NOCSFLOW: LEARNING DENSE CANONICAL FIELDS FOR MONOCULAR CATEGORY-LEVEL 9D POSE ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Category-level monocular 9D pose estimation requires recovering object geometry from RGB, given a predicted mask and camera intrinsics but no external 3D observations. Dense canonical representations such as Normalized Object Coordinate Space (NOCS) are promising for this problem, as they preserve rich pixel-wise 2D–3D correspondence structure for downstream pose recovery, yet learning such representations directly from monocular RGB remains underexplored. In this work, we introduce **NOCSFlow**, a pixel-space flow-matching framework that learns dense NOCS representations by injecting features from a pretrained geometry encoder as implicit geometric conditioning into multiple transformer layers. We further introduce coupled supervision that jointly optimizes the generated canonical fields with direct canonical-space supervision and pose-space supervision backpropagated through a differentiable PnP module. Experiments on CAMERA25 and REAL275 demonstrate substantial improvements over prior methods, including 96.8% on the 10° rotation metric and 85.1% on the joint $(10^\circ, 10\,\mathrm{cm})$ metric on REAL275.

1 INTRODUCTION

Object pose estimation is a fundamental problem in computer vision, with broad applications in robotics, augmented reality, and embodied perception (Tremblay et al., 2018; Xiang et al., 2018). Category-level monocular pose estimation is particularly attractive for real-world deployment because it does not require known object instances, instance-specific CAD models, or multi-view observations at inference (Fan et al., 2022; Liu et al., 2025). However, this setting is also highly challenging: the model must generalize across substantial intra-class shape variation while recovering object geometry and pose from a single RGB observation without externally provided geometric information such as depth, normal maps, or point clouds (Fan et al., 2022; Manhardt et al., 2019).

One line of existing approaches encodes the input observation into a compact feature representation and directly predicts the object pose parameters, for example by iteratively denoising a pose vector (Liu et al., 2025). While this formulation provides a simple and direct solution for pose estimation, it does not explicitly model the dense correspondences between image observations and object geometry that can provide useful spatial constraints for pose recovery. An alternative line of work builds on the *Normalized Object Coordinate Space* (NOCS), which represents each object instance in a shared canonical coordinate system (Wang et al., 2019b). By assigning each observed object pixel a canonical 3D coordinate, NOCS establishes dense image-to-canonical correspondences and provides a structured representation of object geometry for subsequent pose recovery (Ikeda et al., 2024; Krishnan et al., 2024).

Despite this advantage, learning accurate dense canonical representations from monocular RGB remains challenging. A NOCS map requires predicting a 3D canonical coordinate for each observed pixel, while a single RGB image provides no direct geometric measurement to constrain these predictions. Existing NOCS-based pipelines therefore often rely on depth or multi-modal observations for canonical prediction and pose recovery (Wang et al., 2019b; Ikeda et al., 2024). As a result, recovering accurate and pose-consistent canonical representations from RGB alone, without explicit geometric observations at inference, remains a challenging problem.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

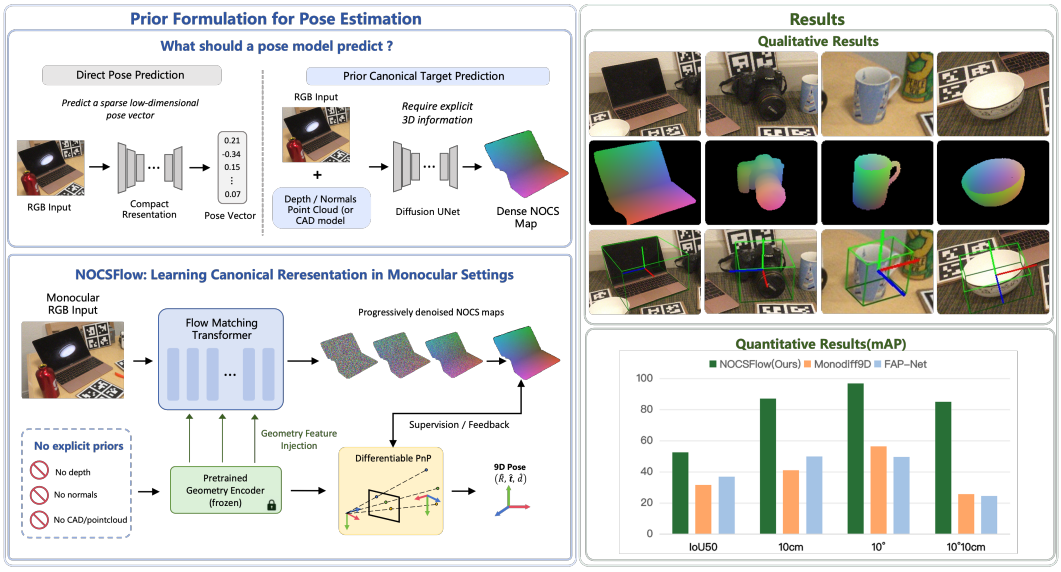

Figure 1: **Dense canonical representation learning from monocular RGB. Top left:** Pose-vector prediction and canonical prediction with explicit geometric inputs. **Bottom left:** NOCSFlow combines flow matching, pretrained geometry features, and differentiable PnP supervision, without external depth, point clouds, or CAD models at inference. **Right:** Predicted NOCS maps, recovered poses, and quantitative comparisons on REAL275.

We therefore study dense canonical representation learning under the monocular setting. The key challenge is how to learn an accurate canonical representation from monocular observations when explicit and accurate 3D information, such as ground truth depth or point clouds, is unavailable. Such a representation should not only recover the underlying object geometry from visual evidence, but also preserve the correspondence structure required for downstream pose estimation.

To this end, we introduce **NOCSFlow**, a pixel-space flow-matching framework for learning dense Normalized Object Coordinate Space (NOCS) representations from monocular RGB images, given a predicted target mask and known camera intrinsics. To better learn the canonical representation, we employ a pretrained geometry encoder to extract dense geometric features and inject them as implicit geometric conditioning into multiple transformer layers of the flow-matching generator. We further introduce coupled supervision to make the learned canonical representation better aligned with the downstream pose estimation task. Specifically, the generated NOCS field is jointly optimized by direct canonical-space supervision and pose-space supervision backpropagated through a differentiable PnP module. Together, these designs encourage NOCSFlow to learn dense canonical fields that are both geometrically informative and directly optimized for reliable pose recovery.

Experiments on CAMERA25 and REAL275 (Wang et al., 2019b) demonstrate the effectiveness of this formulation for category-level monocular 9D pose estimation. On REAL275, NOCSFlow achieves 96.8% on the 10° rotation metric and 85.1% on the joint $(10^\circ, 10\,\mathrm{cm})$ metric, improving the best prior results DMSR (Wei et al., 2024) and GIVEPose (Huang et al., 2025), respectively—by 37.3 and 50.9 percentage points. These results demonstrate the effectiveness of combining implicit geometric guidance with pose-level feedback for learning dense canonical representations from monocular observations.

2 Related Work

2.1 Category-Level Pose Estimation

Object pose estimation has progressed from instance-level settings to category-level scenarios. Instance-level methods Peng et al. (2019); He et al. (2020); Wang et al. (2019a) assume known object identities and accurate CAD models, achieving strong performance on seen instances but

limited generalization to unseen objects. Category-level pose estimation instead targets unseen instances within a semantic category, where intra-class shape variation makes the problem more practical yet more challenging. Existing methods mainly follow two paradigms. Direct regression predicts pose parameters from image features Irshad et al. (2022); Lee et al. (2025); Yang et al. (2025); Xiang et al. (2018), but remains vulnerable to depth–scale ambiguity in monocular settings. Canonical-representation-based methods, represented by NOCS Wang et al. (2019b), estimate dense correspondences in normalized object coordinates and recover pose through geometric alignment Chen & Dou (2021); Liu et al. (2022); Zheng et al. (2024); Lin et al. (2021); Liu et al. (2023); Lin et al. (2023). The final pose is typically obtained by PnP-based optimization or related solvers Umeyama (2002); Lepetit et al. (2009). This formulation preserves dense geometric structure and the solver is often used as post-processing.

2.2 Geometric Representations and Generative Pose Estimation

Geometric representations provide useful constraints for pose recovery. Monocular depth models (Yang et al., 2024) supply inferred geometry, while pose methods reconstruct depth, estimate local geometry, or jointly recover object shape (Fan et al., 2022; Wei et al., 2024; Irshad et al., 2022). Category-level shape priors and geometric alignment further address intra-class variation (Chen & Dou, 2021; Zheng et al., 2024; Huang et al., 2025; Tian et al., 2020). Complementary work reduces dependence on explicit shape priors (Lin et al., 2021; Liu et al., 2023; Lin et al., 2023) or explores robustness, domain adaptation, and learning with limited real-world pose annotations (You et al., 2022; Lee et al., 2022; Fu & Wang, 2022; Wang et al., 2020). These directions emphasize the importance of geometric representations and supervision for reliable pose estimation.

Pose-hypothesis reasoning ranges from object-pose scoring (Okorn et al., 2021) to diffusion-based camera pose estimation (Wang et al., 2023). For category-level objects, GenPose (Zhang et al., 2023) and MonoDiff9D (Liu et al., 2025) generate low-dimensional pose vectors, with the latter also incorporating reconstructed point clouds. DiffusionNOCS (Ikeda et al., 2024) instead generates dense canonical maps, but uses observed depth for subsequent pose recovery. Building on dense canonical representations, NOCSFlow learns NOCS fields from monocular RGB through implicit geometric conditioning and coupled canonical-space and pose-space supervision, aligning field generation with downstream pose recovery.

3 Method

3.1 Overview

Given a full RGB image $I \in \mathbb{R}^{3\times H\times W}$, a predicted object mask $M \in \{0,1\}^{1\times H\times W}$, and camera intrinsics $K \in \mathbb{R}^{3\times 3}$, our goal is to estimate the category-level 9D pose $(\hat{R}, \hat{\boldsymbol{t}}, \hat{d})$, where $\hat{R} \in SO(3)$, $\hat{\boldsymbol{t}} \in \mathbb{R}^3$, and $\hat{d} \in \mathbb{R}^3_{>0}$ denote rotation, translation, and metric size, respectively. We use t for flow time and $\boldsymbol{t}$ for translation.

Instead of directly regressing pose parameters, NOCSFlow factorizes pose estimation into dense canonical-field generation and geometric pose decoding. It first generates a pixel-wise Normalized Object Coordinate Space (NOCS) field from the RGB image and object mask, without using depth, point clouds, or category-level CAD models as explicit conditioning inputs. The generated field establishes dense image-to-object correspondences, which are metrically lifted and decoded into pose by a differentiable weighted PnP solver.

The key to our formulation is that the solver is not merely used as post-processing. During training, each sampled clean-field prediction receives two coupled supervision signals: a direct canonical-space objective that constrains individual NOCS coordinates, and a solver-induced pose-space objective that measures whether the predicted correspondences jointly recover the correct pose. Both signals are backpropagated into the same field-generation process.

3.2 Canonical Field Generation without Explicit 3D Conditioning

Implicit geometric conditioning. Generating a dense canonical field from a single RGB image is challenging because the observation does not explicitly provide the 3D structure required to establish

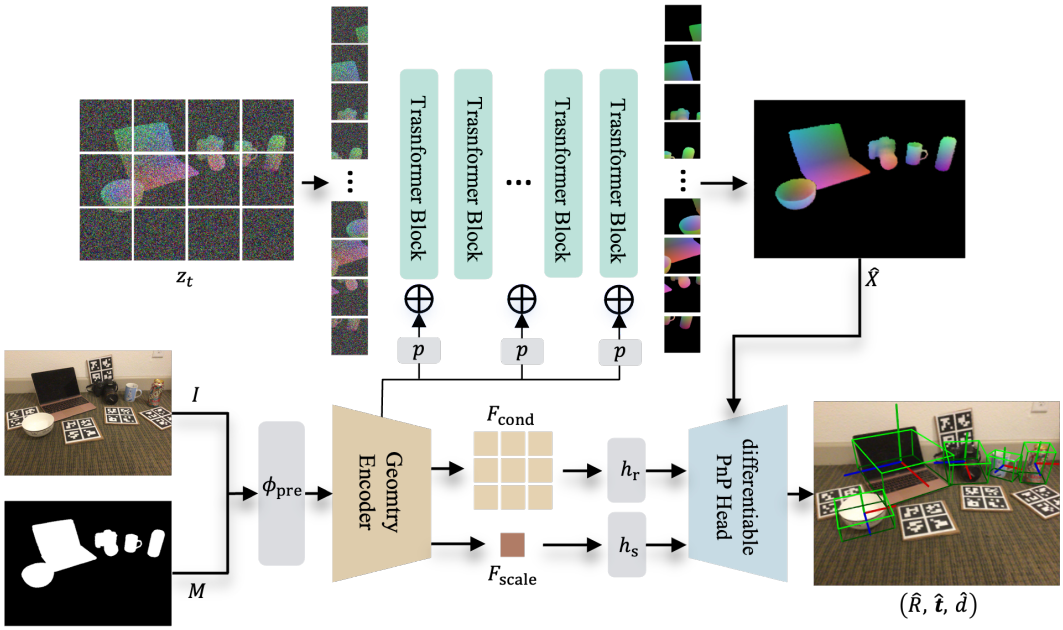

Figure 2: **Overview of NOCSFlow.** Given an RGB image I and object mask M, a trainable pre-encoder adapts the image–mask pair to a frozen monocular geometry encoder, which extracts dense conditioning features F_{cond} and a global scale feature F_{scale}. A pixel-space flow generator produces an intermediate clean NOCS field $\hat{X}_t$, while lightweight heads estimate metric scale, size ratio, and correspondence confidence. During training, the same field prediction is optimized by direct canonical-field supervision and pose supervision backpropagated through a differentiable weighted PnP solver. At inference, the terminal NOCS field is metrically lifted and decoded into the 9D pose $(\hat{R}, \hat{\boldsymbol{t}}, \hat{d})$.

reliable image-to-canonical correspondences. We therefore condition the generation process on implicit geometric features extracted by a pretrained monocular geometry encoder from MoGe2 Wang et al. (2025b), without relying on explicit object-level 3D inputs.

We concatenate the RGB image and object mask and map the resulting 4-channel tensor to the 3-channel input space of the pretrained encoder:

$$I_c = \phi_{\text{pre}}([I, M]), \qquad (F_{\text{cond}}, F_{\text{scale}}) = E(I_c), \tag{1}$$

where ϕ_{pre} is a trainable convolutional pre-encoder and E is a frozen MoGe-2 encoder Wang et al. (2025b). Here E includes the encoder-specific feature interface; $F_{\text{cond}} \in \mathbb{R}^{N \times D}$ denotes its spatial features aligned to the generator patch grid, while the global feature $F_{\text{scale}} \in \mathbb{R}^{D}$ is used for metric-scale prediction. The geometry encoder is evaluated once for each image–mask pair, and its features are reused throughout the generation trajectory.

Geometry-conditioned pixel-space flow. We use a pixel-space transformer generator with DiT-style blocks Peebles & Xie (2023) and clean-target prediction motivated by JiT Li & He (2026). The generator operates on patchified NOCS states, with flow time embedded through adaptive layer normalization. To preserve spatial correspondence between image evidence and canonical coordinates, we inject geometry features into a selected set of transformer blocks $\mathcal{S}$:

$$\bar{H}_t^{(l)} = \begin{cases} H_t^{(l)} + p^{(l)}(F_{\text{cond}}), & l \in \mathcal{S}, \\ H_t^{(l)}, & l \notin \mathcal{S}, \end{cases} \qquad H_t^{(l+1)} = B^{(l)}\left(\bar{H}_t^{(l)}; e(t)\right), \tag{2}$$

where $p^{(l)}$ projects the geometry feature dimension to the hidden dimension of the l-th block, and $e(t)$ denotes the flow-time embedding. Spatial alignment precedes channel projection, so corresponding geometry and NOCS tokens are added (Appendix A.1).

Let $X \in [-1,1]^{3\times H\times W}$ denote the ground-truth canonical field. For an object-frame point x^{obj} centered at the 3D bounding-box center, we use $X = 2x^{\mathrm{obj}}/\|d\|_2$, where $d \in \mathbb{R}^3_{>0}$ contains the three bounding-box side lengths. For Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $t \sim \mathcal{U}(0,1)$, we construct the linear interpolation

$$z_t = tX + (1-t)\varepsilon. \tag{3}$$

Rather than predicting velocity directly, the generator predicts the clean endpoint field together with its output features:

$$(\hat{X}_t, F_t^{\mathrm{out}}) = G_\theta(z_t, t; F_{\mathrm{cond}}). \tag{4}$$

The corresponding predicted and target velocities are

$$\hat{v}_t = \frac{\hat{X}_t - z_t}{1-t}, \qquad v_t^\star = \frac{X - z_t}{1-t} = X - \varepsilon. \tag{5}$$

This clean-target parameterization is important for our coupled training formulation: $\hat{X}_t$ is already an estimate of the final canonical field at any sampled flow time, allowing pose-level supervision to be applied without unrolling the complete inference trajectory.

Metric extent prediction. The canonical field determines dense object-frame correspondences but does not by itself provide metric scale. We predict a positive global scale $\hat{s}$ from the global geometry feature and an axis-wise size ratio $\hat{r}$ from the spatial features:

$$\hat{s} = \mathrm{softplus}\left(h_s(F_{\mathrm{scale}})\right), \qquad a = \mathrm{softplus}\left(h_r(\mathrm{GAP}(F_{\mathrm{cond}}))\right), \qquad \hat{r} = \frac{a}{\|a\|_2}. \tag{6}$$

The predicted metric size is

$$\hat{d} = \hat{s}\hat{r}. \tag{7}$$

Both quantities depend only on the image features and are therefore shared across all sampled flow times.

3.3 Coupled Supervision for Canonical Field Generation

A pixel-wise NOCS objective encourages locally accurate canonical coordinates, but it does not explicitly require the full set of correspondences to be compatible with a single camera pose. Conversely, pose supervision captures global geometric consistency but does not uniquely determine every canonical coordinate. We therefore couple the two objectives on the same intermediate prediction $\hat{X}_t$.

For each sampled flow time, supervision follows two paths:

$$\hat{X}_t \longrightarrow \mathcal{L}_{\mathrm{field}}, \qquad (\hat{X}_t, F_t^{\mathrm{out}}) \longrightarrow (\hat{X}_t^m, w_t) \longrightarrow \mathrm{PnP} \longrightarrow (\hat{R}_t, \hat{\boldsymbol{t}}_t) \longrightarrow \mathcal{L}_{\mathrm{pose}}. \tag{8}$$

Canonical-space supervision. Let V denote the valid NOCS supervision mask, which selects pixels with ground-truth canonical coordinates and is distinct from the predicted input mask M. We supervise the flow velocity using

$$\mathcal{L}_{\mathrm{FM}} = \frac{1}{|V|}\sum_{u,v} V_{u,v}\ \mathrm{SmoothL1}\left(\hat{v}_{t,u,v}, v^\star_{t,u,v}\right), \qquad |V| = \sum_{u,v} V_{u,v} > 0. \tag{9}$$

We additionally regularize the coordinate magnitude of the clean-field prediction:

$$\mathcal{L}_{\mathrm{norm}} = \frac{1}{|V|}\sum_{u,v} V_{u,v}\ \mathrm{SmoothL1}\left(\|\hat{X}_{t,u,v}\|_2, \|X_{u,v}\|_2\right). \tag{10}$$

The direct canonical-field objective is

$$\mathcal{L}_{\mathrm{field}} = \mathcal{L}_{\mathrm{FM}} + \lambda_{\mathrm{norm}}\mathcal{L}_{\mathrm{norm}}. \tag{11}$$

Solver-induced pose supervision. To evaluate whether the generated field forms geometrically valid correspondences, we first lift it to metric object coordinates:

$$\hat{X}_t^m = \frac{\hat{s}}{2}\hat{X}_t. \tag{12}$$

For each object pixel $i \in \Omega_M$, the metric coordinate $\hat{X}_{t,i}^m$ is paired with its image coordinate $p_i = (u_i, v_i)^\top$. A confidence head predicts the reliability of each correspondence from the generator feature:

$$w_{t,i} = \mathrm{sigmoid}\left(h_w(F_{t,i}^{\mathrm{out}})\right). \tag{13}$$

The pose associated with the current canonical field is obtained through weighted reprojection minimization:

$$(\hat{R}_t, \hat{\boldsymbol{t}}_t) = \arg\min_{\substack{R\in SO(3),\\ \boldsymbol{t}\in\mathbb{R}^3}} \sum_{i\in\Omega_M} w_{t,i}\,\rho\Big(\Big\|\pi\Big(K(R\hat{X}_{t,i}^m + \boldsymbol{t})\Big) - p_i\Big\|_2\Big), \tag{14}$$

where π denotes perspective projection and ρ is a robust penalty. Image coordinates and K must undergo the same preprocessing transform (Appendix A.2). We implement this operation using a differentiable weighted PnP module based on EPro-PnP Chen et al. (2022).

Given ground-truth translation $\boldsymbol{t}^\star$, we define

$$\mathcal{L}_t = \mathrm{SmoothL1}\left(\hat{\boldsymbol{t}}_t, \boldsymbol{t}^\star\right). \tag{15}$$

For unit quaternions $\hat{q}_t$ and $q^\star$ corresponding to the predicted and ground-truth rotations, respectively, the rotation loss is

$$\mathcal{L}_r = \left[2\arccos\left(|\langle \hat{q}_t, q^\star\rangle|\right)\right]^2. \tag{16}$$

We combine them into the pose-space objective

$$\mathcal{L}_{\mathrm{pose}} = \lambda_t\mathcal{L}_t + \lambda_r\mathcal{L}_r. \tag{17}$$

Because the solver is differentiable, $\mathcal{L}_{\mathrm{pose}}$ is propagated through PnP, metric lifting, and correspondence weighting to the generated field and its internal features. It therefore favors fields whose correspondences are not only locally plausible, but also jointly consistent with the target pose.

Metric calibration and overall objective. Given ground-truth metric size d, we define

$$s^\star = \|d\|_2, \qquad r^\star = \frac{d}{\|d\|_2}. \tag{18}$$

The metric calibration objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{metric}} = {} & \lambda_{\mathrm{scale}}\,\mathrm{SmoothL1}\left(\hat{s}, s^\star\right) \\ & + \lambda_{\mathrm{size}}\,\mathrm{SmoothL1}\left(\hat{r}, r^\star\right).\end{aligned} \tag{19}$$

The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{field}} + \mathcal{L}_{\mathrm{pose}} + \mathcal{L}_{\mathrm{metric}}. \tag{20}$$

Inference. At inference, the geometry features and metric predictions are computed once. Starting from $z_0 \sim \mathcal{N}(0, \sigma^2 I)$, we integrate the learned velocity field using Euler updates:

$$z_{t_{k+1}} = z_{t_k} + \Delta t\,\hat{v}_{t_k}, \qquad \hat{v}_{t_k} = \frac{\hat{X}_{t_k} - z_{t_k}}{1 - t_k}. \tag{21}$$

Here $t_k = k/L$, $k = 0, \ldots, L-1$, $\Delta t = 1/L$, and $\hat{X}_{t_k}$ is the field component of Equation 4. The terminal field z_1 is metrically lifted and, with confidence weights from the last generator evaluation, passed once through the weighted PnP solver to obtain $(\hat{R}, \hat{\boldsymbol{t}})$. Together with $\hat{d} = \hat{s}\hat{r}$, they form the final 9D pose. The coupled supervision is used only during training and introduces no additional inference-time guidance or sampling trajectory.

Table 1: **Comparison with state-of-the-art methods on CAMERA25 and REAL275.** We report mAP (%) under standard 3D IoU and pose-error thresholds. In the main-text tables, darker green with bold marks the best result and lighter green marks the second best in each metric column. All entries concern monocular RGB pose methods; external-depth DiffusionNOCS is compared separately. The two available GIVEPose CAMERA25 entries are carried over from the original per-category results in Appendix C.2. Gains are discussed in percentage points.

Method	CAMERA25					REAL275				
	IoU_{50}	IoU_{75}	10cm	10°	10°10cm	IoU_{50}	IoU_{75}	10cm	10°	10°10cm
Synthesis (Chen et al., 2020)	-	-	-	-	-	-	-	34.0	14.2	4.8
MSOS (Lee et al., 2021)	32.4	5.1	29.7	60.8	19.2	23.4	3.0	39.5	29.2	9.6
CenterSnap-RGB (Irshad et al., 2022)	-	-	-	-	-	31.5	-	-	-	30.1
OLD-Net (Fan et al., 2022)	32.1	5.4	30.1	74.0	23.4	25.4	1.9	38.9	37.0	9.8
FAP-Net (Li et al., 2024)	39.2	6.7	36.0	80.4	29.8	36.8	5.2	49.7	49.6	24.5
DMSR (Wei et al., 2024)	34.6	6.5	32.3	81.4	27.4	28.3	6.1	37.3	59.5	23.6
LaPose (Zhang et al., 2024)	-	-	-	-	-	17.5	2.6	44.4	-	30.5
MonoDiff9D (Liu et al., 2025)	35.2	6.7	33.6	80.1	28.2	31.5	6.3	41.0	56.3	25.7
DA-Pose (Yang et al., 2025)	41.4	6.1	40.5	60.8	24.6	28.1	3.6	45.8	27.5	13.4
GIVEPose (Huang et al., 2025)	21.3	-	-	-	43.2	20.1	-	45.9	-	34.2
NOCSFlow (Ours)	**46.4**	**12.0**	**74.0**	**88.1**	**69.6**	**52.6**	**13.4**	**87.1**	**96.8**	**85.1**

4 Experiments

4.1 Experimental Setup

Datasets and protocol. We train on the standard CAMERA synthetic and NOCS real training splits and evaluate on CAMERA25 and REAL275 (Wang et al., 2019b), which cover bottle, bowl, camera, can, laptop, and mug. CAMERA25 contains 25K synthetic test images; REAL275 contains 2,750 real test images. Per-category results and input statistics are provided in Appendix C.

Detection and input. We use the Mask R-CNN predicted boxes and masks released by prior work. The encoder receives the full RGB image with a target-specific predicted mask, and PnP uses known camera intrinsics. Inference requires no ground-truth detections, external depth maps, or point clouds.

Metrics and implementation. We report 3D IoU mAP at IoU_{50} and IoU_{75}, translation mAP at $10\,\mathrm{cm}$, rotation mAP at $10°$, and joint pose mAP at $(10°, 10\,\mathrm{cm})$. Joint success requires both pose thresholds. The default model uses a frozen MoGe-2 encoder and 50 Euler steps with one hypothesis. Appendix A details the architecture and training settings; Appendix D reports throughput and sampling-step trade-offs.

4.2 Comparison with Prior Methods

Table 1 compares monocular methods on CAMERA25 and REAL275. On REAL275, NOCSFlow reaches 87.1% translation mAP, 96.8% rotation mAP, and 85.1% joint mAP. Relative to MonoDiff9D (Liu et al., 2025), the translation and rotation gains are 46.1 and 40.5 percentage points. On CAMERA25, joint mAP is 69.6%, 26.4 points above the listed GIVEPose result. The analyses below compare canonical prediction models and examine geometric conditioning and pose supervision.

Figure 3 shows generated canonical maps alongside recovered 3D boxes on REAL275, linking representation quality to downstream pose recovery. Table 3 compares DiffusionNOCS under its external-depth input setting.

4.3 Cross-Domain Evaluation on Wild6D

We evaluate the model trained on the CAMERA and NOCS real training splits directly on Wild6D (Fu & Wang, 2022), without additional training or fine-tuning. This tests transfer to new instances and mobile-captured scenes within five overlapping known categories (bottle, bowl, camera, laptop, and mug). Table 2 reports the aggregate metrics, and Figure 4 shows recovered 3D boxes. NOCSFlow achieves 29.5% IoU_{50} and 60.9% rotation mAP, compared with 24.1% and 40.4% for GIVEPose. The joint mAP is 58.5%, compared with 40.3% for GIVEPose. Per-category results are provided in Appendix C.3.

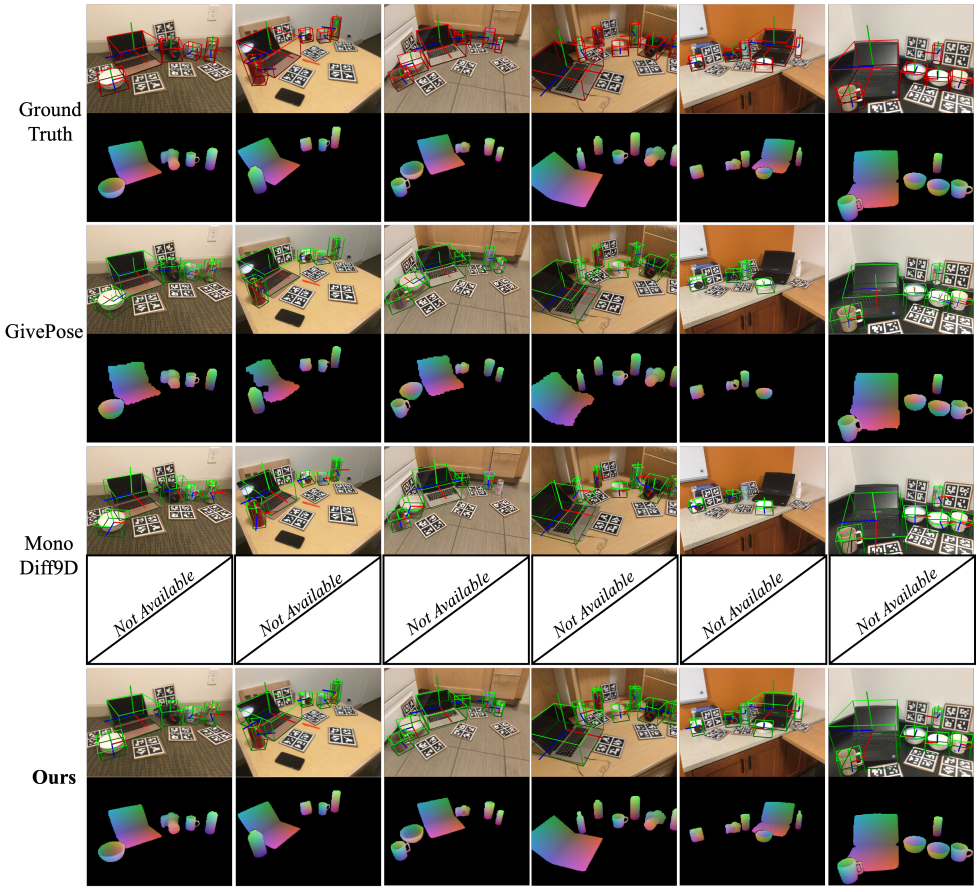

Figure 3: **Qualitative comparison on REAL275.** We show predicted 3D boxes in green, ground-truth boxes in red, and predicted NOCS maps when available. MonoDiff9D (Liu et al., 2025) does not produce NOCS maps because it directly predicts pose parameters. Compared with GIVEPose (Huang et al., 2025) and MonoDiff9D, NOCSFlow yields more coherent NOCS maps and better-aligned 3D boxes.

Table 2: Zero-shot quantitative results on Wild6D. All methods are trained on the CAMERA and NOCS real training splits and evaluated on Wild6D without fine-tuning.

Method	IoU_{50}	IoU_{75}	10cm	10°	10°10cm
MonoDiff9D Liu et al. (2025)	9.6	0.4	36.3	23.4	11.1
GIVEPose Huang et al. (2025)	24.1	**11.2**	52.3	40.4	40.3
NOCSFlow (Ours)	**29.5**	7.9	**61.1**	**60.9**	**58.5**

4.4 Analysis of Canonical Representation Learning

We compare dense generative pose models (Table 3), encoder choices (Appendix B.2), and cumulative conditioning and supervision changes (Table 4).

Comparison with DiffusionNOCS. NOCSFlow outperforms DiffusionNOCS (Ikeda et al., 2024), which also generates dense NOCS but uses observed depth and depth-derived surface normals, by 6.8 percentage points at both 5°5cm and 10°5cm on REAL275 (Table 3).

Implicit geometric conditioning. Across the tested frozen encoder configurations, VGGT and MoGe-2 average 84.2% joint mAP, compared with 39.5% averaged over MAE, CLIP, DINOv2, and DINOv3; the default MoGe-2 reaches 85.1%. Appendix B.2 reports the complete encoder/interface comparison. Within the cumulative study, multi-layer injection reduces NOCS MAE from 0.104 to

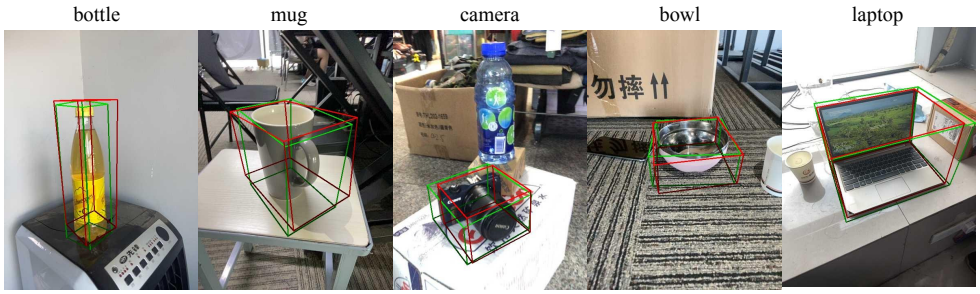

Figure 4: **Qualitative results on Wild6D (Fu & Wang, 2022).** Predicted 3D bounding boxes (green) and ground truth (red) are shown for five categories. Examples are from the model evaluated without pose-task fine-tuning on Wild6D; they illustrate recovered poses on new instances within the five known categories.

Table 3: **Dense generative pose comparison on REAL275.** Input modalities and pose recovery differ. Values are mAP (%).

Method	Geometry at inference	5°5cm	10°5cm	15°5cm
DiffusionNOCS	observed depth	35.0	66.6	77.1
NOCSFlow	RGB + predicted mask	**41.8**	**73.4**	**78.9**

Table 4: **Canonical-field and pose accuracy under cumulative design changes on REAL275.** The baseline uses the pretrained encoder, a dense NOCS target, input-level conditioning, post-hoc PnP, and NOCS-extrema-based size. MAE measures NOCS coordinate error over valid foreground pixels; remaining columns are mAP (%).

Variant	NOCS MAE ↓	IoU_{50}	IoU_{75}	10cm	10°	10°10cm
Input-level baseline	0.104	41.8	6.6	73.2	75.4	72.6
+ Multi-layer injection	0.060	48.9	9.7	80.1	82.6	79.5
+ Differentiable PnP	0.051	50.9	10.8	83.9	85.6	83.2
+ Size prediction (full)	**0.045**	**52.6**	**13.4**	**87.1**	**96.8**	**85.1**

0.060 and raises joint mAP from 72.6% to 79.5%, linking conditioning throughout the generator to more accurate canonical fields.

Coupled field and pose supervision. Adding differentiable PnP supervision reduces NOCS MAE from 0.060 to 0.051 and increases joint mAP from 79.5% to 83.2% (Table 4), indicating that pose feedback improves the generated correspondences. Adding metric-size supervision and prediction gives the final NOCS MAE of 0.045 and joint mAP of 85.1%. This branch supervises shared image features as well as metric size, allowing the additional training signal to affect both field generation and pose recovery.

Appendix B provides the NOCSformer reproduction and encoder protocols, additional conditioning and mask-placement comparisons, and sensitivity to sparsifying direct NOCS losses.

5 Conclusion

We introduced **NOCSFlow**, a framework for learning dense canonical fields for category-level monocular 9D pose estimation. Given RGB, a predicted target mask, and known camera intrinsics, it combines pixel-space flow matching with pretrained geometric conditioning and coupled canonical-field and pose supervision. The generated correspondences are metrically lifted and decoded by PnP, without externally supplied depth, point clouds, or CAD models at inference. Experiments on CAMERA25 and REAL275 demonstrate the effectiveness of this formulation, with additional transfer results on Wild6D. These results support dense canonical-field generation as a structured approach to monocular pose estimation.

AI USE STATEMENT

Generative AI tools assisted with language editing, manuscript organization, and consistency checks of mathematical notation and reported tables. They were not used to generate figures or experimental measurements. The reported experiments are the authors' work, and the authors are responsible for the final manuscript and its verification.

REFERENCES

Michael S Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. *arXiv preprint arXiv:2209.15571*, 2022.

Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. 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*, pp. 2781–2790, 2022.

Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2773–2782, 2021.

Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges. Category level object pose estimation via neural analysis-by-synthesis. In *European Conference on Computer Vision*, pp. 139–156. Springer, 2020.

Zhaoxin Fan, Zhenbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Object level depth reconstruction for category level 6d object pose estimation from monocular rgb image. In *European Conference on Computer Vision*, pp. 220–236. Springer, 2022.

Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *Advances in Neural Information Processing Systems*, 35: 27469–27483, 2022.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11632–11641, 2020.

Ziqin Huang, Gu Wang, Chenyangguang Zhang, Ruida Zhang, Xiu Li, and Xiangyang Ji. Givepose: Gradual intra-class variation elimination for rgb-based category-level object pose estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 22055–22066, 2025.

Takuya Ikeda, Sergey Zakharov, Tianyi Ko, Muhammad Zubair Irshad, Robert Lee, Katherine Liu, Rares Ambrus, and Koichi Nishiwaki. 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. IEEE, 2024.

Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 10632–10640. IEEE, 2022.

Akshay Krishnan, Abhijit Kundu, Kevis-Kokitsi Maninis, James Hays, and Matthew Brown. OmniNOCS: A unified nocs dataset and model for 3d lifting of 2d objects. In *European Conference on Computer Vision*, 2024. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/papers/09562.pdf.

Hakjin Lee, Junghoon Seo, and Jaehoon Sim. You only pose once: A minimalist's detection transformer for monocular rgb category-level 9d multi-object pose estimation. *arXiv preprint arXiv:2508.14965*, 2025.

Taeyeop Lee, Byeong-Uk Lee, Myungchul Kim, and In So Kweon. Category-level metric scale object shape and pose estimation. *IEEE Robotics and Automation Letters*, 6(4):8575–8582, 2021.

Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. Uda-cope: Unsupervised domain adaptation for category-level object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14891–14900, 2022.

Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81(2):155–166, 2009.

Jia Li, Li Jin, Xibin Song, Yeheng Chen, Nan Li, and Xueying Qin. Implicit coarse-to-fine 3d perception for category-level object pose estimation from monocular rgb image. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2043–2050. IEEE, 2024.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise, 2026. URL https://arxiv.org/abs/2511.13720.

Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3560–3569, 2021.

Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 14001–14011, 2023.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Jian Liu, Wei Sun, Hui Yang, Jin Zheng, Zichen Geng, Hossein Rahmani, and Ajmal Mian. Monodiff9d: Monocular category-level 9d object pose estimation via diffusion model. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8687–8694. IEEE, 2025.

Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13978–13988, 2023.

Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. Catre: Iterative point clouds alignment for category-level object pose refinement. In *European Conference on Computer Vision*, pp. 499–516. Springer, 2022.

Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6841–6850, 2019.

Brian Okorn, Qiao Gu, Martial Hebert, and David Held. Zephyr: Zero-shot pose hypothesis rating. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 14141–14148. IEEE, 2021.

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

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Semantic_Image_Synthesis_With_Spatially-Adaptive_Normalization_CVPR_2019_paper.html.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4195–4205, 2023.

Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. URL https://arxiv.org/abs/2508.10104.

Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pp. 530–546. Springer, 2020.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018.

Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002.

Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3343–3352, 2019a.

Gu Wang, Fabian Manhardt, Jianzhun Shao, Xiangyang Ji, Nassir Navab, and Federico Tombari. Self6d: Self-supervised monocular 6d object pose estimation. In *European Conference on Computer Vision*, pp. 108–125. Springer, 2020.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2642–2651, 2019b.

Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9773–9783, 2023.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. *arXiv preprint arXiv:2503.11651*, 2025a. URL https://arxiv.org/abs/2503.11651.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025b.

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

Y Xiang, T Schmidt, V Narayanan, and D Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes robotics: Science and systems xiv. *Robotics: Science and Systems Foundation*, 2018.

Hui Yang, Wei Sun, Jian Liu, Jin Zheng, Zhiwen Zeng, and Ajmal Mian. Rgb-based category-level object pose estimation via depth recovery and adaptive refinement. *IEEE Robotics and Automation Letters*, 2025.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10371–10381, 2024.

Yang You, Ruoxi Shi, Weiming Wang, and Cewu Lu. Cppf: Towards robust category-level 9d pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6866–6875, 2022.

Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems*, 36:54627–54644, 2023.

Ruida Zhang, Ziqin Huang, Gu Wang, Chenyangguang Zhang, Yan Di, Xingxing Zuo, Jiwen Tang, and Xiangyang Ji. Lapose: Laplacian mixture shape modeling for rgb-based category-level object pose estimation. In *European Conference on Computer Vision*, pp. 467–484. Springer, 2024.

Linfang Zheng, Tze Ho Elden Tse, Chen Wang, Yinghan Sun, Hua Chen, Ales Leonardis, Wei Zhang, and Hyung Jin Chang. Georef: Geometric alignment across shape variation for category-level object pose refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10693–10703, 2024.