, we apply controlled rotational variations to each base template. For each template $\mathbf{I}_i$, we generate T rotations around the camera's principal axis, each rotation defined by an angle $\theta_k = \frac{2\pi k}{T}$, where $k = 0, \ldots, T-1$, covering the full 360° range. In our experiments, we set $T = 5$ to balance efficiency and accuracy, yielding a set of rotationally-augmented templates $\{\mathbf{I}_{i,k}\}_{k=1}^{5}$. In the following, we discuss the template selection process in detail.

3.2.2. Image Matching

MASt3R as an Image Matcher. Our approach builds upon MASt3R [24], a model for joint local 3D reconstruction and pixelwise matching between two input images $\mathbf{I}_a$ and $\mathbf{I}_b \in \mathbb{R}^{H \times W \times 3}$. Conceptually, MASt3R operates as a mapping function $\mathbf{f}(\mathbf{I}_a, \mathbf{I}_b) = \text{Dec}(\text{Enc}(\mathbf{I}_a), \text{Enc}(\mathbf{I}_b))$, where the encoder $\text{Enc}(\mathbf{I}) \rightarrow \mathbf{F}$ is a Siamese Vision Transformer (ViT) that processes image $\mathbf{I}$ into token vectors of size $N \times d$, with $N = w \times h$, yielding $\mathbf{F} \in \mathbb{R}^{N \times d}$. The decoder, $\text{Dec}(\mathbf{F}_a, \mathbf{F}_b)$, employs twin ViT modules that produce pixelwise pointmaps $\mathbf{X}$, local feature maps $\mathbf{D}$ for each image.

Using these local feature representations, correspondences between images are identified via the fastNN algorithm [24]. It efficiently establishes reciprocal matches between the feature maps $\mathbf{D}_a$ and $\mathbf{D}_b$ by initially seeding points across a uniform pixel grid and iteratively refining these seeds to form high-quality mutual correspondences. Through this process, fastNN accurately aligns key features across images. The resulting reciprocal pixel pairs between $\mathbf{I}_a$ and $\mathbf{I}_b$ are represented as $M_{a,b} = \{(\mathbf{y}_a^c, \mathbf{y}_b^c)\}_{c=1}^{|M_{a,b}|}$, where $\mathbf{y}_a^c$ and $\mathbf{y}_b^c \in \mathbb{N}^2$ denote the coordinates of matched pixels in each respective image.

Similarity-Based Template Selection. Unlike methods that require hundreds of templates [22, 36, 41], Pos3R simplifies template selection by needing only forty templates, significantly reducing the complexity of the selection process. Rather than relying on a trained template selection network, we employ a straightforward, training-free approach based on the similarity of matched correspondences. Specifically, given a target segment $\mathbf{I}_m$ and a set of eight base templates $\{\mathbf{I}_i\}_{i=1}^{8}$, each augmented with rotational variations $\{\mathbf{I}_{i,k}\}_{k=1}^{5}$, we identify the most similar template for pose estimation by calculating correspondences between the target segment and each template.

For each rotationally-augmented template $\mathbf{I}_{i,k}$, we obtain reciprocal pixel pairs between $\mathbf{I}_m$ and $\mathbf{I}_{i,k}$, represented as: $M_{m,i,k} = \{(\mathbf{y}_m^p, \mathbf{y}_{i,k}^p)\}_{p=1}^{|M_{m,i,k}|}$, where $M_{m,i,k}$ is the set of all reciprocal pixel pairs between $\mathbf{I}_m$ and $\mathbf{I}_{i,k}$, and $\mathbf{y}_m^p$ and $\mathbf{y}_{i,k}^p \in \mathbb{N}^2$ denote the coordinates of matched pixels in the target segment and the template variant, respectively.

For each matched pair $(\mathbf{y}_m^p, \mathbf{y}_{i,k}^p)$ in $M_{m,i,k}$, we retrieve the corresponding local features from the feature maps $\mathbf{D}_m$ and $\mathbf{D}_{i,k}$ generated by MASt3R. Specifically, we obtain the feature vector $\mathbf{f}_m^p$ at coordinate $\mathbf{y}_m^p$ in the target segment from $\mathbf{D}_m$, and the feature vector $\mathbf{f}_{i,k}^p$ at coordinate $\mathbf{y}_{i,k}^p$ in the template from $\mathbf{D}_{i,k}$. We compute the feature similarity for each matched pair as: $S(\mathbf{f}_m^p, \mathbf{f}_{i,k}^p) = \mathbf{f}_m^p \cdot \mathbf{f}_{i,k}^p$, where $\cdot$ denotes the dot product.

To calculate an overall similarity score between the target segment $\mathbf{I}_m$ and each template variant $\mathbf{I}_{i,k}$, we aggregate the similarity scores across all matched pairs as follows:

$$\text{sim}(\mathbf{I}_m, \mathbf{I}_{i,k}) = \sum_{p=1}^{|M_{m,i,k}|} S(\mathbf{f}_m^p, \mathbf{f}_{i,k}^p), \tag{1}$$

where $M_{m,i,k}$ is the set of all matched pairs between $\mathbf{I}_m$ and $\mathbf{I}_{i,k}$. After computing $\text{sim}(\mathbf{I}_m, \mathbf{I}_{i,k})$ for each pair, we select the template with the highest similarity score:

$$(i_{\text{opt}}, k_{\text{opt}}) = \underset{i\in\{1,\ldots,8\},\, k\in\{1,\ldots,5\}}{\arg\max} \text{sim}(\mathbf{I}_m, \mathbf{I}_{i,k}). \tag{2}$$

The selected template $\mathbf{I}_{i_{\text{opt}},k_{\text{opt}}}$ is then used as the closest match to the target segment for the pose estimation process.

3.2.3. Pose Fitting

After selecting the suitable template $\mathbf{I}_{i_{\text{opt}},k_{\text{opt}}}$, we proceed to estimate pose $\mathbf{T}_m = (\mathbf{R}_m, \mathbf{t}_m)$, where $\mathbf{R}_m$ is a 3D rotation matrix and $\mathbf{t}_m$ is a 3D translation vector that transforms the object from model space to camera space. It relies on a set of 2D-3D correspondences $\mathcal{C}_{t_{\text{final}}} = \{(\mathbf{y}_m^j, \mathbf{P}^j)\}_{j=1}^{|M|}$, where $\mathbf{y}_m^j \in \mathbb{R}^2$ represents the coordinates of matched pixels in the target segment $\mathbf{I}_m$, $\mathbf{P}^j \in \mathbb{R}^3$ denotes the corresponding 3D points in model space from the selected template, and there are $|M|$ pairs. To determine $\mathbf{T}_m$, we solve the Perspective-n-Point (P*n*P) problem, which minimizes the reprojection error:

$$\underset{\mathbf{R}_m,\mathbf{t}_m}{\arg\min} \sum_{j=1}^{|M|} \left\| \mathbf{y}_m^j - \pi(\mathbf{R}_m\mathbf{P}^j + \mathbf{t}_m) \right\|^2, \tag{3}$$

where π is the projection function that maps 3D points to 2D image points according to the camera intrinsics.

To enhance robustness against outliers, we employ the Efficient P*n*P (EP*n*P) algorithm [23] combined with a RANSAC-based fitting strategy [11]. In this approach, P*n*P is applied iteratively to random subsets of four correspondences from $\mathcal{C}_{t_{\text{final}}}$, generating multiple pose hypotheses. For each hypothesis, we count the number of inliers, defined as correspondences where the reprojection error falls below a predefined threshold ϵ:

$$\text{inliers} = \left| \left\{ j : \left\| \mathbf{y}_m^j - \pi(\mathbf{R}_m\mathbf{P}^j + \mathbf{t}_m) \right\| < \epsilon \right\} \right|. \tag{4}$$

The hypothesis with the highest inlier count is selected as the final coarse pose estimate $\mathbf{T}_m$.

4. Experiments

In this section, we first outline the experimental setup (Section 4.1). We then evaluate our method's performance in comparison with previous approaches on the seven core datasets of the BOP challenge [17], examining accuracy, runtime efficiency, and effectiveness when using predicted 3D models (Section 4.2). This evaluation highlights the strengths and unique contributions of our approach. Lastly, we present an ablation study to investigate the impact of various configurations in our method (Section 4.3).

4.1. Experimental Setup

Evaluation Datasets. We evaluate our approach on the seven core datasets of the BOP challenge [17]: LineMod Occlusion (LM-O) [3], T-LESS [15], TUD-L [16], ICBIN [8], ITODD [9], HomebrewedDB (HB)[21], and YCB-Video (YCB-V) [57]. Together, these datasets feature 132 distinct objects and $19,048$ testing instances, each in complex, cluttered scenes with partial occlusions. Table 1 provides the instance count for each dataset. It is worth noting that the *unseen* object pose estimation task remains challenging, with significant room for improvement.

Evaluation Metrics. We use the BOP evaluation protocol [17] for 6D object localization, which assesses pose accuracy with three error metrics: Visible Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD). VSD evaluates only the visible part of the object to handle ambiguities, MSSD measures 3D surface deviation with global symmetries, and MSPD assesses perceivable deviation using object symmetries in projection. A pose is considered correct for a metric e if $e < \theta_e$, with θ_e as the correctness threshold. The Average Recall (AR) for each metric e, denoted AR_e, is the mean Recall over various θ_e thresholds, and for VSD, multiple misalignment tolerances τ. The overall AR score is the average of the three metrics: $\text{AR} = (\text{AR}_{\text{VSD}} + \text{AR}_{\text{MSSD}} + \text{AR}_{\text{MSPD}})/3$.

Pose Refinement. To show Pos3R can integrate with render-and-compare refinement techniques, we apply the refinement method from MegaPose [22] to our result. Given an input image and an estimated pose, the refiner predicts the relative transformation between the initial and ground-truth pose. The refiner is trained on a large-scale dataset with several well-designed techniques [22]. We use the similarity score Eq. 1 to select the top-5 templates and each test crop-template pair gives a pose with EPnP. Then, we use the refiner to update these five pose hypotheses.

4.2. Comparison With the State of the Art

We compare Pos3R with both training-free and training-based methods. The training-free methods include FoundPose [41] and ZS6D [2], while the training-based methods include MegaPose [22], GigaPose [39], OSOP [48], and GenFlow [34]. For pose refinement, we apply the refiner from MegaPose [22]. We use the publicly available code from [39] with the default hyperparameters.

#	Method	Training-Free	Refinement	Datasets (num. instances)							Mean	Time
				LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V		
				1445	6423	600	1786	3041	1630	4123		
Coarse Pose Estimation:												
1	GigaPose [39]	✗	–	29.9	27.3	30.2	23.1	18.8	34.8	29.0	27.6	0.9
2	GenFlow [34]	✗	–	25.0	21.5	30.0	16.8	15.4	28.3	27.7	23.5	3.8
3	MegaPose [22]	✗	–	22.9	17.7	25.8	15.2	10.8	25.1	28.1	20.8	15.5
4	OSOP [48]	✗	–	31.2	–	–	–	–	49.2	33.2	–	–
5	ZS6D [2]	✓	–	29.8	21.0	–	–	–	–	32.4	–	–
6	FoundPose [41]	✓	–	**39.6**	**33.8**	46.7	23.9	20.4	50.8	45.2	37.2	1.7
7	Baseline	✓	–	30.5	23.6	43.2	21.4	15.4	42.5	49.3	32.3	0.7
7	Pos3R (Ours)	✓	–	32.3	31.5	**47.3**	**33.1**	**25.1**	**53.7**	**53.9**	**39.5**	1.4
With Pose Refinement (5 hypotheses):												
8	FoundPose	✗	MegaPose	58.6	54.9	65.7	44.4	36.1	70.3	**67.3**	56.8	11.2
9	GigaPose	✗	MegaPose	**59.9**	57.0	64.5	**46.7**	**39.7**	72.2	66.3	**57.9**	7.3
10	GenFlow	✗	GenFlow	56.3	52.3	**68.4**	45.3	39.5	**73.9**	63.3	57.1	20.9
11	MegaPose	✗	MegaPose	56.0	50.7	**68.4**	41.4	33.8	70.4	62.1	54.7	47.4
12	Pos3R (ours)	✗	MegaPose	56.5	**57.5**	66.8	46.0	37.5	70.1	66.4	57.3	8.0

Table 1. **Performance Comparison on the Seven Datasets of BOP.** This table reports the Average Recall (AR) scores per dataset, the mean AR score across all datasets, and the time required to estimate poses for all objects in an image (in seconds). The runtime data of other methods are sourced from FoundPose [41]. The upper section lists methods for coarse pose estimation without refinement, while the lower section presents methods that incorporate a refinement stage using multiple pose hypotheses, reporting the best refined pose. Methods without task-specific training are marked with a green check mark (✓). In the coarse estimation category, Pos3R achieves the highest AR scores on several datasets and demonstrates good overall generalization with a competitive mean AR score. While refinement-based methods yield slightly higher accuracy in certain cases, Pos3R remains a robust and efficient choice for coarse pose estimation and performs comparably to top refined methods when combined with a refinement stage. The best results in the coarse estimation section are highlighted in **green**, and the best results in the refinement section are highlighted in **blue**. The second-best results are underlined.

Table 1 provides a detailed comparison of Pos3R (Pos3R) with other leading 6D pose estimation methods across the seven primary datasets in the BOP challenge. The upper portion of the table showcases results for coarse pose estimation without refinement, while the lower portion highlights performance after incorporating a pose refinement step with five hypotheses. It is important to note that pose refinement involves task-specific training on extensive pose datasets, making it non-training-free [22].

In the coarse pose estimation category, Pos3R demonstrates superior performance over other training-free methods, including FoundPose and ZS6D, achieving the highest mean accuracy and consistently strong results across most datasets. For instance, Pos3R attains the highest AR (Average Recall) on the TUD-L, HB, and YCB-V datasets, along with a mean AR of 39.5. This high performance across varied datasets reflects Pos3R 's robust generalization capabilities for different object poses and scenarios. Additionally, with a runtime of only 1.4 seconds, Pos3R stands out as an efficient alternative to methods that either require substantial computational resources or are not training-free.

Pos3R remains competitive after pose refinement using MegaPose's refine. Although Pos3R is not inherently designed for refinement, it achieves comparable accuracy to MegaPose and other refined methods across several datasets. These findings underscore the adaptability of Pos3R, as it provides accurate initial pose estimates and benefits from additional refinement techniques—demonstrating its potential for scalable, practical, and flexible applications in 6D pose estimation.

Moreover, we observe that Pos3R faces challenges in achieving competitive accuracy on the LM-O dataset, where significant occlusion poses a major difficulty. Occlusion disrupts the initial image matching process, often resulting in suboptimal pose estimates that refinement methods struggle to correct. FoundPose [41] addresses this issue by employing a bag-of-words strategy for template selection, which has shown greater resilience to occlusion. To improve robustness in such scenarios, incorporating the local contrastive learning approach used by GigaPose [39] could be a promising direction. This method is specifically designed to enhance feature matching in occluded regions, and we consider this an area for future exploration.

4.3. Component Analysis

Pose Estimation Using Predicted 3D Models. Model-based methods for unseen object pose estimation typically

Table 2. **Comparison with Predicted 3D Models on LM-O [3].** Average Recall is reported. All methods use 3D models predicted by Wonder3D [32]. Results with pose refinement, applied to the generated 3D models, are also included for comparison.

#	Method	Single Image-to-3D		GT 3D model
		Coarse	Refined	w/o refinement
1	MegaPose [22]	15.4	25.2	22.7
2	GigaPose [39]	17.6	27.2	29.4
3	Pos3R	**19.5**	**28.5**	**32.3**

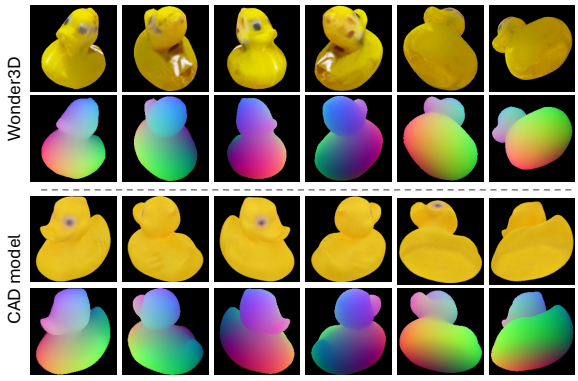

Figure 3. **Rendered Images and 3D Coordinate Maps from Wonder3D and CAD Model.** The top two rows show rendered images and 3D coordinate maps from Wonder3D [32], while the bottom two rows show those from the real CAD model.

require accurate, textured 3D models. To relax this requirement, we follow the approach in [39] and use 3D models predicted by Wonder3D [32] from a single reference image. The predicted 3D models use the same object coordinate system as the ground truth (GT) frame, ensuring aligned axes. We then assess the performance of our method using these reconstructed models in place of the high-fidelity CAD models provided by the dataset. For a fair comparison, we adopt the same settings as [39], including the reference image for Wonder3D and post-processing steps, and report results on the LM-O dataset [3]. Examples of rendered views from generated 3D model are shown in Fig. 3. As shown in Table 2, Pos3R achieves higher AR scores than both MegaPose [22] and GigaPose [39], both before and after pose refinement, showing the effectiveness of Pos3R.

Template Selection Technique. In our experiments, we use the similarity of correspondence matches (Eq. 1) as the primary method for template selection. Additionally, we consider two alternative approaches: 1) Inliers: selecting templates based on the number of inliers (Eq. 4); 2) Confidence: leveraging the 3D pointmaps and confidence maps provided by MASt3R for each target segment-template pair [24], where the average confidence score is used for selection. As illustrated in Fig. 4, template selection based on inliers proves effective, while selection based on similarity score yields the best performance.

Table 3. **Comparison of Pose Estimation Method.** We evaluate our proposed method (Pos3R), a variant without axial rotation augmentation (*w/o* axial rotation), a variant that use face centers as base templates, and a version that replaces the 3D foundation model MASt3R with the 2D foundation model DINOv2. We also replace MASt3R with another dense image matcher RoMa [10]. Pos3R achieves the highest accuracy across all datasets, highlighting the importance of axial rotation handling and the advantage of using 3D foundation models over DINOv2.

#	Method	LM-O	TUD-L	YCB-V	T-Less	IC-BIN	ITODD	HB
1	Pos3R	**32.3**	**47.3**	**53.9**	**31.5**	**33.1**	**25.1**	**53.7**
2	*w/o* Ax. Rot.	30.5	43.2	49.3	23.6	21.4	15.4	42.5
3	Face Center	31.5	45.5	53.3	28.0	29.4	22.4	52.5
4	DINOv2 (L11)	18.5	25.0	22.5	14.2	15.6	14.1	32.1
5	DINOv2 (L9)	20.5	26.0	24.0	15.6	16.6	14.8	32.9
6	RoMa [10]	20.8	25.0	32.0	16.2	17.0	15.5	25.1

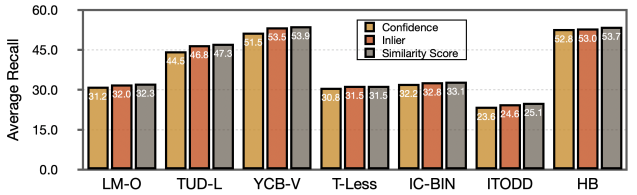

Figure 4. **Comparison of Template Selection Techniques.** We compare template selection methods based on confidence, inliers, and similarity score. The similarity score consistently achieves the highest average recall (AR) across all datasets.

Impact of In-Plane Rotation and 3D Consistency on Pose Estimation. Table 3 evaluates the impact of incorporating in-plane (axial) rotations and 3D-consistent features on pose estimation performance across seven datasets. Our method, Pos3R (row 1), achieves the highest accuracy on all datasets by combining controlled in-plane rotations with the 3D-consistent features provided by the MASt3R foundation model. In contrast, row 2 presents results without in-plane rotation, where only eight base templates are used to account for out-of-plane rotations. The performance drops considerably, highlighting the importance of addressing in-plane rotations for improving correspondence quality and overall pose accuracy. Rows 4 and 5 present a further variation in which we replace MASt3R with the 2D foundation model DINOv2. Despite following the same pipeline (*e.g.*, using the same 40 templates), this substitution leads to a notable performance decline across all datasets. Additionally, in row 6, replacing MASt3R with the dense image matcher RoMa [10] also results in lower accuracy. We speculate that MASt3R's improved performance stems from its large-scale and 3D-aware training, which may enhance its ability to produce features that are well-suited for pose estimation. The above analysis highlights the importance of incorporating both in-plane rotations and 3D-consistent features to achieve robust and accurate 6D pose estimation.

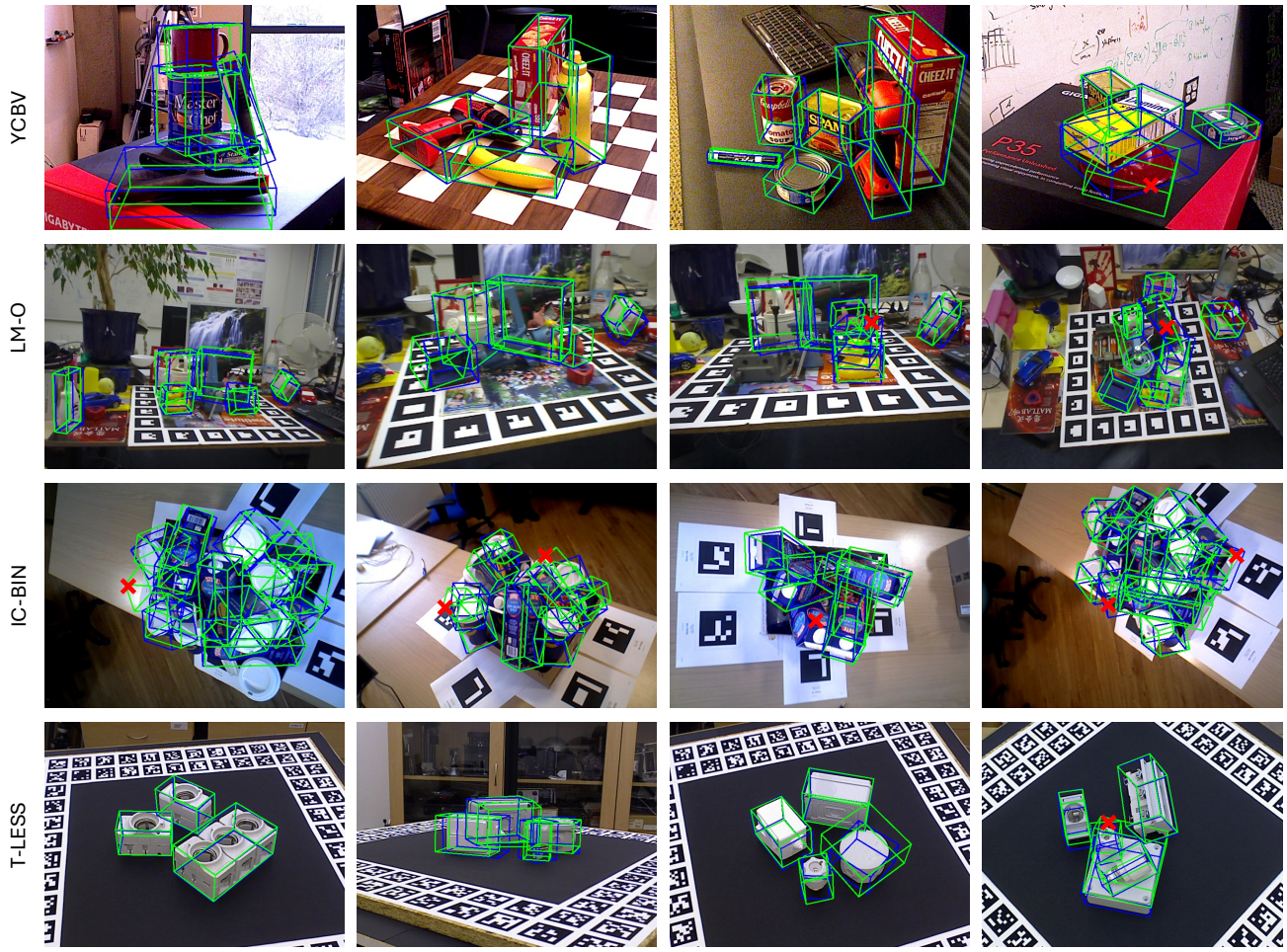

Figure 5. **Qualitative Results of 6D Pose Estimation.** Visualization of 6D object pose estimation of Pos3R across LM-O [3], T-LESS [15], IC-BIN [8], and YCB-V [57]. Pos3R effectively handles a variety of objects and maintains accuracy even in very crowded scenes (*e.g.*, IC-BIN), where predictions remain reasonably robust. Furthermore, Pos3R adapts well to texture-less objects, as demonstrated in the T-LESS dataset. However, as marked by red crosses (✗), heavy occlusion poses a challenge, with reduced accuracy in heavily occluded regions, illustrating both the strengths and limitations of our approach.

Qualitative Results of 6D Pose Estimation. Figure 5 shows pose estimation results of Pos3R across challenging datasets, including LM-O [3], T-LESS [15], IC-BIN [8], and YCB-V [57]. Pos3R demonstrates robust performance across diverse object types, sizes, and textures, handling complex environments effectively. In cluttered scenes like IC-BIN, Pos3R accurately localizes multiple, tightly packed objects. It also performs well on textureless objects in T-LESS, a challenge for vision-based methods due to minimal surface features. However, Figure 5 also reveals a limitation: in heavily occluded scenes, marked by red crosses, Pos3R cannot well handle. Reduced visibility of occluded objects results in deviations in the predicted bounding boxes. This limitation suggests future directions for improvement, such as integrating occlusion-aware methods (*e.g.*, GigaPose [39]) or leveraging multi-view information to enhance robustness in these challenging cases.

5. Conclusion

This work introduces Pos3R, a training-free, RGB-only framework for 6D pose estimation of unseen objects. By leveraging the 3D foundation model MASt3R, Pos3R generates robust, 3D-consistent features that effectively handle both in-plane and out-of-plane rotations. Without relying on extensive datasets or object-specific training, Pos3R establishes a strong baseline for training-free research. Using a minimal set of templates captured from eight cube vertex viewpoints with controlled rotational variations, Pos3R achieves accurate and efficient pose estimation. Experiments on the BOP challenge demonstrate that Pos3R outperforms other training-free methods in coarse pose estimation and achieves competitive results with refined methods when combined with MegaPose refinement. Future work will focus on incorporating occlusion-aware techniques to enhance robustness in occluded settings.

Acknowledgments. We thank all reviewers and ACs for their constructive comments. This work was generously supported through research collaboration with RIOS Intelligent Machines.

References

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2(3):4, 2021. 1, 3

[2] Philipp Ausserlechner, David Haberger, Stefan Thalhammer, Jean-Baptiste Weibel, and Markus Vincze. Zs6d: Zero-shot 6d object pose estimation using vision transformers. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 463–469. IEEE, 2024. 1, 3, 5, 6

[3] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II 13*, pages 536–551. Springer, 2014. 5, 7, 8

[4] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 1, 2

[5] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao–blackwellized particle filter for 6-d object pose tracking. *IEEE Transactions on Robotics*, 37(5):1328–1342, 2021. 1, 2

[6] Francesco Di Felice, Salvatore D'Avella, Alberto Remus, Paolo Tripicchio, and Carlo Alberto Avizzano. One-shot imitation learning with graph neural networks for pick-and-place manipulation tasks. *IEEE Robotics and Automation Letters*, 2023. 1

[7] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020. 1, 2

[8] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3583–3592, 2016. 5, 8

[9] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd-a dataset for 3d object recognition in industry. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2200–2208, 2017. 5

[10] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 7

[11] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 3, 5

[12] Ge Gao, Mikko Lauri, Yulong Wang, Xiaolin Hu, Jianwei Zhang, and Simone Frintrop. 6d object pose regression via supervised learning on point clouds. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3643–3649. IEEE, 2020. 2

[13] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1580–1589, 2020. 3

[14] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022. 2

[15] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 5, 8

[16] 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 *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018. 5

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5610–5619, 2024. 1, 2, 3, 5

[18] Sabera Hoque, Shuxiang Xu, Ananda Maiti, Yuchen Wei, and Md Yasir Arafat. Deep learning for 6d pose estimation of objects—a case study for autonomous driving. *Expert Systems with Applications*, 223:119838, 2023. 1

[19] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2930–2939, 2020. 2

[20] Junwen Huang, Hao Yu, Kuan-Ting Yu, Nassir Navab, Slobodan Ilic, and Benjamin Busam. Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10095–10105, 2024. 2

[21] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 5

[22] Yann Labbé, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *arXiv preprint arXiv:2212.06870*, 2022. 2, 3, 4, 5, 6, 7

[23] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009. 3, 5

[24] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 1, 2, 3, 4, 7

[25] Hongyang Li, Jiehong Lin, and Kui Jia. Dcl-net: Deep correspondence learning network for 6d pose estimation. In *European Conference on Computer Vision*, pages 369–385. Springer, 2022. 2

[26] Yuelong Li, Yafei Mao, Raja Bala, and Sunil Hadap. Mrc-net: 6-dof pose estimation with multiscale residual correlation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10476–10486, 2024. 1, 2

[27] Ruyi Lian and Haibin Ling. Checkerpose: Progressive dense keypoint localization for object pose estimation with graph neural network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14022–14033, 2023. 2

[28] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 2

[29] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, and Qiang Fu. Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation. *IEEE Transactions on Industrial Informatics*, 19 (11):11171–11181, 2023. 1

[30] 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

[31] 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 *European Conference on Computer Vision*, pages 298–315. Springer, 2022. 2

[32] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9970–9980, 2024. 7

[33] Ningkai Mo, Wanshui Gan, Naoto Yokoya, and Shifeng Chen. Es6d: A computation efficient and symmetry-aware 6d pose regression framework. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6718–6727, 2022. 1, 2

[34] Sungphill Moon, Hyeontae Son, Dongcheol Hur, and Sangwook Kim. Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2024. 3, 5, 6

[35] Fengjun Mu, Rui Huang, Ao Luo, Xin Li, Jing Qiu, and Hong Cheng. Temporalfusion: Temporal motion reasoning with multi-frame fusion for 6d object pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5930–5936. IEEE, 2021. 2

[36] Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6771–6780, 2022. 4

[37] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2134–2140, 2023. 2, 3

[38] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Yinlin Hu, Renaud Marlet, Mathieu Salzmann, and Vincent Lepetit. Nope: Novel object pose estimation from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17923–17932, 2024. 2

[39] Van Nguyen Nguyen, Thibault Groueix, Mathieu Salzmann, and Vincent Lepetit. Gigapose: Fast and robust novel object pose estimation via one correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9903–9913, 2024. 2, 3, 4, 5, 6, 7, 8

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3

[41] Evin Pınar Örnek, Yann Labbé, Bugra Tekin, Lingni Ma, Cem Keskin, Christian Forster, and Tomas Hodan. Foundpose: Unseen object pose estimation with foundation features. In *European Conference on Computer Vision*, pages 163–182. Springer, 2025. 1, 3, 4, 5, 6

[42] Jaewoo Park and Nam Ik Cho. Dprost: Dynamic projective spatial transformer network for 6d pose estimation. In *European Conference on Computer Vision*, pages 363–379. Springer, 2022. 2

[43] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10710–10719, 2020. 2

[44] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4561–4570, 2019. 3

[45] Fabio Poiesi and Davide Boscaini. Learning general and distinctive 3d local deep descriptors for point cloud registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3979–3985, 2022. 3

[46] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d

poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017. 1, 2

[47] Caner Sahin and Tae-Kyun Kim. Category-level 6d object pose recovery in depth images. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 1, 2

[48] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6835–6844, 2022. 5, 6

[49] Yongzhi Su, Jason Rambach, Nareg Minaskan, Paul Lesur, Alain Pagani, and Didier Stricker. Deep multi-state object pose estimation for augmented reality assembly. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 222–227. IEEE, 2019. 1

[50] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 2

[51] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 2

[52] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, pages 699–715, 2018. 2

[53] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 1, 2

[54] Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 108–125. Springer, 2020. 1, 2

[55] Jiaxin Wei, Xibin Song, Weizhe Liu, Laurent Kneip, Hongdong Li, and Pan Ji. Rgb-based category-level object pose estimation via decoupled metric scale recovery. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2036–2042. IEEE, 2024. 1, 2

[56] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 1

[57] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 5, 8

[58] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 3

[59] Jiyao 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 *European Conference on Computer Vision*, pages 199–216. Springer, 2025. 1, 2

[60] Heng Zhao, Shenxing Wei, Dahu Shi, Wenming Tan, Zheyang Li, Ye Ren, Xing Wei, Yi Yang, and Shiliang Pu. Learning symmetry-aware geometry correspondences for 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14045–14054, 2023. 2

[61] Wanqing Zhao, Shaobo Zhang, Ziyu Guan, Wei Zhao, Jinye Peng, and Jianping Fan. Learning deep network for detecting 3d object keypoints and 6d poses. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 14134–14142, 2020. 1, 2, 3