

UF0-6D: Unified Flow-based Zero-Shot 6D Object Pose Estimation without Refinement

Yingnan Guo¹, Chun Her Lim¹, Yu Feng², and Yu Zhang^{1, (✉)}

¹ State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, China

zhangyu80@zju.edu.cn

² Information Engineering College, Zhejiang University of Technology, Hangzhou, China

Abstract. Existing zero-shot 6D object pose estimation methods predominantly rely on multi-stage pipelines with iterative render-and-compare refinement or coarse-to-fine hypothesis selection, leading to substantial computational overhead and implementation complexity. Generative modeling on geometric manifolds offers an attractive alternative by directly representing pose posteriors. In this work, we present **UF0-6D**, a **Unified Flow-based Zero-shot** framework for 6D object pose estimation without refinement. UF0-6D reformulates pose estimation as conditional Riemannian flow matching on $SE(3)$, learning an instance-conditioned posterior via geodesic-consistent bridge velocities and symmetry-aware Riemannian optimal transport supervision. A single ODE integration pass replaces coarse-to-refine cascades and iterative alignment, unifying model-based and model-free settings via a geometry-conditioning design that maps both CAD renders and reference-view renders into a shared point-token representation. Our RGB-D variant achieves a mean AR of **81.2** on BOP benchmark datasets, surpassing the previous state-of-the-art by **+6.0** AR, while our RGB variant reaches **70.6** AR at only **0.379s** per image, establishing a new accuracy-speed Pareto frontier.

Keywords: Object 6D pose estimation · Conditional flow matching · Zero-shot learning

1 Introduction

Estimating the rigid 6D transformation remains a long-standing and challenging problem in computer vision. Accurate 6D object pose estimation underpins a wide range of downstream applications, including robotic manipulation [24, 53, 54] and mixed reality [34]. Despite remarkable progress, real-world pose estimation remains challenging due to occlusion, clutter, textureless surfaces, specularities, and symmetries that induce inherent pose ambiguities.

Recent research [7, 15, 28, 45, 50, 52, 60, 61] has increasingly pivoted towards *zero-shot* pose estimation for unseen objects, generally categorized into model-based and model-free paradigms. In the model-based regime, state-of-the-art

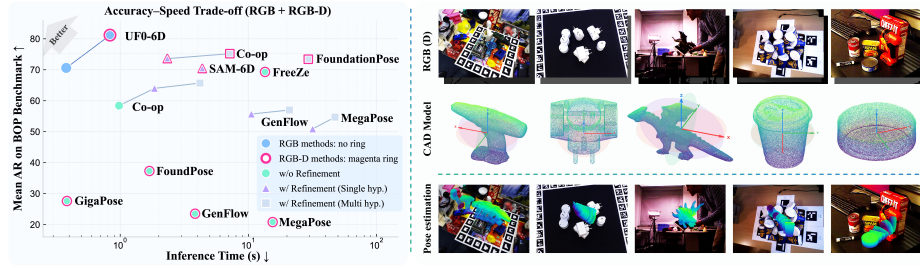


Fig. 1: UF0-6D: accuracy-speed trade-off and qualitative pose-flow visualizations. **Left:** Accuracy-speed trade-off (mean AR vs. inference time) on BOP for representative RGB and RGB-D methods. **Right:** Qualitative examples: RGB (or RGB-D) inputs (top), corresponding CAD models (middle), and the pose-flow trajectory on $SE(3)$ visualized along ODE integration (bottom).

systems typically rely on template retrieval followed by iterative render-and-compare refinement to mitigate discretization errors [27, 35, 55]. Correspondence-driven alternatives attempt to streamline this process by matching query images to rendered templates via PnP /RANSAC [1, 31, 38, 39, 41]. Similarly, in the model-free regime, few-shot pipelines [13, 16, 48] often degrade when facing textureless objects or severe occlusions, limiting their generalization capabilities [33]. Consequently, the prevailing dependence on multi-stage inference and iterative refinement remains a critical bottleneck, hindering efficient and reliable deployment in real-time scenarios.

As shown in Fig. 2, we argue that a key reason refinement is so prevalent is that many methods do not *explicitly model* the conditional posterior over poses under ambiguity. Generative modeling offers an attractive alternative by directly representing pose distributions. Score-based approaches such as GenPose [59] learn pose posteriors via score matching, but estimating scores on manifolds like $SE(3)$ is non-trivial and can be sensitive under symmetry-induced multi-modality. Flow matching [32] provides a deterministic probability-flow formulation that avoids explicit score estimation, and recent Riemannian extensions [8, 42] have shown potential for learning distributions on geometric spaces. However, existing methods mainly do not yet provide a unified, instance-level, zero-shot formulation that works in both model-based and model-free scenarios.

In this work, we present **UF0-6D**, a **U**nified **F**low-based **Z**ero-shot framework for 6D object pose estimation **without refinement**. UF0-6D reformulates pose estimation as learning a *conditional Riemannian probability flow* on $SE(3)$, where conditioning captures both the observed query evidence and an object-specific representation that can originate from either CAD templates or reference-view templates. This unified formulation enables a *single-stage* inference procedure: UF0-6D integrates a learned ODE once to sample from the posterior, without coarse-to-refine design, render-and-compare refinement, or iterative optimization. Importantly, we incorporate symmetry-aware learning by selecting symmetry-consistent transport targets using a Riemannian optimal-

transport-inspired rule, avoiding ad-hoc symmetry-specific heads while respecting the evaluation protocols based on symmetry-aware metrics in BOP.

We train UF0-6D from large-scale synthetic data and evaluate on the seven core datasets of the BOP benchmark. UF0-6D achieves state-of-the-art performance with a superior accuracy-speed trade-off, demonstrating that directly learning the conditional pose posterior on $SE(3)$ can *remove* the long-standing reliance on refinement while improving robustness to occlusion and symmetry.

In summary, our main contributions are:

- **Instance-level conditional flow formulation on $SE(3)$.** To the best of our knowledge, UF0-6D is the first method to formulate zero-shot 6D object pose estimation as conditional Riemannian flow matching on $SE(3)$.
- **Single-stage inference without refinement.** By directly learning the instance-conditioned posterior pose distribution via geodesic-consistent bridge velocities and symmetry-aware Riemannian optimal transport supervision, UF0-6D eliminates coarse-to-refine cascades and iterative render-and-compare refinement, performing robust inference in a single ODE integration pass.
- **State-of-the-art zero-shot results.** As shown in Fig. 1, we demonstrate SOTA zero-shot performance across all seven core BOP datasets, achieving strong accuracy-speed trade-offs without any refinement stage.

2 Related Works

2.1 Model-based Pose Estimation of Novel Objects

Model-based 6D object pose estimation assumes access to accurate CAD models and trains/testing on the same instances, solved via direct pose regression [30,56], 2D–3D correspondence with PnP [44,51], or 3D–3D correspondence and least-squares fitting [14, 15]. While category-level methods relax instance specificity [7,28,50,52,60,61], they typically generalize only within predefined categories and thus cannot handle arbitrary novel objects. Template-based methods render dense pose hypotheses and retrieve the best match [27,35,39,40,46,55], often followed by render-and-compare refinement to improve accuracy [27, 35, 55]. Correspondence-driven methods establish 2D–2D or 2D–3D matches between query images and rendered templates using learned descriptors, then recover pose by PnP/RANSAC [1, 31, 38, 39, 41]. Co-op [36] uses a learned refinement module that estimates probabilistic flow between query and rendered views, coupled with a differentiable PnP layer. FoundationPose [55] leverages large-scale synthetic data generation, transformer-based modules, and a coarse-to-refine strategy with pose selection/ranking for improved generalization.

2.2 Few-shot Model-free Pose estimation of Novel Objects

Model-free approaches remove the need for explicit CAD models and instead rely on a small set of reference images captured at test time [13, 16, 43, 48].

Instance-specific pipelines such as RLLG [4] and NeRF-Pose [29] reconstruct object representations (e.g., radiance fields) and estimate pose, but typically require per-object training and are not designed for broad generalization to arbitrary novel objects. For generalizable novel-object settings, Gen6D [33] introduces a detection–retrieval–refinement pipeline, but still relies on fine-tuning to mitigate distribution shift. OnePose [48] and OnePose++ [13] build object models via SfM and solve pose by matching learned correspondences; FS6D [16] extends similar ideas to RGB-D. However, correspondence quality can degrade significantly for textureless objects and under heavy occlusion, motivating alternative formulations that can directly model the pose posterior under ambiguity.

A closely related direction is to unify model-based and model-free pose estimation within a single system. FoundationPose [55] explicitly supports both scenarios by using an implicit object representation for efficient novel-view synthesis when CAD models are unavailable.

2.3 Generative Modeling for Object Pose Estimation

Generative modeling has recently emerged as a promising paradigm for a variety of tasks beyond traditional generation tasks, such as classification [6], perception [55, 57], and robotics action planning [9, 58]. One of the pioneering works in this area, GenPose [59], proposed learning the 6D pose distribution through score matching. Diff-Reg [23] formulates 6D pose estimation as $SE(3)$ diffusion-based point-cloud registration with reverse transformation denoising. However, score matching struggles with resolving pose ambiguities caused by object symmetries. Recent advancements in Riemannian manifolds [2, 8, 22, 42] have demonstrated the power of flow matching in modeling complex geometric transformations.

3 Method

3.1 Unified Pose Flow Formulation

Problem formulation Given an observation of a target object instance and an object-specific representation, our goal is to infer its 6D pose $g = [\mathbf{R}, \mathbf{t}] \in SE(3)$, where $\mathbf{R} \in SO(3)$ and $\mathbf{t} \in \mathbb{R}^3$. We consider two evaluation regimes under a unified formulation:

Model-based. At test time, a CAD model (mesh) $\mathcal{M}$ is available. We assume a detector provides an object crop (and optionally depth) from the input image $\mathcal{I}$, denoted as $\mathcal{I}^{\text{crop}}$ (and $\mathcal{D}^{\text{crop}}$). The task is to estimate g for the detected instance.

Model-free. At test time, a small set of reference views $\mathcal{R} = \{(\mathcal{I}_k^{\text{ref}}, g_k^{\text{ref}})\}_{k=1}^K$ is provided, with known poses. These references may come from sparse-view capture and can be used to build an implicit or explicit geometry proxy. The task remains estimating g for the query crop.

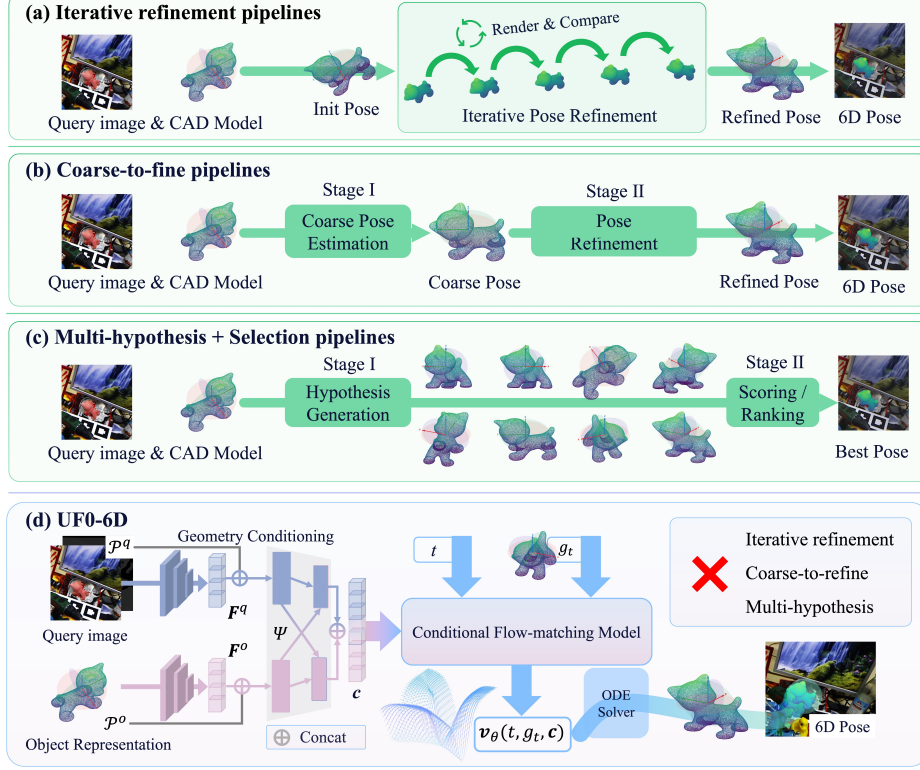


Fig. 2: Overview of UF0-6D and existing paradigms for zero-shot 6D pose estimation. From top to bottom: iterative refinement, coarse-to-fine, multi-hypothesis selection, and **UF0-6D**, which performs single-stage zero-shot 6D pose estimation via a conditional flow on $SE(3)$ without refinement.

We unify both regimes by introducing a *geometry conditioning variable* $\mathbf{c}$ that summarizes the observation and the object representation:

$$\mathbf{c} = \Phi(\mathcal{I}^{\text{crop}}, \mathcal{D}^{\text{crop}}, \mathcal{G}), \quad \mathcal{G} \in \{\text{CAD templates, reference-view templates}\}. \quad (1)$$

The learning target is the conditional posterior $q(g | \mathbf{c})$ over poses on $SE(3)$.

Conditional continuous normalizing flows on $SE(3)$ We model $q(g | \mathbf{c})$ via a conditional continuous normalizing flow (CNF) defined by an ODE-driven probability flow [32]:

$$\frac{dg_t}{dt} = \mathbf{v}_\theta(t, g_t, \mathbf{c}), \quad t \in [0, 1], \quad (2)$$

where $\mathbf{v}_\theta$ is a learnable time-dependent *velocity field* conditioned on $\mathbf{c}$. Starting from a tractable base distribution $\rho_0(g)$ (we use an isotropic base on the Lie

algebra, cf. Sec. 3.3), the flow induces a path of densities $\{\rho_t(\cdot \mid \mathbf{c})\}_{t \in [0,1]}$ whose endpoint $\rho_1(\cdot \mid \mathbf{c})$ approximates $q(\cdot \mid \mathbf{c})$.

The density evolution along the flow satisfies the continuity equation on manifolds [8]:

$$\partial_t \rho_t(g \mid \mathbf{c}) + \nabla_g \cdot (\rho_t(g \mid \mathbf{c}) \mathbf{v}_\theta(t, g, \mathbf{c})) = 0, \quad (3)$$

with $\nabla_g \cdot (\cdot)$ denoting the (Riemannian) divergence operator on $SE(3)$.

3.2 Feature Extraction and Geometry Conditioning

UF0-6D requires a conditioning representation $\mathbf{c}$ that captures (i) *what is observed in the query* and (ii) *what the object is*, in a way that works for both CAD-available and CAD-free settings. We construct $\mathbf{c}$ from two point sets with aligned token features: a query set $\mathcal{P}^q$ from the detected crop and an object template set $\mathcal{P}^o$ from either CAD-rendered templates (model-based) or reference-view templates (model-free).

Query crop tokens $\rightarrow$ point-wise query features Let $\mathcal{I}^{\text{crop}}$ be the RGB crop of the detected object. We extract dense patch tokens using a vision transformer backbone (ViT) [10] and reshape them into a per-pixel embedding map $\mathbf{E}^q \in \mathbb{R}^{H \times W \times d}$. We then sample a set of 2D (or 2.5D with depth) points from the crop:

$$\mathcal{P}^q = \{\mathbf{p}_i^q\}_{i=1}^{N_q}, \quad \mathbf{p}_i^q \in \mathbb{R}^2 \text{ (RGB) or } \mathbb{R}^3 \text{ (RGB-D)}. \quad (4)$$

Each point inherits its feature by indexing the corresponding pixel token:

$$\mathbf{f}_i^q = \mathbf{E}^q(\mathbf{p}_i^q) \in \mathbb{R}^d, \quad \mathbf{F}^q = \{\mathbf{f}_i^q\}_{i=1}^{N_q}. \quad (5)$$

This produces a *point-token* representation of the observed crop that is agnostic to whether depth is present.

Object templates $\rightarrow$ point-wise object features To represent the object instance, we build a template bank that can be instantiated from either CAD or reference views.

Model-based templates. Given a mesh $\mathcal{M}$, we render V views with known camera extrinsics $\{\pi_j\}_{j=1}^V$, producing template images $\{\mathcal{I}_j^o\}_{j=1}^V$ (and optionally depth). We extract ViT [10] tokens $\mathbf{E}_j^o$ and aggregate visible object pixels across views into a pooled point set $\mathcal{P}^o$ by sampling from the union of visible pixels (or back-projected surface points when depth is available):

$$\mathcal{P}^o = \{\mathbf{p}_m^o\}_{m=1}^{N_o}, \quad \mathbf{f}_m^o = \mathbf{E}_{j(m)}^o(\mathbf{p}_m^o), \quad \mathbf{F}^o = \{\mathbf{f}_m^o\}_{m=1}^{N_o}. \quad (6)$$

Model-free templates. Given reference images $\mathcal{R}$, we first reconstruct an object-centric 3D proxy using fast sparse-view 3D Gaussian Splatting [26] (see Appendix for details). We then render template views from this proxy and apply the same token extraction and visible-pixel/point aggregation as in the model-based case, producing $\mathcal{P}^o$ and $\mathbf{F}^o$ in an identical format.

Geometry conditioning vector Finally, we fuse query and object tokens into a conditioning representation $\mathbf{c}$ through a cross-attention encoder Ψ :

$$\mathbf{c} = \Psi(\mathcal{P}^q, \mathbf{F}^q, \mathcal{P}^o, \mathbf{F}^o). \quad (7)$$

Crucially, $\mathbf{c}$ is *instance-specific* (depends on $\mathcal{M}$ or $\mathcal{R}$) but *pipeline-invariant* across model-based and model-free regimes.

3.3 Learning Pose Distribution via Conditional Flow Matching

Flow matching learns $\mathbf{v}_\theta$ by matching it to a known *bridge velocity* that transports a base distribution to the data distribution without requiring score estimation [8, 32]. For 6D pose, the key is to enforce manifold-consistent trajectories on $SE(3)$. We follow the Riemannian flow matching principle and adapt it to *instance-level* pose posteriors by making the conditioning $\mathbf{c}$ explicitly instance-dependent (Sec. 3.2), changing the learning target from category priors to instance posteriors conditioned on object templates and observations.

Riemannian structure on $SE(3)$ We represent poses as elements of the Lie group $SE(3)$. Let $\mathfrak{se}(3)$ be its Lie algebra. We use the matrix logarithm and exponential maps:

$$\log : SE(3) \rightarrow \mathfrak{se}(3), \quad \exp : \mathfrak{se}(3) \rightarrow SE(3). \quad (8)$$

In practice, we decompose $SE(3) \simeq SO(3) \times \mathbb{R}^3$ using the standard disintegration of measures for rigid motions (Haar $\times$ Lebesgue), which allows constructing coupled flows on rotation and translation while maintaining manifold validity.

Geodesic bridge and target velocity Given a ground-truth pose $g_1 = [\mathbf{R}_1, \mathbf{t}_1]$ and a base pose sample $g_0 = [\mathbf{R}_0, \mathbf{t}_0]$, we define a time- t bridge $u(t)$ that lies on the shortest path under the chosen geometry:

Rotation geodesic on $SO(3)$. We use Lie-algebra geodesic interpolation:

$$\mathbf{R}_t = \mathbf{R}_0 \exp\left(t \log(\mathbf{R}_0^\top \mathbf{R}_1)\right). \quad (9)$$

Translation interpolation on $\mathbb{R}^3$. We use Euclidean interpolation:

$$\mathbf{t}_t = (1 - t)\mathbf{t}_0 + t\mathbf{t}_1. \quad (10)$$

The corresponding *target velocity* (bridge vector field) is defined as the time derivative of the bridge at $(\mathbf{R}_t, \mathbf{t}_t)$. For translation, $\dot{\mathbf{t}}_t = \mathbf{t}_1 - \mathbf{t}_0$. For rotation, using the left-trivialized tangent on $SO(3)$, one obtains a constant body-frame angular velocity:

$$\boldsymbol{\omega} = \log(\mathbf{R}_0^\top \mathbf{R}_1) \in \mathfrak{so}(3), \quad \dot{\mathbf{R}}_t = \mathbf{R}_t \boldsymbol{\omega}. \quad (11)$$

Thus the bridge velocity on $SE(3)$ is

$$\mathbf{u}(t; g_0, g_1) = (\mathbf{u}_{SO(3)}(t), \mathbf{u}_{\mathbb{R}^3}(t)) = (\mathbf{R}_t \boldsymbol{\omega}, \mathbf{t}_1 - \mathbf{t}_0). \quad (12)$$

Bi-invariant Riemannian metric and matching objective To compare velocities on the manifold, we introduce a product Riemannian metric on $SE(3)$, combining a bi-invariant metric on $SO(3)$ (e.g., induced by the Killing form) with the Euclidean inner product on $\mathbb{R}^3$ for geometric consistency. We denote the induced norm by $\|\cdot\|_{SE(3)}$.

Conditional Riemannian flow matching trains $\mathbf{v}_\theta$ to match the bridge velocity $\mathbf{u}$ in expectation:

$$\mathcal{L}_{\text{CRFM}}(\theta) = \mathbb{E}_{(\mathbf{c}, g_1) \sim \mathcal{D}} \mathbb{E}_{g_0 \sim \rho_0(\cdot)} \mathbb{E}_{t \sim \mathcal{U}(0,1)} \left[\|\mathbf{v}_\theta(t, g_t, \mathbf{c}) - \mathbf{u}(t; g_0, g_1)\|_{SE(3)}^2 \right], \quad (13)$$

where g_t is the interpolated pose at time t from Eqs. (9)–(10).

Symmetry-aware learning via Riemannian optimal transport A central difficulty in instance-level 6D pose estimation is symmetry-induced pose multiplicity: for an object with symmetry group $\mathcal{S} \subset SE(3)$, the set of equivalent ground-truth poses is

$$\mathcal{Y}(g_1) = \{g_1 \circ s \mid s \in \mathcal{S}\}. \quad (14)$$

In BOP, symmetries are explicitly accounted for by MSSD/MSPD metrics; we incorporate them during learning by adapting the Riemannian optimal transport (ROT) strategy in RFMPose [42].

Concretely, for a fixed source g_0 and a multi-hypothesis target distribution supported on $\mathcal{Y}(g_1)$, we define a Riemannian transport cost using the $\mathfrak{se}(3)$ logarithm:

$$c(g_a, g_b) = \|\log(g_a^{-1}g_b)\|_{\mathfrak{se}(3)}. \quad (15)$$

Following the symmetry-aware target selection strategy used in RFMPose [42], we select the symmetry-consistent pose that minimizes the geodesic cost. In our implementation, this yields a simple and effective symmetry handling rule at training time:

$$\tilde{g}_1 = \arg \min_{g \in \mathcal{Y}(g_1)} c(g_0, g), \quad (16)$$

and we use $\tilde{g}_1$ in the bridge construction (Eqs. (9)–(12)). This replaces ad-hoc symmetry-specific heads/losses with a geometry-first principle: always transport along the shortest symmetry-consistent geodesic.

Single-stage inference: sampling by ODE integration At test time, UF0-6D performs *single-stage* inference by integrating the learned probability flow ODE:

$$\frac{dg_t}{dt} = \mathbf{v}_\theta(t, g_t, \mathbf{c}), \quad g_0 \sim \rho_0, \quad \Rightarrow \quad \{g_1^{(m)}\}_{m=1}^M. \quad (17)$$

In contrast to coarse-to-refine pipelines, there is no separate refinement stage; the flow directly generates 6D pose from the learned posterior.

Pose likelihood and final pose readout In the **model-based** (BOP Benchmark [18, 19, 49]) setting, we use a single hypothesis for efficiency ($M=1$) and directly output the resulting pose. In the **model-free** setting, we draw a small set of hypotheses ($M > 1$) and compute a symmetry-aware final estimate.

To score multiple hypotheses without training an additional energy network, we adopt the end-to-end likelihood estimation used in continuous normalizing flows: the log-density evolves according to

$$\frac{\partial}{\partial t} \log p_t(g_t) = -\nabla_g \cdot \mathbf{v}_\theta(t, g_t, \mathbf{c}), \quad (18)$$

and thus

$$\log p_1(g_1) - \log p_0(g_0) = -\int_0^1 \nabla_g \cdot \mathbf{v}_\theta(t, g_t, \mathbf{c}) dt. \quad (19)$$

Following RFMPose [42], we approximate the divergence via Hutchinson’s trace estimator to avoid explicit Jacobians, enabling practical likelihood evaluation for each hypothesis. For random vectors $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$,

$$\nabla_g \cdot \mathbf{v}_\theta(t, g_t, \mathbf{c}) = \text{tr}(\mathbf{J}_t) \approx \mathbb{E} [\boldsymbol{\epsilon}^\top \mathbf{J}_t \boldsymbol{\epsilon}], \quad (20)$$

where $\mathbf{J}_t = \frac{\partial \mathbf{v}_\theta}{\partial g}$ and the Jacobian-vector product is computed by automatic differentiation.

Given hypothesis set $\{g_1^{(m)}\}_{m=1}^M$ with likelihoods $\{\ell_m\}$, we filter low-likelihood samples and compute a weighted estimate. For translation we use Euclidean weighted mean:

$$\hat{\mathbf{t}} = \frac{\sum_m w_m \mathbf{t}^{(m)}}{\sum_m w_m}, \quad w_m = \exp(\ell_m). \quad (21)$$

For rotation, we use a likelihood-weighted Fréchet mean on $SO(3)$:

$$\hat{\mathbf{R}} = \arg \min_{\mathbf{R} \in SO(3)} \sum_m w_m d_{SO(3)}^2(\mathbf{R}, \mathbf{R}^{(m)}), \quad (22)$$

where $d_{SO(3)}$ is the geodesic distance induced by the chosen bi-invariant metric. Finally, we output $\hat{g} = [\hat{\mathbf{R}}, \hat{\mathbf{t}}]$.

This readout is *not* an iterative refinement: it is a closed-form aggregation of samples from the learned posterior and does not alternate between rendering and alignment (render-and-compare) nor run a coarse-to-fine cascade.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate our method on the seven core datasets of the BOP Challenge [20]: LineMod Occlusion (LM-O) [3], T-LESS [17], TUD-L [18], IC-BIN [11], ITODD [12], HomebrewedDB (HB) [25], and YCB-Video (YCB-V)

Table 1: Results on the BOP challenge datasets. The table shows the Average Recall (AR) score for each dataset, the mean AR score across all datasets, and the time required to estimate the poses of all objects within an image. From top to bottom, the results are ordered as follows: With pose refinement (single hypothesis), With pose refinement (multi hypothesis), and Pose Estimation without Refinement.

#	Coarse Pose Est.	Pose Refinement	Input	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Mean	Time(s)
With pose refinement (single hypothesis):												
1	MegaPose [27]	MegaPose [27]	RGB	49.9	47.7	65.3	36.7	31.5	65.4	60.1	50.9	31.724
2	GenFlow [35]	GenFlow [35]	RGB	54.7	51.4	67.0	43.7	38.4	73.0	61.9	55.7	10.553
3	GigaPose [38]	MegaPose [27]	RGB	55.7	54.1	58.0	45.0	37.6	69.3	63.2	54.7	2.301
4	FoundPose [41]	MegaPose [27]	RGB	55.6	51.1	63.3	43.3	35.7	69.7	66.1	55.0	6.385
5	GigaPose [38]	GenFlow [35]	RGB	59.5	55.0	60.7	47.8	41.3	72.2	60.8	56.8	2.213
6	GigaPose [38]	Co-op [36]	RGB	60.2	<u>56.0</u>	62.8	50.2	43.4	73.9	61.9	58.3	1.149
7	Co-op [36]	Co-op [36]	RGB	64.2	63.5	71.7	51.2	47.3	<u>83.2</u>	67.0	64.0	<u>1.852</u>
8	SAM-6D [31]	SAM-6D [31]	RGB-D	<u>69.9</u>	51.5	<u>90.4</u>	58.8	60.2	77.6	<u>84.5</u>	<u>70.4</u>	4.367
9	Co-op [36]	Co-op [36]	RGB-D	71.5	64.6	90.5	<u>57.5</u>	<u>58.2</u>	85.7	87.4	73.6	2.331
With pose refinement (multi hypothesis):												
10	MegaPose [27]	MegaPose [27]	RGB	56.0	50.7	68.4	41.4	33.8	70.4	62.1	54.7	47.386
11	GenFlow [35]	GenFlow [35]	RGB	56.3	52.3	68.4	45.3	39.5	73.9	63.3	57.0	20.890
12	GigaPose [38]	MegaPose [27]	RGB	59.8	56.5	63.1	47.3	39.7	72.2	66.1	57.8	7.682
13	FoundPose [41]	MegaPose [27]	RGB	61.0	57.0	69.3	47.9	40.7	72.3	69.0	59.6	20.523
14	Co-op [36]	Co-op [36]	RGB	65.5	<u>64.8</u>	72.9	<u>54.2</u>	49.0	<u>84.9</u>	68.9	65.7	4.186
16	FoundationPose [55]	FoundationPose [55]	RGB-D	75.6	64.6	92.3	50.8	<u>58.0</u>	83.5	88.9	73.4	29.317
15	Co-op [36]	Co-op [36]	RGB-D	<u>73.0</u>	66.4	<u>90.5</u>	59.7	61.3	87.1	88.7	75.2	<u>7.162</u>
Pose Estimation without Refinement:												
17	ZS6D [1]	—	RGB-D	29.8	21.0	—	—	—	32.4	—	—	—
18	MegaPose [27]	—	RGB-D	22.9	17.7	25.8	15.2	10.8	25.1	28.1	20.8	15.465
19	GenFlow [35]	—	RGB-D	25.0	21.5	30.0	16.8	15.4	28.3	27.7	23.5	3.839
20	GigaPose [38]	—	RGB-D	29.9	27.3	30.2	23.1	18.8	34.8	29.0	27.6	<u>0.384</u>
21	FoundPose [41]	—	RGB-D	39.7	33.8	46.9	23.9	20.4	50.8	45.2	37.3	1.690
22	Co-op [36]	—	RGB	59.7	59.2	64.2	45.8	39.1	78.1	62.6	58.4	0.979
23	FreeZe [5]	—	RGB-D	<u>68.9</u>	52.0	<u>93.6</u>	49.9	<u>56.1</u>	79.0	<u>85.3</u>	69.3	13.474
24	UF0-6D(ours)	—	RGB	66.2	<u>62.3</u>	86.9	<u>60.2</u>	54.2	<u>80.3</u>	84.1	<u>70.6</u>	0.379
25	—	—	RGB-D	75.9	73.3	95.9	71.2	73.2	90.0	88.7	81.2	0.832

[56]. For the model-free setup, we follow [16, 55] and select a small set of reference images with known poses from the training split of each dataset to construct the object representation at test time. Our model is not trained on any images from these datasets; training relies solely on synthetic data. Additional implementation details are provided in the appendix.

Metrics.

Model-based setup (BOP benchmark). For the model-based evaluation, we strictly follow the BOP Challenge protocol [18, 19, 49]. Pose accuracy is assessed using three error functions: Visible Surface Discrepancy (VSD), Maximum Symmetry-Aware Surface Distance (MSSD), and Maximum Symmetry-Aware Projection Distance (MSPD). The overall performance is reported as the Average Recall (AR), which is the mean of the recall rates across these three metrics: $AR = (AR_{VSD} + AR_{MSSD} + AR_{MSPD})/3$.

Table 2: Model-free pose estimation results on YCB-Video measured by AUC of ADD and ADD-S. “Finetune-free” indicates no fine-tuning on the target (test) instances.

Method	Ref. images	Finetune-free	ADD-S	ADD
PREDATOR [21]	16	✓	71.0	24.3
LoFTR [47]	16	✓	52.5	26.2
FS6D-DPM [16]	16	×	88.4	42.1
FoundationPose [55]	16	✓	<u>97.4</u>	91.5
UF0-6D(ours)	3		93.8	86.1
	6	✓	96.8	<u>91.0</u>
	12		97.6	91.8

Model-free setup. For the model-free setting, we adopt the standard ADD and ADD-S metrics [56]. We report the Area Under the Curve (AUC) of ADD and ADD-S, following prior works [13, 16, 55].

4.2 Quantitative results

Model-based. Table 1 showcases the performance of UF0-6D on the seven core BOP datasets under the zero-shot, model-based setting. For a fair comparison with previous works [1, 27, 35, 38, 41], we adopt the default CNOS [37] detections. Our single-stage, refinement-free UF0-6D establishes a new state of the art, demonstrating a remarkable combination of accuracy and speed.

Specifically, our RGB-D variant achieves a mean AR of **81.2**, significantly outperforming all prior methods, including those that rely on computationally expensive multi-hypothesis refinement pipelines like Co-op [36] (75.2 AR) and FoundationPose [55] (73.4 AR). This result is a strong testament to the power of our flow-based formulation, which can model complex pose distributions more effectively than conventional coarse-to-fine approaches.

The RGB-only version of UF0-6D is equally compelling. With a mean AR of **70.6**, surpassing recent works like FreeZe [5] (69.3 AR) and Co-op (coarse) [36] (58.4 AR), but also outperforms strong RGB-D methods with dedicated refinement stages like SAM-6D [31] (70.4 AR). Crucially, this top-tier accuracy is achieved with exceptional efficiency. Our RGB model is the fastest among all high-performing coarse methods at just **0.379s**, while our RGB-D model (0.832s) remains significantly faster than refinement-based competitors.

Model-free. Table 2 presents our model-free pose estimation results on the YCB-Video dataset. All methods are evaluated in a finetune-free setting, ensuring a fair comparison of zero-shot generalization capabilities. Our UF0-6D demonstrates exceptional data efficiency and accuracy. Notably, with only 12 reference views, our method achieves state-of-the-art performance (**97.6** ADD-S / **91.8** ADD), surpassing FoundationPose [55] which requires 16 views.

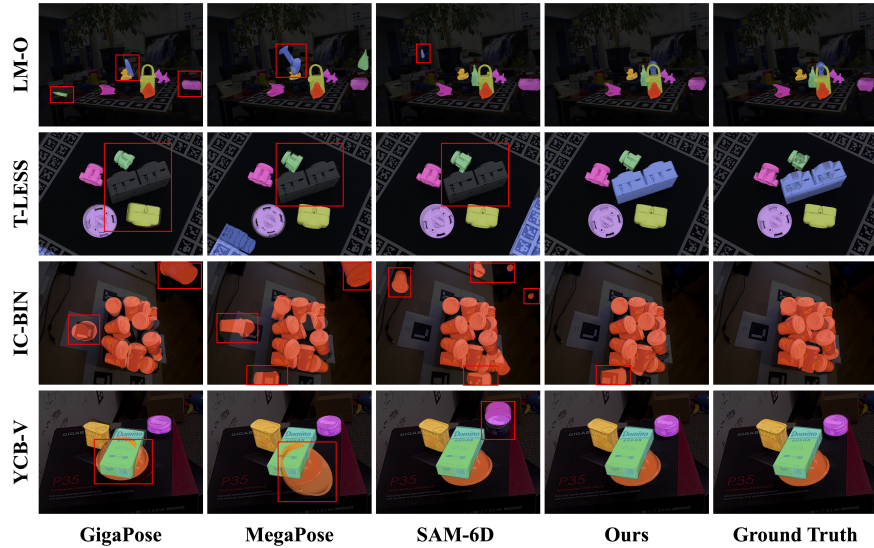


Fig. 3: Qualitative comparison on core BOP datasets. UF0-6D (ours) achieves precise alignment in a single stage without iterative refinement, outperforming MegaPose [27], GigaPose [38], and SAM-6D [31] across LM-O, T-LESS, IC-BIN, and YCB-V. Ground truth is shown for reference.

The ability to generate accurate poses from a compact geometric representation derived from minimal data demonstrates that UF0-6D is a highly practical solution, obviating the need for dense reconstructions or the per-object fine-tuning required by methods like FS6D [16].

4.3 Qualitative results

Comparison with other methods. Fig. 3 presents qualitative comparisons on representative samples from the BOP benchmark [20].

Across datasets, UF0-6D produces more consistent 3D placements, particularly in challenging cases with heavy occlusion (LM-O, IC-BIN), textureless objects and severe symmetries (T-LESS), and cluttered tabletop scenes (YCB-V). More qualitative comparisons are provided in the appendix.

Visualizing the single-stage pose flow on $SE(3)$. Fig. 4 visualizes the intermediate poses along the ODE integration trajectory at $t \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. We show results on the five BOP datasets that provide ground-truth annotations for visualization. Starting from a base pose sample at $t=0.0$, UF0-6D progressively transports the pose towards the high-probability region conditioned on the observation, and the rendered object aligns increasingly well with the image until reaching the final estimate at $t=1.0$.

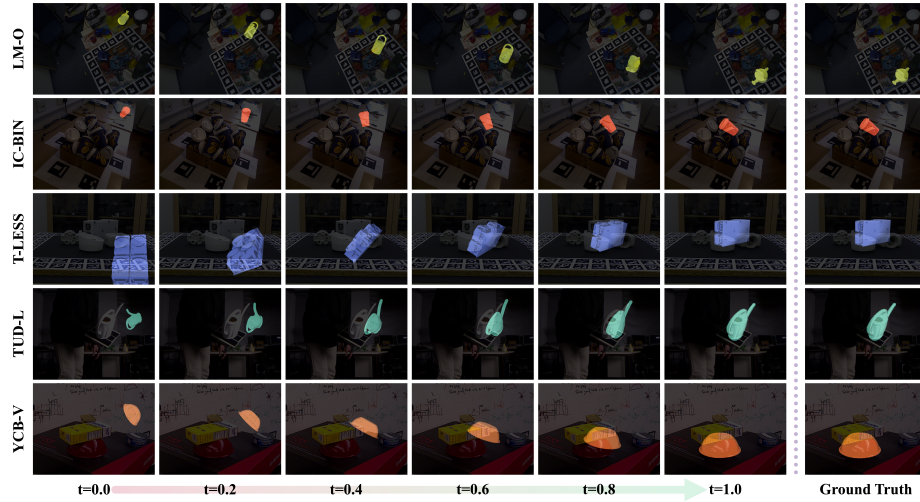


Fig. 4: Pose flow visualization. UF0-6D trajectory at $t \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ on LM-O, T-LESS, IC-BIN, TUD-L, and YCB-V.

While the learned flow is robust under typical initialization noise, we observe that UF0-6D may fail when the initial pose deviation is excessively large, leading the trajectory to converge to an incorrect mode. Qualitative failure cases and further analysis are provided in the appendix.

4.4 Ablation Studies and Analyses

Ablation on geometry conditioning components Our unified formulation leverages an *instance-specific* conditioning vector $\mathbf{c}$ that fuses evidence from the query crop and the object representation (CAD- or reference-based templates), cf. Eq. (7). To assess the contribution of each component to the learned posterior flow on $SE(3)$, as shown in Tab. 3, we ablate the geometry conditioning by selectively removing (i) query features $\mathbf{F}^q$, (ii) object/template features $\mathbf{F}^o$, and (iii) their geometric coordinates $\mathcal{P}^q, \mathcal{P}^o$.

Key observations. (1) **Both query and object features are essential.** Removing either $\mathbf{F}^q$ or $\mathbf{F}^o$ causes a substantial performance drop, while removing both features leads to a collapse (RGB: $70.6 \rightarrow 42.7$, RGB-D: $81.2 \rightarrow 55.4$). This indicates that UF0-6D does not merely learn a category-level pose prior; instead, it relies on *instance-conditioned* evidence to shape the posterior flow, consistent with the objective in Eq. (13).

(2) **Query features contribute more than object features, especially for RGB.** Discarding $\mathbf{F}^q$ degrades performance more than discarding $\mathbf{F}^o$ (RGB: -14.7 vs. -9.8 ; RGB-D: -12.1 vs. -8.3). We attribute this to $\mathbf{F}^q$ encoding appearance-dependent cues (e.g., occlusion and illumination) that determine

Table 3: Ablation on geometry conditioning vector $\mathbf{c}$. Mean AR (%) on the seven core BOP datasets [20]. We toggle the usage of query/object point coordinates ($\mathcal{P}^q, \mathcal{P}^o$) and query/object features ($\mathbf{F}^q, \mathbf{F}^o$) in the cross-set encoder Ψ .

Input	Setting	Conditioning inputs to Ψ				AR $\uparrow$
		$\mathcal{P}^q$	$\mathbf{F}^q$	$\mathcal{P}^o$	$\mathbf{F}^o$	
RGB	w/o both feats.	✓		✓		42.7
	w/o query feat.	✓		✓	✓	55.9
	w/o object feat.	✓	✓	✓		60.8
	Full UF0-6D	✓	✓	✓	✓	70.6
RGB-D	w/o both feats.	✓		✓		55.4
	w/o query feat.	✓		✓	✓	69.1
	w/o object feat.	✓	✓	✓		72.9
	Full UF0-6D	✓	✓	✓	✓	81.2

Table 4: Ablation on ODE integration steps on LM-O. AR is evaluated under the standard BOP protocol [18, 19, 49]. Step=20 corresponds to our default setting used in the main results (Tab. 1).

Input	ODE steps	AR $\uparrow$	Time (s) $\downarrow$
RGB	1	57.8	0.154
	5	62.9	0.226
	10	65.4	0.342
	20	<u>66.2</u>	0.379
	40	66.3	0.637
RGB-D	1	66.4	0.357
	5	72.8	0.469
	10	75.1	0.641
	20	<u>75.9</u>	0.832
	40	76.0	1.342

which pose modes remain plausible, whereas $\mathbf{F}^o$ primarily supports instance identification and view-consistent geometric matching. The effect is stronger for RGB due to the absence of depth constraints.

(3) RGB-D is more resilient to weakened conditioning but still benefits from full fusion. Depth reduces ambiguity in translation and partially constrains rotation, yielding higher AR than RGB even under ablations.

Effect of ODE Integration Steps UF0-6D performs single-stage inference by integrating the learned probability-flow ODE on $SE(3)$ (Eq. (2)). More integration steps improve numerical accuracy but increase runtime due to additional network evaluations. Tab. 4 shows a clear accuracy–speed trade-off. Using few steps noticeably degrades AR (e.g., step=1 is -8.4 AR for RGB and -9.5 AR for RGB-D compared to step=20). Accuracy increases with more steps but quickly saturates: 20→40 steps yields only $+0.1$ AR for both inputs (RGB and RGB-D) while substantially increasing runtime. Based on this observation, we use **20 ODE steps** as default.

5 Conclusion

In this paper, we presented **UF0-6D**, a unified, flow-based framework for zero-shot 6D object pose estimation that eliminates the need for iterative refinement. By reformulating pose estimation as learning a conditional Riemannian probability flow directly on the $SE(3)$ manifold, we address the inherent challenges of multi-modality and symmetry without relying on the computational burden of coarse-to-fine cascades or render-and-compare loops. Our approach unifies model-based and model-free settings through a robust geometry conditioning mechanism and leverages large-scale synthetic data to achieve strong Sim2Real generalization. Extensive evaluations on the seven core datasets of the BOP benchmark demonstrate that UF0-6D significantly outperforms existing methods in both accuracy and inference speed. These results validate that theoretically grounded generative modeling on geometric manifolds offers a superior alternative to traditional regression-and-refinement paradigms. UF0-6D offers a scalable single-stage framework for robust open-world 6D pose estimation, enabling efficient deployment in latency-sensitive applications such as real-time robotic manipulation and mixed reality.

Acknowledgements

This research was supported by the Baima Lake Laboratory Joint Funds of the Zhejiang Provincial Natural Science Foundation of China under Grant No. LBMHD25F030001, in part by the National Natural Science Foundation of China (Grant No. 62576311), and in part by Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F030001.

References

1. Ausserlechner, P., Habegger, D., Thalhammer, S., Weibel, J.B., Vincze, M.: Zs6d: Zero-shot 6d object pose estimation using vision transformers. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 463–469. IEEE (2024)
2. Bose, J., Akhound-Sadegh, T., Huguet, G., Fatras, K., Rector-Brooks, J., Liu, C., Nica, A., Korablyov, M., Bronstein, M., Tong, A.: Se (3)-stochastic flow matching for protein backbone generation. In: International Conference on Learning Representations. vol. 2024, pp. 22590–22621 (2024)
3. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6d object pose estimation using 3d object coordinates. In: European conference on computer vision. pp. 536–551. Springer (2014)
4. Cai, M., Reid, I.: Reconstruct locally, localize globally: A model free method for object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3153–3163 (2020)
5. Caraffa, A., Boscaini, D., Hamza, A., Poiesi, F.: Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In: European Conference on Computer Vision. pp. 414–431. Springer (2024)

6. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niebner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. In: 2017 International Conference on 3D Vision (3DV). pp. 667–676 (2017). <https://doi.org/10.1109/3DV.2017.00081>
7. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11973–11982 (2020)
8. Chen, R.T., Lipman, Y.: Flow matching on general geometries. In: International Conference on Learning Representations. vol. 2024, pp. 47922–47945 (2024)
9. Chi, C., Xu, Z., Feng, S., Cousineau, E., Du, Y., Burchfiel, B., Tedrake, R., Song, S.: Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* **44**(10-11), 1684–1704 (2025)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
11. Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.K.: Recovering 6d object pose and predicting next-best-view in the crowd. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3583–3592 (2016)
12. Drost, B., Ulrich, M., Bergmann, P., Hartinger, P., Steger, C.: Introducing mvtec itodd-a dataset for 3d object recognition in industry. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 2200–2208 (2017)
13. He, X., Sun, J., Wang, Y., Huang, D., Bao, H., Zhou, X.: OnePose++: Keypoint-free one-shot object pose estimation without CAD models. *Advances in Neural Information Processing Systems (NeurIPS)* **35**, 35103–35115 (2022)
14. He, Y., Huang, H., Fan, H., Chen, Q., Sun, J.: Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3003–3013 (2021)
15. He, Y., Sun, W., Huang, H., Liu, J., Fan, H., Sun, J.: Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11632–11641 (2020)
16. He, Y., Wang, Y., Fan, H., Sun, J., Chen, Q.: Fs6d: Few-shot 6d pose estimation of novel objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 6814–6824 (2022)
17. Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: 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). pp. 880–888. IEEE (2017)
18. Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., et al.: Bop: Benchmark for 6d object pose estimation. In: Proceedings of the European conference on computer vision (ECCV). pp. 19–34 (2018)
19. Hodaň, T., Sundermeyer, M., Drost, B., Labbé, Y., Brachmann, E., Michel, F., Rother, C., Matas, J.: Bop challenge 2020 on 6d object localization. In: European Conference on Computer Vision. pp. 577–594. Springer (2020)
20. Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J.: 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. pp. 5610–5619 (2024)

21. Huang, S., Gojcic, Z., Usvyatsov, M., Wieser, A., Schindler, K.: Predator: Registration of 3d point clouds with low overlap. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 4267–4276 (2021)
22. Huguet, G., Vuckovic, J., Fatras, K., Thibodeau-Laufer, E., Lemos, P., Islam, R., Liu, C.H., Rector-Brooks, J., Akhound-Sadegh, T., Bronstein, M., et al.: Sequence-augmented se (3)-flow matching for conditional protein generation. *Advances in neural information processing systems* **37**, 33007–33036 (2024)
23. Jiang, H., Salzmann, M., Dang, Z., Xie, J., Yang, J.: Se (3) diffusion model-based point cloud registration for robust 6d object pose estimation. *Advances in Neural Information Processing Systems* **36**, 21285–21297 (2023)
24. Kappler, D., Meier, F., Issac, J., Mainprice, J., Cifuentes, C.G., Wüthrich, M., Berenz, V., Schaal, S., Ratliff, N., Bohg, J.: Real-time perception meets reactive motion generation. *IEEE Robotics and Automation Letters* **3**(3), 1864–1871 (2018)
25. Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
26. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., et al.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4), 139–1 (2023)
27. Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: Megapose: 6d pose estimation of novel objects via render & compare. In: Proceedings of the 6th Conference on Robot Learning (CoRL) (2022)
28. Lee, T., Tremblay, J., Blukis, V., Wen, B., Lee, B.U., Shin, I., Birchfield, S., Kweon, I.S., Yoon, K.J.: Tta-cope: Test-time adaptation for category-level object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21285–21295 (2023)
29. Li, F., Vutukur, S.R., Yu, H., Shugurov, I., Busam, B., Yang, S., Ilic, S.: Nerf-pose: A first-reconstruct-then-regress approach for weakly-supervised 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2123–2133 (2023)
30. Li, Z., Wang, G., Ji, X.: Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7678–7687 (2019)
31. Lin, J., Liu, L., Lu, D., Jia, K.: 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. pp. 27906–27916 (2024)
32. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022)
33. Liu, Y., Wen, Y., Peng, S., Lin, C., Long, X., Komura, T., Wang, W.: Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In: European Conference on Computer Vision. pp. 298–315. Springer (2022)
34. Marchand, E., Uchiyama, H., Spindler, F.: Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics* **22**(12), 2633–2651 (2015)
35. Moon, S., Son, H., Hur, D., Kim, S.: Genflow: Generalizable recurrent flow for 6d pose refinement of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10039–10049 (2024)
36. Moon, S., Son, H., Hur, D., Kim, S.: Co-op: Correspondence-based novel object pose estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11622–11632 (2025)

37. Nguyen, V.N., Groueix, T., Ponimatin, G., Lepetit, V., Hodan, T.: Cnos: A strong baseline for cad-based novel object segmentation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2134–2140 (2023)
38. Nguyen, V.N., Groueix, T., Salzmann, M., Lepetit, V.: Gigapose: Fast and robust novel object pose estimation via one correspondence. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9903–9913 (2024)
39. Nguyen, V.N., Hu, Y., Xiao, Y., Salzmann, M., Lepetit, V.: 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*. pp. 6771–6780 (2022)
40. Okorn, B., Gu, Q., Hebert, M., Held, D.: Zephyr: Zero-shot pose hypothesis rating. In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 14141–14148. IEEE (2021)
41. Örneke, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: Foundpose: Unseen object pose estimation with foundation features. In: *European Conference on Computer Vision*. pp. 163–182. Springer (2024)
42. Ouyang, W., Ye, Q., Wang, J., Xu, Z., Chen, J.: Rfm-pose: Generative category-level object pose estimation via riemannian flow matching. *Advances in Neural Information Processing Systems* **38**, 67591–67610 (2026)
43. Park, K., Mousavian, A., Xiang, Y., Fox, D.: 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*. pp. 10710–10719 (2020)
44. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 7668–7677 (2019)
45. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnnet: Pixel-wise voting network for 6dof pose estimation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 4561–4570 (2019)
46. Shugurov, I., Li, F., Busam, B., Ilic, S.: Osop: A multi-stage one shot object pose estimation framework. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6835–6844 (2022)
47. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: Loftr: Detector-free local feature matching with transformers. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 8922–8931 (2021)
48. Sun, J., Wang, Z., Zhang, S., He, X., Zhao, H., Zhang, G., Zhou, X.: Onepose: One-shot object pose estimation without cad models. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 6825–6834 (2022)
49. Sundermeyer, M., Hodaň, T., Labbe, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J.: Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 2785–2794 (2023)
50. Tian, M., Ang Jr, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: *European Conference on Computer Vision*. pp. 530–546. Springer (2020)
51. Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. In: *Conference on Robot Learning (CoRL)*. pp. 306–316 (2018)

52. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2642–2651 (2019)
53. Wen, B., Lian, W., Bekris, K., Schaal, S.: Catgrasp: Learning category-level task-relevant grasping in clutter from simulation. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 6401–6408. IEEE (2022)
54. Wen, B., Lian, W., Bekris, K., Schaal, S.: You only demonstrate once: Category-level manipulation from single visual demonstration. RSS (2022)
55. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17868–17879 (2024)
56. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In: RSS (2018)
57. Xie, F., Wang, Z., Ma, C.: Diffusiontrack: Point set diffusion model for visual object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19124 (2024)
58. Ze, Y., Zhang, G., Zhang, K., Hu, C., Wang, M., Xu, H.: 3d diffusion policy: Generalizable visuomotor policy learning via simple 3d representations. arXiv preprint arXiv:2403.03954 (2024)
59. Zhang, J., Wu, M., Dong, H.: Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems* **36**, 54627–54644 (2023)
60. Zhang, R., Di, Y., Manhardt, F., Tombari, F., Ji, X.: Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7452–7459. IEEE (2022)
61. Zheng, L., Wang, C., Sun, Y., Dasgupta, E., Chen, H., Leonardis, A., Zhang, W., Chang, H.J.: Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 17163–17173 (2023)