G2FM: A Geodesic Flow Matching Framework with Geometric Prior for Category-Level 9-DoF Pose Estimation

Qianyu Chen[1], Xiaogang Zhang[2], Yangyi Wan[2], Zhengzhao Pan[2], Kai Wang[1], Yuqi Cai[2], Wenbin Yan[4], Feng Yang[1], and Hua Chen[3]

[1] School of Artificial Intelligence and Robotics, Hunan University, Changsha, China
[2] College of Electrical and Information Engineering, Hunan University, Changsha, China
zhangxg@hnu.edu.cn
[3] College of Computer Science and Electronic Engineering, Hunan University, Changsha, China
chua@hnu.edu.cn
[4] Electric Power Research Institute, State Grid Hunan Electric Power Co., Ltd., Changsha, China
Corresponding authors: Xiaogang Zhang and Hua Chen.

Abstract. Estimating category-level 9-DoF object poses (rotation, translation, scale) is fundamental for robotic manipulation and autonomous systems. Generative approaches, particularly diffusion-based models, have achieved strong accuracy by effectively bridging the sim-to-real domain gap through large-scale synthetic training. However, it is difficult for diffusion-based approaches to achieve both high accuracy and real-time performance due to their reliance on a multi-step iterative inference process. To address this limitation, we propose Geometric Prior Geodesic Flow Matching (G2FM), an efficient, simulation-free training objective that learns a deterministic vector field to directly map noise to pose distributions. G2FM learns a vector field mapping noise to pose distributions with only three ODE steps, achieving high accuracy and real-time performance (> 30 FPS). To avoid the structural distortion and singularities caused by Euclidean approximations, we construct a geometrically-consistent trajectory that operates directly on the $SO(3)$ manifold by transporting rotations along geodesic paths. Furthermore, to dynamically refine the generative flow during inference and enhance accuracy without retraining, we design a training-free geometric guidance mechanism by leveraging Chamfer distance gradients to project pose corrections directly onto the $SO(3)$ tangent space. Extensive experiments show that G2FM delivers a favorable accuracy-efficiency-generalization trade-off for category-level RGB-D 9-DoF pose estimation. Code will be released upon publication.

Keywords: 9-D pose estimation · Flow matching · Training-free guidance · SO(3) manifold · Geodesic path

1 Introduction

9-DoF pose estimation, which involves predicting an object's rotation, translation, and scale relative to a camera's coordinate system, is a cornerstone technology for numerous applications, including augmented reality [40, 45], robotic manipulation [1, 74], and autonomous driving [68].

With traditional image processing methods, object pose estimation relied on template matching and feature point correspondences [45], which struggled with texture-less objects and complex backgrounds. Recently, deep learning has revolutionized this field by directly learning robust feature representations from data. These data-driven methods can generally be categorized into two main paradigms: discriminative and generative approaches. Discriminative methods model pose estimation as a deterministic mapping process. Early discriminative works directly regressed pose parameters or keypoints [61, 62]. However, these models often exhibited limited generalization capabilities when faced with novel objects or severe Sim-to-Real domain shifts. To address this, driven by the advancement of Vision Foundation Models, modern discriminative approaches [4, 6, 10, 11, 14, 24, 32, 41, 43, 67] leverage robust, large-scale feature extractors to establish zero-shot dense correspondences or perform template matching. While these models significantly improve generalization and inference efficiency, their deterministic nature fundamentally limits their ability to capture the inherent ambiguities and severe occlusions often encountered in real-world scenes.

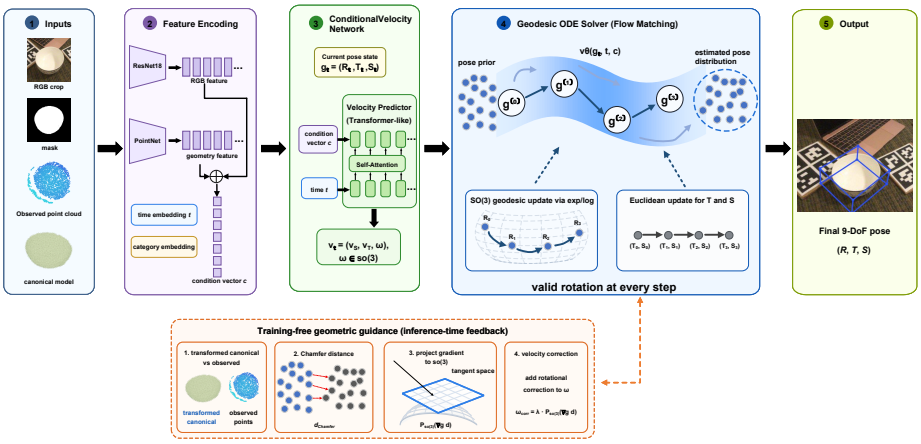

Fig. 1: Overview of G2FM. RGB, depth point cloud, and category prior features condition a velocity network. The ODE solver updates rotation through an $SO(3)$ geodesic path and updates translation/scale in Euclidean space, with training-free geometric guidance applied at inference time.

To overcome these limitations and probabilistically model observational uncertainties, generative approaches have gained prominence. By learning the con-

ditional probability distribution of poses rather than a single point estimate, these methods can naturally manage ambiguities such as object symmetries [25]. Early generative works explored VAEs [56] and GANs [50], though they frequently suffered from mode collapse. Building on the capability to model complex data distributions, diffusion models [25, 39, 78] have emerged as a powerful paradigm, treating pose estimation as a generative sampling process. However, diffusion models face a critical bottleneck: their reliance on a discrete, multi-step iterative denoising process incurs substantial computational overhead. This leads to slow inference speeds, making it exceedingly difficult to meet the strict real-time requirements of practical applications such as robotic manipulation.

Recently, Flow Matching (FM) [37] has become a novel and promising generative backbone. Its efficient simulation-free objective and deterministic straight probability paths endow it with high-speed generation capabilities, making it an ideal candidate to overcome the diffusion bottleneck. However, most existing FM frameworks primarily focus on standard Euclidean data spaces. In object pose estimation, modeling 3D rotation is mathematically constrained by strict orthogonality and determinant properties, meaning that valid rotations naturally reside on the non-Euclidean Special Orthogonal group, or $SO(3)$ manifold. Consequently, directly applying standard Euclidean FM models to the 9-DoF pose space leads to an inherent geometric mismatch. As illustrated in Fig. 2, standard linear interpolation cuts blindly through the ambient space, corrupting the rotational structure by violating these orthogonality constraints. In an iterative ODE process, these structural corruptions accumulate rapidly, causing the entire trajectory to drift and leading to severe generative collapse.

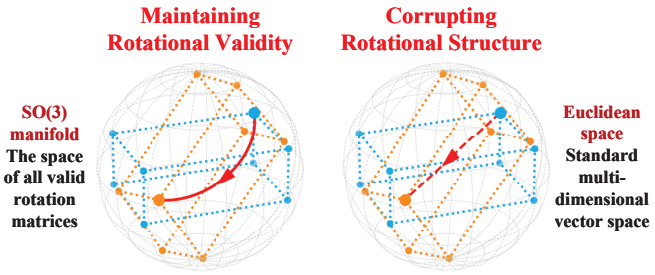

Fig. 2: Comparison of interpolation paths on the $SO(3)$ manifold. **(Left)** Our geodesic path strictly operates on the manifold surface, preserving rotational validity. **(Right)** Naive Euclidean interpolation cuts through the space, corrupting structural orthogonality and yielding invalid rotations.

To address the above challenges, we propose G2FM, a novel category-level 9-DoF pose estimation architecture based on continuous-time Flow Matching. Firstly, considering the inherent non-Euclidean nature of rotation, G2FM presents an innovative Geodesic Flow formulation to decouple the pose into a $SO(3) \times$

$\mathbb{R}^3 \times \mathbb{R}^3$ product manifold. It strictly parameterizes the rotational flow as a valid angular velocity in the $\mathfrak{so}(3)$ tangent space, completely resolving structural corruption and guaranteeing geometric consistency at every ODE step. In addition, we design an innovative Training-Free Geometric Guidance mechanism. By analytically projecting Chamfer distance gradients onto the tangent space, this mechanism performs preferential refinement across the generative trajectory, adaptively injecting structural priors to enhance pose accuracy without incurring any retraining overhead.

Our main contributions are summarized as follows:

- **1)** We propose G2FM, a highly efficient continuous generative architecture specifically tailored for category-level 9-DoF pose estimation. By establishing a deterministic, manifold-aware probability path, G2FM alleviates the speed-accuracy trade-off of existing diffusion models, enabling robust real-time perception.
- **2)** We propose a Geodesic Flow Matching formulation specifically designed for complex high-dimensional pose spaces. By explicitly transporting rotations along geodesic paths on the $SO(3)$ manifold, it strictly preserves geometric validity during the entire generation process.
- **3)** We propose a novel Training-Free Geometric Guidance mechanism. It efficiently maps structural point-cloud constraints into valid tangent-space angular velocities, continuously refining the generative flow on the fly.
- **4)** Comprehensive evaluations on synthetic and real-world datasets demonstrate that G2FM achieves a strong accuracy-efficiency-generalization trade-off while running at > 30 FPS. This work provides valuable guidance for the efficient modeling of non-Euclidean geometric data in generative pose estimation.

2 Related Work

2.1 Pose Estimation

Object pose estimation comprises two paradigms. Instance-level methods [5, 23, 61, 62] rely on precise CAD models, limiting generalization. In contrast, the category-level paradigm is significantly more challenging as it must generalize to unseen object instances, handling large intra-class variations in geometry and texture. Early methods pioneered this area by defining shared canonical spaces, such as NOCS [63], to create a common representation for all instances within a category. Subsequent research has explored diverse strategies, including voting schemes based on geometric features [73], fusing semantic priors with geometric information [9], or leveraging stereo vision for more reliable 3D data [76]. Many deterministic approaches [15, 53, 66, 73, 75] often depend on synthetic training, but this reliance on the Sim-to-Real paradigm introduces a significant domain gap and observational uncertainty.

Beyond category-level pose estimation, broader 3D scene understanding also studies object geometry in full scenes. Indoor reconstruction methods such as

Total3DUnderstanding [46] and From Points to Multi-Object 3D Reconstruction [13] recover room layouts, object meshes, and multi-object scene structure. CAD alignment and object-lifting methods, including Scan2CAD [2], ROCA [21], SPARC [28], Vid2CAD [44], DiffCAD [19], and OmniNOCS [27], retrieve or align CAD models, or lift image observations into object-centric 3D representations. Scene synthesis datasets and methods, such as 3D-FRONT [16], ATISS [51], and DiffuScene [57], focus on generating plausible indoor layouts. Related scene-level benchmarks such as ScanNet++ [71], Omni3D [3], and Hypersim [54] emphasize full-scene geometry, object detection, or holistic indoor understanding. In contrast, G2FM follows the category-level RGB-D 9-DoF pose-and-size setting used by NOCS/REAL275, where each detected object is evaluated by canonical pose, translation, and scale metrics.

For category-level pose estimation under Sim-to-Real domain shift, diffusion models have emerged as a powerful paradigm because they can model uncertainty and generate robust pose hypotheses. These diffusion-based approaches generally follow two main strategies:

- Denoising Intermediate Representations: This strategy focuses on refining the intermediate outputs (like coordinates or heatmaps) required for pose estimation, rather than the pose itself. DiffusionNOCS [25] iteratively denoises NOCS (Normalized Object Coordinate Space) maps to improve 2D-3D correspondence robustness. 6D-Diff [69] models the detection of 2D keypoints as a reverse diffusion process. It generates precise keypoint heatmaps by denoising from an initial noisy prediction, thereby enhancing the accuracy of 2D keypoint detection which is crucial for pose estimation. Similarly, Diff-COPE [58] refines a sparse set of keypoints using a diffusion process conditioned on category information, specifically to better handle intra-class shape variations. Taking a different approach, EA6D [80] introduces an Environment Decoupling Diffusion Model (EDDM) to generate environment-independent object representations in a latent space, aiming to decouple the object's features from environmental factors like background clutter and illumination.
- Operating Directly on Pose Space: This strategy treats the pose itself as the target data to be generated or refined, operating directly on the manifold or in a continuous parameter space. GenPose [78] models the pose distribution on the SE(3) manifold to generatively sample plausible poses conditioned on the input image. Diff9D [39] refines an initial coarse pose by performing diffusion directly in the 9-DoF space to achieve a highly accurate result. Other works optimize this process, such as [29], which proposes a joint learning (regression + diffusion) strategy and score-scaling guidance to accelerate training and replace the extra outlier-filtering network used in models like GenPose [78]. This direct strategy is advantageous as it models the target pose distribution holistically, bypassing the potential error accumulation from intermediate representations like keypoints or coordinate maps. By operating directly in the pose space, these methods can better learn the complex

correlations between pose parameters and inherently handle the geometric properties of the pose manifold.

2.2 Flow Matching

Flow Matching (FM) [37] and Rectified Flow [42] are efficient generative frameworks that learn direct, ODE-based trajectories. Recent work further studies how model-aligned source-target coupling affects FM trajectories [36]. This efficiency has led to its rapid adoption in image and layout generation.

In image synthesis, Yazdani *et al.* [70] used optimal transport to create a straighter mapping for medical images, which significantly accelerates sampling and enhances quality compared to diffusion models. Oh *et al.* [47] introduced FlawMatch for industrial defect generation, using spatial attributes as discrete conditions, which achieved a 17x sampling speed-up and better stability over DDPM. Similarly, Andrade Guerreiro *et al.* [20] highlighted FM's geometrical interpretation in layout generation, where smooth, directed paths lead to a faster sampling process than the noisy trajectories of diffusion models.

Building on this foundation, FM has also been applied to complex geometric data. Chen and Lipman [8] extended FM to general Riemannian manifolds, which has been utilized for rigid-body transformations in physical sciences [26, 72] and robotics [18, 79]. Recently, concurrent work RFMPose [49] explored Riemannian flow matching on the $SE(3)$ manifold to resolve symmetric ambiguities for 6-DoF instance-level pose estimation.

In contrast to the aforementioned methods that primarily focus on 6-DoF rigid-body motion, the proposed G2FM aims to tackle the more complex category-level 9-DoF pose estimation task. By decoupling the pose space into a $SO(3) \times \mathbb{R}^3 \times \mathbb{R}^3$ product manifold, it efficiently accommodates the intra-category scale variations. Furthermore, rather than modifying the training-time objective, G2FM introduces a plug-and-play training-free geometric guidance mechanism during inference. This design allows the model to dynamically inject structural priors to refine the generative trajectory, achieving high-precision pose estimation while maintaining real-time inference speed (> 30 FPS).

2.3 Rotation Representations and Geometric Refinement

Rotation parameterization is closely related to category-level pose estimation. Continuous representations such as 6D rotation [81] reduce discontinuities for neural rotation regression and are effective for single-pose prediction. For generative pose modeling, however, the representation also interacts with the sampling trajectory. A rotation can be normalized at the endpoint, while the intermediate states of a Euclidean flow may still leave the rotation manifold. This motivates manifold-aware updates that model rotational motion through Lie-algebra velocities and exp/log maps. This view is complementary to neural descriptor and refinement methods such as NRDF [55], which use learned geometric fields for local pose refinement.

3 Proposed Method

Our G2FM framework, illustrated in Fig. 1, is a continuous-time generative model for 9-DoF pose estimation. This section first introduces the overall framework and its continuous-time formulation, clarifying the relationship between the continuous-time variable t and the discrete inference steps (Sec. 3.1). We then detail our core contributions: the Geodesic Flow Matching (GFM) formulation (Sec. 3.2) and the training-free geometric guidance mechanism (Sec. 3.3).

3.1 Overall Framework and Continuous-Time Formulation

As illustrated in Fig. 1, G2FM processes multimodal inputs using a ResNet-18 and a PointNet to extract global RGB and geometric point cloud features, respectively. These features are then concatenated to guide a Transformer-based velocity prediction module. The estimation process is iterative: starting from a near-identity pose, a numerical ODE solver applies this predicted velocity over a few discrete steps to guide the pose along a geodesic flow, converging at the final ground-truth pose. Detailed network architectures are provided in the supplementary.

The core of G2FM is a continuous-time generative framework parameterizing the trajectory from a simple prior distribution p_0 at $t = 0$ to the target pose distribution p_1 at $t = 1$. Unlike models trained to reverse a fixed stochastic process, we use Conditional Flow Matching (CFM) [37] to efficiently regress a neural network $v_t(x; \theta)$ towards a target vector field $u_t(x|x_1)$. The CFM loss is defined as:

$$\mathcal{L}_{CFM}(\theta) = \mathbb{E}_{t, p_t(x_1), p_t(x|x_1)} \left[\| v_t(x; \theta) - u_t(x|x_1) \| \right] \tag{1}$$

This simulation-free training allows the model to learn deterministic, straight trajectories, enabling high-quality synthesis in just a few ODE steps during inference.

3.2 Geometric Prior Geodesic Flow Matching

Applying standard Euclidean Flow Matching to the non-Euclidean $SO(3)$ manifold violates orthogonality constraints ($R^T R = I$), as illustrated in Fig. 2. Furthermore, while the $SE(3)$ manifold is typically utilized for 6-DoF rigid-body motion, it fundamentally cannot accommodate the scale variations (S) inherent in our category-level 9-DoF task. To simultaneously maintain geometric validity and task compatibility, we decouple the pose space into a product manifold $SO(3) \times \mathbb{R}^3 \times \mathbb{R}^3$. This allows us to define independent, closed-form generation paths: geodesic paths for rotation on $SO(3)$, and standard linear paths for translation and scaling in Euclidean space.

We represent a 9-DoF pose as $g = (R, T, S) \in SO(3) \times \mathbb{R}^3 \times \mathbb{R}^3$. We define the conditional probability path $p_t(g|g_1)$ from a randomly sampled noisy pose $g_0 \sim p_0$ to a real data pose $g_1 \sim p_1$. Crucially, unlike standard generative models

that often use a wide-variance Gaussian noise prior, we adopt a "warm start" strategy for p_0. We define p_0 as a distribution of small random perturbations centered around the identity pose ($R = I, T = 0, S = 0$). This prior is sampled independently of the input observation, rather than obtained from a detector, PnP solver, or learned refiner. The network still learns $\log(R_1 R_0^T)$ toward arbitrary target rotations, while the near-canonical start keeps the few-step ODE trajectory stable and efficient. The points $g_t = (R_t, T_t, S_t)$ along this path are generated through the following geodesic interpolation method:

$$\begin{aligned} R_t &= \exp(t \cdot \log(R_1 R_0^T)) R_0 \\ T_t &= (1-t)T_0 + tT_1 \\ S_t &= (1-t)S_0 + tS_1 \end{aligned} \tag{2}$$

This path ensures that during interpolation, R_t is always a valid rotation matrix, thereby maintaining pose geometry without introducing structural distortion.

The condition vector field $u_t(g_t|g_1)$ corresponding to the geodesic path g_t is the velocity vector that drives g_0 to g_1. For the mixed path we defined, this vector field has a clear closed-form solution:

$$u_t(g_t|g_1) = \big(\log(R_1 R_0^T), T_1 - T_0, S_1 - S_0\big) \tag{3}$$

Here, $\log(R_1 R_0^T)$ calculates the minimum rotation vector (angular velocity) in the tangent space $\mathfrak{so}(3)$ required to rotate from R_0 to R_1. The exponential map then projects a scaled portion of this vector back to the $SO(3)$ manifold, ensuring smooth traversal along the geodesic. Notably, for this specifically constructed linear-geodesic formulation, the conditional vector field $u_t(g_t|g_1)$ remains constant in the tangent-space representation along the entire continuous trajectory. Similar to flow matching in Euclidean space, our goal is to train a neural network $v_\theta(g, t, c)$, conditioned on the observed point cloud and other information c, to approximate this vector field. Therefore, we propose the geodesic flow matching loss $\mathcal{L}_{GFM}$ (Eq. (4)), which follows the principle of Conditional Flow Matching [37]:

$$\mathcal{L}_{GFM}(\theta) = \mathbb{E}_{t, g_1, g_0, c}\left[\|v_\theta(g_t, t, c) - u_t(g_t|g_1)\|\right] \tag{4}$$

where $t \in \mathcal{U}[0, 1]$, $g_1 \sim p_{data}$, and $g_0 \sim p_{noise}$, with g_t obtained through the aforementioned interpolation. By optimizing this mathematically tractable objective, our network learns a deterministic, manifold-constrained vector field capable of smoothly guiding any noisy initialization directly to the ground-truth pose.

Beyond maintaining geometric validity, our geodesic formulation inherently adopts a *Tangent Space Linearization* strategy. As formulated in Eq. (3), projecting the flow into the tangent space $\mathfrak{so}(3)$ via the logarithmic map converts the complex non-linear manifold regression into a linear learning task of a constant angular velocity. Consequently, compared to the stochastic gradients in diffusion models [39], our deterministic geodesic flow guarantees consistent and straight-line gradients. This analytical stability establishes the essential mathematical foundation for the training-free geometric guidance mechanism.

3.3 Training-free Geometric Guidance

The guidance mechanism computes a geometric loss $\mathcal{L}_{geom}$ at each ODE step t. We define this loss using the Chamfer Distance to measure the discrepancy between the transformed canonical model P_{model} (using the current pose g_t) and the observed points P_{obs}:

$$\mathcal{L}_{geom}(g_t) = \text{Chamfer}(\tau(P_{model}, g_t), P_{obs}) \tag{5}$$

where $\tau(P, g) = R(P \odot S) + T$ is the pose transformation function.

We then compute the gradients of this loss with respect to the pose components: $\nabla_{R_t}\mathcal{L}_{geom}$, $\nabla_{T_t}\mathcal{L}_{geom}$, and $\nabla_{S_t}\mathcal{L}_{geom}$. A critical step is to project the ambient space gradient $\nabla_{R_t}\mathcal{L}_{geom}$ onto the tangent space $\mathfrak{so}(3)$ at R_t to obtain a valid angular velocity correction, ω_{corr} . This ensures the update preserves the $SO(3)$ manifold structure.

Finally, we use the calculated gradient to correct the vector $v_\theta(g_t, t, c) = (v_S, v_T, \omega_{model})$ predicted by the neural network, obtaining a geometrically-guided final velocity $v'_t = (v'_S, v'_T, \omega')$:

$$\begin{aligned} v'_S &= v_S - \gamma_S \nabla_{S_t} L_{geom}, \\ v'_T &= v_T - \gamma_T \nabla_{T_t} L_{geom}, \\ \omega' &\leftarrow \omega_{model} - \gamma_R \omega_{corr} \end{aligned} \tag{6}$$

where γ_S, γ_T, and γ_R control the guiding strength. The corrected velocity vector v'_t is used to perform the single-step update of the ODE solver, integrating the geometric constraints into the entire generation process. Due to space limits, the comprehensive mathematical proofs regarding Euclidean structural corruption and the discrete ODE solver implementations are detailed in the supplementary material.

4 Experiment Results

To evaluate domain generalization, we train the proposed G2FM exclusively on synthetic data and test it on challenging real-world and zero-shot benchmarks.

4.1 Dataset and Evaluation Metrics

For the synthetic datasets, we use the CAMERA25 dataset [63]. For real-world evaluation, we use REAL275 [63], Wild6D [17], and Omni6DPose [77]. REAL275 and Wild6D follow category-level RGB-D 9-DoF pose-and-size evaluation with object-centric masks/point clouds. Omni6DPose is used as a zero-shot cross-domain test without fine-tuning.

We use the standard IoU3D and n°-$m$cm metrics. IoU3D measures 3D bounding-box overlap, while n°-$m$cm reports the percentage of predictions whose rotation and translation errors are below the given thresholds.

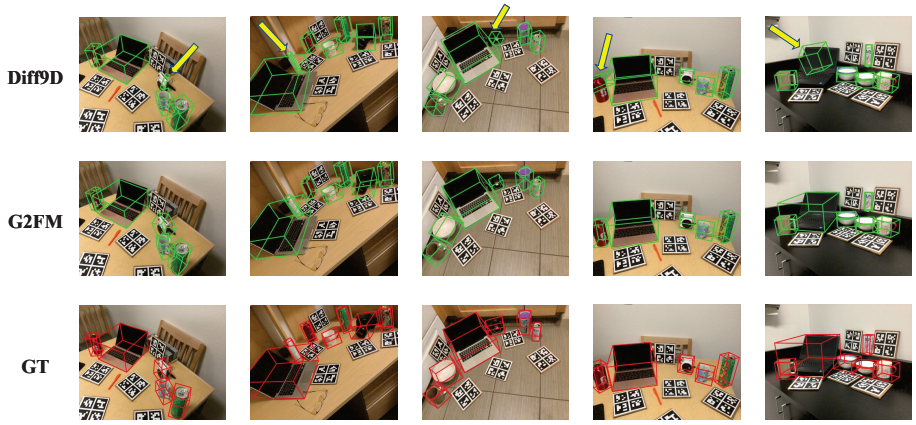

Fig. 3: Qualitative results. From top to bottom: predicted poses by Diff9D [39], predicted poses by G2FM, and ground-truth poses. Trained only on synthetic data, G2FM demonstrates excellent generalization, accurately estimating object poses in cluttered real-world environments.

4.2 Evaluation Results

Evaluation on the REAL275 Dataset We compare G2FM with several state-of-the-art methods on the REAL275 dataset. The detailed results are presented in Tab. 1.

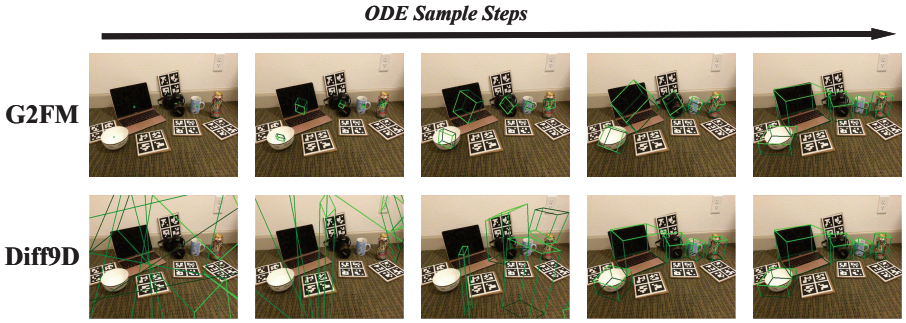

Fig. 4: Visualization of the G2FM and Diff9D sampling process. Our method generates geometrically valid poses from the very first step, which then smoothly converge to the final estimation.

As shown in Tab. 1, G2FM improves the accuracy-efficiency trade-off on REAL275. When using an ImageNet-pretrained segmenter, our method outperforms the strong diffusion baseline Diff9D on the challenging IoU-based metrics ($3D_{50}$ and $3D_{75}$), which demonstrates strong generalization from synthetic to real-world data. Under the stricter setting with a segmenter trained on synthetic

Table 1: Comparison on REAL275 using IoU3D (%) and $n°$-$m$cm (%) metrics. "Syn", "Real w Label", "Real w/o Label", and "Shape Prior" indicate whether each method uses synthetic training data, labeled real training data, unlabeled real training data, and category shape priors. "✓" and "-" indicate with and without the corresponding setting. The best and second-best results are bolded and underlined. Results for competing methods are adopted from Diff9D [39].

Method	Syn	Real w Label	Real w/o Label	Shape Prior	mean Average Precision (mAP)					
					$3D_{50}$	$3D_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
Segmented by Mask R-CNN [22] Pretrained on ImageNet Dataset										
SPD [59]	✓	-	-	✓	55.1	17.1	11.4	12.0	33.5	37.8
SGPA [7]	✓	-	-	✓	53.1	17.8	19.8	27.7	36.5	62.6
STG6D [38]	✓	-	-	✓	48.4	14.4	20.1	27.9	39.0	63.7
SAR-Net [33]	✓	-	-	✓	-	-	31.6	<u>42.3</u>	50.3	68.3
SPD [59]	✓	✓	-	✓	68.5	27.5	19.3	21.4	43.2	54.1
CR-Net [64]	✓	✓	-	✓	-	33.2	27.8	34.3	47.2	60.8
SGPA [7]	✓	✓	-	✓	68.8	36.6	<u>35.9</u>	39.6	**61.3**	**70.7**
RePoNet [17]	✓	-	✓	✓	76.0	-	29.1	31.3	48.5	56.8
SSC6D+ICP [52]	✓	-	✓	-	72.7	-	28.6	33.4	51.8	62.9
DiffusionNOCS [25]	✓	-	-	-	-	-	-	35.0	-	66.6
Diff9D [39]	✓	-	-	-	<u>76.5</u>	<u>41.7</u>	35.3	**43.9**	54.8	<u>70.0</u>
G2FM (Ours)	✓	-	-	-	**80.0**	**44.8**	**36.5**	41.2	<u>57.5</u>	67.6
Segmented by Mask R-CNN [22] Pretrained on CAMERA25 Dataset										
DPDN [34]	✓	-	-	✓	67.2	39.8	29.7	37.3	53.7	67.0
UDA-COPE [30]	✓	-	✓	-	**75.5**	34.4	<u>30.5</u>	34.9	57.0	66.1
TTA-COPE [31]	✓	-	-	-	69.1	39.7	30.2	35.9	**61.7**	**73.2**
VI-Net [35]	✓	-	-	-	33.3	11.3	11.7	19.9	14.4	29.1
SecondPose [9]	✓	-	-	-	34.9	14.7	12.7	21.2	17.6	34.2
Diff9D [39]	✓	-	-	-	69.2	<u>44.1</u>	**36.5**	**45.2**	<u>57.7</u>	<u>72.2</u>
G2FM (Ours)	✓	-	-	-	<u>71.0</u>	**44.9**	**36.5**	<u>42.1</u>	56.8	68.2
VI-Net [35]	✓	✓	-	-	-	48.3	50.0	**57.6**	70.8	**82.1**
Diff9D [39]	✓	✓	-	-	79.8	55.8	**50.5**	57.1	**72.1**	81.5
G2FM (Ours)	✓	✓	-	-	**82.7**	**57.1**	48.9	53.5	70.6	77.7

data CAMERA25, G2FM maintains its competitive edge, achieving top performance on the $3D_{75}$ metric. The qualitative results in Fig. 3 further validate these quantitative improvements, showing that our predicted poses align more closely with the ground truth in challenging, cluttered scenes.

The ODE visualization in Fig. 4 further shows that G2FM keeps poses geometrically valid from early steps, starting from a plausible initial pose and smoothly evolving toward the final solution. This contrasts with diffusion sampling, where intermediate boxes can be unstable or geometrically invalid before convergence.

Evaluation on the Wild6D Dataset We also evaluate G2FM on the Wild6D dataset. As shown in Tab. 2, G2FM achieves highly competitive results across multiple key metrics. Notably, while competing methods often utilize real-world training data, our model is trained solely on synthetic data and still demonstrates strong domain generalization. This validates the effectiveness of our framework in addressing the Sim-to-Real domain shift.

Table 2: Comparison on Wild6D using IoU3D (%) and $n°$-$m$cm (%) metrics. "Syn", "Real w Label", "Real w/o Label", and "Shape Prior" follow the same definitions as in Tab. 1. * denotes training with both CAMERA25 and REAL275. The best and second-best results are bolded and underlined. Results for competing methods are adopted from Diff9D [39].

Method	Syn	Real w Label	Real w/o Label	Shape Prior	mean Average Precision (mAP)					
					$3D_{50}$	$3D_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
NOCS* [63]	✓	✓	-	-	0.00	0.00	0.04	0.04	0.04	0.04
SPD* [59]	✓	✓	-	✓	52.64	20.26	6.90	9.26	20.10	27.84
SGPA* [7]	✓	✓	-	✓	63.48	34.46	**26.28**	29.18	**33.80**	39.52
GPV-Pose* [12]	✓	✓	-	-	-	-	14.10	21.50	23.80	<u>41.10</u>
Diff9D [39]	✓	-	-	-	<u>76.48</u>	**38.16**	<u>25.52</u>	<u>30.46</u>	<u>32.48</u>	40.94
G2FM (Ours)	✓	-	-	-	**77.54**	<u>36.20</u>	23.18	**30.62**	29.96	**41.76**

Zero-shot Generalization on Omni6DPose To validate robustness beyond REAL275, we evaluate our CAMERA25-trained model directly on the large-scale Omni6DPose dataset (ROPE subset) [77] without fine-tuning.

Table 3: Zero-shot comparison on the Omni6DPose (ROPE subset). Methods in the top section are supervised on the target domain (149 categories) and are listed solely as Oracle References. The bottom section evaluates pure Zero-Shot transfer (trained only on 6 categories of CAMERA25).

Method	$3D_{25}$	$3D_{50}$	$3D_{75}$	5°2cm	5°5cm
Supervised on Target Domain (149 classes) - Reference Upper-bound					
GenPose++ [77]†	39.0	19.1	2.0	10.0	15.1
RFMPose [49]†	42.1	21.0	2.2	10.4	15.7
Zero-Shot Sim-to-Real Transfer from CAMERA25 (6 classes)					
SecondPose (Regression Baseline) [9]	94.1	13.9	0.15	0.00	0.00
Diff9D (Diffusion Baseline) [39]	0.72	0.45	0.08	0.02	0.03
Standard FM (Euclidean)	0.17	0.12	0.02	0.01	0.01
FM + $SO(3)$ (Geometric)	40.16	5.75	0.77	0.59	0.73
G2FM (Ours)	**52.91**	**18.94**	**5.45**	**5.85**	**13.87**

†: Cited from original papers.

As shown in Tab. 3, G2FM achieves robust generalization (52.91% in $3D_{25}$), whereas Diff9D shows significant degradation under this extreme domain shift and produces unstable geometries (0.72%). This confirms our $SO(3)$ geodesic flow prevents the geometric degradation inherent to Euclidean diffusion methods. While the regression baseline SecondPose [9] achieves nontrivial accuracy ($3D_{50}$ of 13.9%), its slow inference (16.4 FPS, Tab. 4) limits real-time utility. Operating at 31.5 FPS, G2FM establishes a stronger accuracy-efficiency trade-off.

Efficiency Analysis We compare inference speed in Tab. 4. G2FM reaches 31.5 FPS with only 3 ODE steps, outperforming Diff9D and DiffusionNOCS because flow matching learns a direct ODE trajectory rather than a long denoising chain. This makes G2FM practical for latency-sensitive robotic grasping.

Table 4: Comparison of inference speed (FPS).

Method	**G2FM (Ours)**	Diff9D [39]	SecondPose [9]	DiffusionNOCS [25]
FPS	**31.5**	19.2	16.4	6.77

4.3 Ablation Studies

Core Components (Tab. 3) The zero-shot setting in Tab. 3 also serves as a stress test for the core components. In the Standard FM baseline, the pose is represented by the same 12D vector used by our network (3D scale, 3D translation, and 6D rotation), but the rotation path and update are Euclidean rather than geodesic. This baseline nearly collapses ($3D_{50}$ of 0.12%). Explicitly modeling the $SO(3)$ path and update yields a large jump to 5.75%, while the full G2FM with geometric guidance reaches 18.94%. These results show that 6D rotation parameterization alone is not sufficient for flow matching. The learned trajectory and intermediate ODE states also need to respect the $SO(3)$ geometry.

Rotation Path Ablation Table 5 further isolates whether the improvement comes from a continuous rotation representation alone or from the manifold-aware path and update. All variants use the same backbone and training data. The Euclidean 6D-FM variant keeps the 6D rotation representation but learns the flow in Euclidean space. The matrix-FM + SVD variant predicts a matrix trajectory and projects the endpoint to $SO(3)$ using SVD. G2FM instead predicts Lie-algebra velocity and updates rotations with exp/log maps along the $SO(3)$ geodesic.

Table 5: Rotation-path ablation on REAL275. All variants use the same training data, backbone, and comparable ODE setting, while changing only the rotation path and update.

Method	$3D_{50}$	$3D_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
Euclidean 6D-FM	76.6	38.2	**34.7**	**44.5**	54.0	**71.2**
Matrix-FM + SVD projection	72.0	21.3	19.9	38.9	30.6	66.6
G2FM with $SO(3)$ geodesic update	**79.4**	**42.6**	33.9	39.2	**55.3**	67.2

The results show a clear trade-off across metrics. The $SO(3)$ geodesic update gives the best IoU-oriented accuracy and the best 10°2cm result, indicating better global pose and size consistency. Euclidean 6D-FM remains competitive on

several rotation-translation thresholds, but this does not contradict the larger-domain Omni6DPose stress test, where the Euclidean path nearly collapses. Matrix-FM with SVD projection performs worse than both, suggesting that endpoint projection alone is not enough to replace a manifold-aware update.

Encoder Ablation Table 6 examines whether stronger feature extractors directly improve G2FM. The original model uses ResNet18 (R18) for RGB features and PointNet (PN) for point-cloud features. We compare it with variants that replace the RGB encoder with DINOv2 [48] or SigLIP2 [60], or replace the point encoder with DGCNN [65]. All variants keep the same G2FM solver, training data, and loss, so the comparison focuses on the encoder choice.

Table 6: Encoder ablation on REAL275. All variants keep the same G2FM solver, training data, and loss. R18 denotes ResNet18 and PN denotes PointNet. "w/" means replacing the corresponding encoder in G2FM, for example, w/ DINOv2 replaces R18 with DINOv2 while keeping PN.

Method	RGB Encoder	Point Encoder	$3D_{50}$	$3D_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
G2FM	R18	PN	**80.0**	**44.8**	**36.5**	**41.2**	**57.5**	**67.6**
w/ DINOv2	DINOv2	PN	74.9	42.2	32.9	36.3	56.8	65.2
w/ SigLIP2	SigLIP2	PN	74.9	42.0	35.3	39.0	52.0	62.0
w/ DGCNN	R18	DGCNN	61.2	32.1	35.0	**41.2**	47.8	59.8

The completed encoder ablation shows that replacing the RGB backbone with DINOv2 or SigLIP2 does not surpass the original G2FM. This is reasonable for 9-DoF pose estimation. DINOv2 and SigLIP2 provide strong semantic RGB features, but the task here depends heavily on fine object geometry, depth-aligned cues, and stable pose-space updates. Replacing PointNet with DGCNN also gives mixed results: it stays close on the stricter rotation-translation metrics, but drops on the IoU-oriented metrics that depend more on the full 9-DoF pose and size. Under the same solver, data, and loss, the original ResNet18+PointNet design remains a better fit for the current RGB-D pipeline. The result supports our choice to keep the backbone fixed in the main comparisons, since the main gain of G2FM comes from geodesic flow modeling and geometric guidance rather than from a stronger visual encoder. It also suggests that foundation RGB features may need task-specific adapters, spatial token fusion, or partial fine-tuning before they can consistently improve category-level 9-DoF pose estimation.

Generality of Geometric Guidance (Tab. 7) We integrate our plug-and-play guidance into Diff9D without retraining. It improves Diff9D on REAL275 and Omni6DPose, even though Omni6DPose lacks canonical models and must use CAMERA25 source models. On Wild6D, severe depth noise slightly hurts $3D_{50}$, showing that stable trajectories are important for robust geometric guidance.

Table 7: Guidance generalizes to Diff9D.

Method	Omni6DPose			REAL275		Wild6D	
	$3D_{25}$(%)	$3D_{50}$(%)	$3D_{75}$(%)	$3D_{50}$(%)	$3D_{75}$(%)	$3D_{50}$(%)	$3D_{75}$(%)
Diff9D	0.72	0.45	0.08	76.5	41.7	76.48	38.16
Diff9D + **Guidance**	**8.45** (+7.73)	**2.83** (+2.38)	**1.41** (+1.33)	**77.4** (+0.9)	**42.3** (+0.6)	74.36 (-2.1)	**38.26** (+0.1)
G2FM (Ours)	**52.91**	**18.94**	**5.45**	**80.0**	**44.8**	**77.54**	36.20

Inference Settings Supplementary experiments show that 3 ODE steps and guidance strength $\gamma = 0.01$ give the best speed-accuracy balance.

5 Conclusion

We propose G2FM, a continuous flow matching framework for category-level 9-DoF pose estimation. By learning a direct vector field and modeling rotations through $SO(3)$ geodesic paths, G2FM achieves a strong accuracy-efficiency trade-off on real-world benchmarks. Training-free geometric guidance further improves pose refinement without retraining. The experiments show that the main benefit comes from respecting the pose geometry during the flow trajectory, rather than merely changing the final rotation representation. Together with the zero-shot Omni6DPose results, this suggests that geodesic flow modeling is a useful direction for efficient and generalizable RGB-D pose estimation.

6 Limitations and Future Work

Several limitations remain. Severe occlusion, noisy depth, or imperfect segmentation can weaken geometric guidance, motivating robust or confidence-weighted guidance. For symmetric objects, Chamfer guidance tolerates geometry-equivalent poses, while symmetry-aware velocity losses remain valuable future work. G2FM also relies on an upstream segmenter, and end-to-end segmentation-pose generation remains future work.

Acknowledgements

This work was funded by the National Natural Science Foundation of China U23A20385 and 62273139. The authors would like to thank the Editor and anonymous reviewers for their valuable comments.

References

1. An, B., Geng, Y., Chen, K., Li, X., Dou, Q., Dong, H.: Rgbmanip: Monocular image-based robotic manipulation through active object pose estimation. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 7748–7755 (2024). https://doi.org/10.1109/ICRA57147.2024.10610690
2. Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Nießner, M.: Scan2cad: Learning cad model alignment in rgb-d scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2614–2623 (2019)
3. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
4. Caraffa, A., Boscaini, D., Hamza, A., Poiesi, F.: Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 414–431. Springer Nature Switzerland, Cham (2025)
5. Chen, H., Wang, P., Wang, F., Tian, W., Xiong, L., Li, H.: Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2781–2790 (June 2022)
6. Chen, J., Sun, M., Zheng, Y., Bao, T., He, Z., Li, D., Jin, G., Rui, Z., Wu, L., Jiang, X.: Geo6d: Geometric-constraints-guided direct object 6d pose estimation network. IEEE Transactions on Multimedia **27**, 5770–5783 (2025). https://doi.org/10.1109/TMM.2025.3543083
7. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2773–2782 (October 2021)
8. Chen, R.T.Q., Lipman, Y.: Flow matching on general geometries. In: The Twelfth International Conference on Learning Representations (2024), https://openreview.net/forum?id=g7ohDlTITL, accessed 28 Jun 2026
9. Chen, Y., Di, Y., Zhai, G., Manhardt, F., Zhang, C., Zhang, R., Tombari, F., Navab, N., Busam, B.: Secondpose: Se(3)-consistent dual-stream feature fusion for category-level pose estimation. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9959–9969 (2024). https://doi.org/10.1109/CVPR52733.2024.00950
10. Corsetti, J., Boscaini, D., Giuliari, F., Oh, C., Cavallaro, A., Poiesi, F.: High-resolution open-vocabulary object 6d pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence pp. 1–12 (2025). https://doi.org/10.1109/TPAMI.2025.3624589
11. Corsetti, J., Boscaini, D., Oh, C., Cavallaro, A., Poiesi, F.: Open-vocabulary object 6d pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18071–18080 (June 2024)
12. Di, Y., Zhang, R., Lou, Z., Manhardt, F., Ji, X., Navab, N., Tombari, F.: Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6781–6791 (June 2022)
13. Engelmann, F., Rematas, K., Leibe, B., Ferrari, V.: From points to multi-object 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4588–4597 (2021)

14. Felice, F.D., Remus, A., Gasperini, S., Busam, B., Ott, L., Thalhammer, S., Tombari, F., Avizzano, C.A.: Instantpose: Zero-shot instance-level 6d pose estimation from a single view. IEEE Robotics and Automation Letters **10**(6), 6023–6030 (2025). https://doi.org/10.1109/LRA.2025.3562788
15. Fischer, T., Zhang, X., Ilg, E.: Unified category-level object detection and pose estimation from rgb images using 3d prototypes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9790–9800 (October 2025)
16. Fu, H., Cai, B., Gao, L., Zhang, L.X., Wang, C., Zeng, Q., Sun, C., Fei, Y., Zheng, Y., Li, Y., Liu, Y.: 3d-front: 3d furnished rooms with layouts and semantics. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) pp. 10933–10942 (2021)
17. Fu, Y., Wang, X.: Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=FgDzS8_Fz7c, accessed 28 Jun 2026
18. Funk, N., Urain, J., Carvalho, J., Prasad, V., Chalvatzaki, G., Peters, J.: Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching. CoRR **abs/2409.04576** (2024), https://doi.org/10.48550/arXiv.2409.04576, accessed 28 Jun 2026
19. Gao, D., Rozenberszki, D., Leutenegger, S., Dai, A.: Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. ACM Transactions on Graphics **43**(4), 106 (2024). https://doi.org/10.1145/3658236
20. Guerreiro, J.J.A., Inoue, N., Masui, K., Otani, M., Nakayama, H.: Layoutflow: Flow matching for layout generation. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 56–72. Springer Nature Switzerland, Cham (2025)
21. Gümeli, C., Dai, A., Nießner, M.: Roca: Robust cad model retrieval and alignment from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4012–4021 (2022). https://doi.org/10.1109/CVPR52688.2022.00399
22. He, K., Gkioxari, G., Dollar, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
23. 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 (CVPR). pp. 3003–3013 (June 2021)
24. Huang, J., Yu, H., Yu, K.T., Navab, N., Ilic, S., Busam, B.: 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 (CVPR). pp. 10095–10105 (June 2024)
25. Ikeda, T., Zakharov, S., Ko, T., Irshad, M.Z., Lee, R., Liu, K., Ambrus, R., Nishiwaki, K.: Diffusionnocs: Managing symmetry and uncertainty in sim2real multi-modal category-level pose estimation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7406–7413 (2024). https://doi.org/10.1109/IROS58592.2024.10802487
26. Klein, L., Krämer, A., Noe, F.: Equivariant flow matching. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 59886–59910. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/bc827452450356f9f558f4e4568d553b-Paper-Conference.pdf, accessed 28 Jun 2026

27. Krishnan, A., Murez, Z., Li, Z., Fidler, S.: Omninocs: A unified nocs dataset and model for 3d lifting of 2d objects. In: European Conference on Computer Vision (ECCV) (2024)
28. Langer, F., Bae, G., Budvytis, I., Cipolla, R.: Sparc: Sparse render-and-compare for cad model alignment in a single rgb image (2022)
29. Lee, S., Kim, T.K.: Joint learning of pose regression and denoising diffusion with score scaling sampling for category-level 6d pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 5757–5768 (October 2025)
30. Lee, T., Lee, B.U., Shin, I., Choe, J., Shin, U., Kweon, I.S., Yoon, K.J.: Uda-cope: Unsupervised domain adaptation for category-level object pose estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 14871–14880 (2022). https://doi.org/10.1109/CVPR52688.2022.01447
31. 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: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21285–21295 (2023). https://doi.org/10.1109/CVPR52729.2023.02039
32. Lee, T., Wen, B., Kang, M., Kang, G., Kweon, I.S., Yoon, K.J.: Any6d: Model-free 6d pose estimation of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11633–11643 (June 2025)
33. Lin, H., Liu, Z., Cheang, C., Fu, Y., Guo, G., Xue, X.: Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6697–6707 (2022). https://doi.org/10.1109/CVPR52688.2022.00659
34. Lin, J., Wei, Z., Ding, C., Jia, K.: Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 19–34. Springer Nature Switzerland, Cham (2022)
35. Lin, J., Wei, Z., Zhang, Y., Jia, K.: Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13955–13965 (2023). https://doi.org/10.1109/ICCV51070.2023.01287
36. Lin, Y., Yao, Y., Zhou, Y., Liu, T.: Beyond optimal transport: Model-aligned coupling for flow matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings. pp. 3955–3964 (June 2026)
37. Lipman, Y., Chen, R.T.Q., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (ICLR) (2023), https://openreview.net/forum?id=PqvMRDCJT9t, accessed 28 Jun 2026
38. Liu, J., Sun, W., Liu, C., Zhang, X., Fu, Q.: 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). https://doi.org/10.1109/TII.2023.3244348
39. Liu, J., Sun, W., Yang, H., Deng, P., Liu, C., Sebe, N., Rahmani, H., Mian, A.: Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(7), 5520–5537 (2025). https://doi.org/10.1109/TPAMI.2025.3552132

40. Liu, J., Sun, W., Yang, H., Zeng, Z., Liu, C., Zheng, J., Liu, X., Rahmani, H., Sebe, N., Mian, A.: Deep learning-based object pose estimation: A comprehensive survey. arXiv preprint arXiv:2405.07801 (2024)
41. Liu, M., Li, S., Chhatkuli, A., Truong, P., Van Gool, L., Tombari, F.: One2any: One-reference 6d pose estimation for any object. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6457–6467 (June 2025)
42. Liu, X., Gong, C., qiang liu: Flow straight and fast: Learning to generate and transfer data with rectified flow. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=XVjTT1nw5z, accessed 28 Jun 2026
43. Liu, Y., Jiang, Z., Xu, B., Wu, G., Ren, Y., Cao, T., Liu, B., Yang, R.H., Rasouli, A., Shan, J.: Hippo: Harnessing image-to-3d priors for model-free zero-shot 6d pose estimation. IEEE Robotics and Automation Letters **10**(8), 8284–8291 (2025). https://doi.org/10.1109/LRA.2025.3585384
44. Maninis, K.K., Popov, S., Nießner, M., Ferrari, V.: Vid2cad: Cad model alignment using multi-view constraints from videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2023)
45. Marchand, E., Uchiyama, H., Spindler, F.: Pose estimation for augmented reality: A hands-on survey. IEEE TVCG **22**(12), 2633–2651 (2016). https://doi.org/10.1109/TVCG.2015.2513408
46. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.J.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 55–64 (2020)
47. Oh, H., Choi, S., Baek, J., Kim, D.J., Joung, J.: Flawmatch: Conditional defect image generation via flow matching for improved surface defect classification. Advanced Engineering Informatics **68**, 103704 (2025). https://doi.org/10.1016/j.aei.2025.103704
48. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024)
49. Ouyang, W., Ye, Q., Wang, J., Xu, Z., Chen, J.: RFMPose: Generative category-level object pose estimation via riemannian flow matching. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems (2025), https://openreview.net/forum?id=6aaixHco6C, accessed 28 Jun 2026
50. Park, K., Patten, T., Vincze, M.: Pix2pose: Pixel-wise coordinate regression of objects given 2d bounding boxes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 7668–7677 (2019)
51. Paschalidou, D., Kar, A., Shugrina, M., Kreis, K., Geiger, A., Fidler, S.: Atiss: Autoregressive transformers for indoor scene synthesis. In: Advances in Neural Information Processing Systems (NeurIPS) (2021)
52. Peng, W., Yan, J., Wen, H., Sun, Y.: Self-supervised category-level 6d object pose estimation with deep implicit shape representation. Proceedings of the AAAI Conference on Artificial Intelligence **36**(2), 2082–2090 (Jun 2022). https://doi.org/10.1609/aaai.v36i2.20104

53. Ren, H., Yang, W., Zhang, S., Zhang, T.: Rethinking correspondence-based category-level object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1170–1179 (June 2025)
54. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2021)
55. Simeonov, A., Du, Y., Tagliasacchi, A., Tenenbaum, J.B., Rodriguez, A., Agrawal, P., Sitzmann, V.: Neural descriptor fields: Se(3)-equivariant object representations for manipulation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2022)
56. Sundermeyer, M., Marton, Z.C., Durner, M., Brucker, M., Triebel, R.: Implicit 3d orientation learning for 6d object detection from rgb images. In: Computer Vision–ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI 15. pp. 699–715. Springer (2018)
57. Tang, J., Nie, Y., Markhasin, L., Dai, A., Thies, J., Nießner, M.: Diffuscene: Denoising diffusion models for generative indoor scene synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20507–20518 (2024)
58. Tang, Y., Cao, Z., Guan, P., Gong, X., Wang, M., Yu, J.: Diff-cope: Diffusion-based category-level 6d object pose estimation. IEEE Transactions on Circuits and Systems for Video Technology pp. 1–1 (2025). https://doi.org/10.1109/TCSVT.2025.3613733
59. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds.) Computer Vision – ECCV 2020. pp. 530–546. Springer International Publishing, Cham (2020)
60. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., Hénaff, O., Harmsen, J., Steiner, A., Zhai, X.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features (2025)
61. Wang, C., Xu, D., Zhu, Y., Martin-Martin, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
62. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16611–16621 (June 2021)
63. 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 (CVPR) (June 2019)
64. Wang, J., Chen, K., Dou, Q.: Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 4807–4814 (2021). https://doi.org/10.1109/IROS51168.2021.9636212
65. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics **38**(5), 146 (2019)

66. Wei, J., Song, X., Liu, W., Kneip, L., Li, H., Ji, P.: Rgb-based category-level object pose estimation via decoupled metric scale recovery. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 2036–2042 (2024). https://doi.org/10.1109/ICRA57147.2024.10611723
67. 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 (CVPR). pp. 17868–17879 (June 2024)
68. Wu, D., Zhuang, Z., Xiang, C., Zou, W., Li, X.: 6d-vnet: End-to-end 6-dof vehicle pose estimation from monocular rgb images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (June 2019)
69. Xu, L., Qu, H., Cai, Y., Liu, J.: 6d-diff: A keypoint diffusion framework for 6d object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9676–9686 (June 2024)
70. Yazdani, M., Medghalchi, Y., Ashrafian, P., Hacihaliloglu, I., Shahriari, D.: Flow matching for medical image synthesis: Bridging the gap between speed and quality. In: Gee, J.C., Alexander, D.C., Hong, J., Iglesias, J.E., Sudre, C.H., Venkataraman, A., Golland, P., Kim, J.H., Park, J. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2025. pp. 216–226. Springer Nature Switzerland, Cham (2026)
71. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
72. Yim, J., Campbell, A., Foong, A.Y.K., Gastegger, M., Jimenez-Luna, J., Lewis, S., Satorras, V.G., Veeling, B.S., Barzilay, R., Jaakkola, T., Noe, F.: Fast protein backbone generation with se(3) flow matching (2023), https://arxiv.org/abs/2310.05297, accessed 28 Jun 2026
73. You, Y., Shi, R., Wang, W., Lu, C.: Cppf: Towards robust category-level 9d pose estimation in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6866–6875 (June 2022)
74. Yu, Q., Hao, C., Wang, J., Liu, W., Liu, L., Mu, Y., You, Y., Yan, H., Lu, C.: Manipose: A comprehensive benchmark for pose-aware object manipulation in robotics. CoRR **abs/2403.13365** (2024), https://doi.org/10.48550/arXiv.2403.13365, accessed 28 Jun 2026
75. Yu, S., Zhai, D.H., Xia, Y.: Catformer: Category-level 6d object pose estimation with transformer. Proceedings of the AAAI Conference on Artificial Intelligence **38**(7), 6808–6816 (Mar 2024). https://doi.org/10.1609/aaai.v38i7.28505
76. Zhang, C., Ling, Y., Lu, M., Qin, M., Wang, H.: Category-level object detection, pose estimation and reconstruction from stereo images. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) Computer Vision – ECCV 2024. pp. 332–349. Springer Nature Switzerland, Cham (2025)
77. Zhang, J., Huang, W., Peng, B., Wu, M., Hu, F., Chen, Z., Zhao, B., Dong, H.: Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In: European Conference on Computer Vision. pp. 199–216. Springer (2024)
78. Zhang, J., Wu, M., Dong, H.: Generative category-level object pose estimation via diffusion models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 54627–54644. Curran Associates, Inc. (2023), https://proceedings.neurips.cc/

paper_files/paper/2023/file/ab59d149fc0c2c9039d3e3049f7914b1-Paper-Conference.pdf, accessed 28 Jun 2026

79. Zhang, Q., Liu, Z., Fan, H., Liu, G., Zeng, B., Liu, S.: Flowpolicy: Enabling fast and robust 3d flow-based policy via consistency flow matching for robot manipulation. Proceedings of the AAAI Conference on Artificial Intelligence **39**(14), 14754–14762 (Apr 2025). https://doi.org/10.1609/aaai.v39i14.33617
80. Zhang, S., Huang, Y., Zhao, W., Zhao, W., Guan, Z., Peng, J.: Environment-agnostic pose: Generating environment-independent object representations for 6d pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 8678–8687 (October 2025)
81. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5745–5753 (2019)