Boosting 6D Object Pose Estimation via Monocular Depth Cues

Fengda Hao[1], Rui Song[1], Qingyuan Wang[1], Jiaojiao Li[1], Zhiyong Hu[1], David Ferstl[2], and Yinlin Hu[2]

[1] State Key Laboratory of Integrated Services Networks, Xidian University
[2] Magic Leap

Abstract. We present an RGB-only method for 6D object pose estimation that leverages monocular depth cues from a single image. Although monocular depth estimation has advanced substantially, its predictions remain scale-ambiguous and locally unreliable, limiting their use in metric pose refinement. We address this gap by closing the loop between depth correction and pose refinement: monocular depth is not only a regularizer but is iteratively calibrated and filtered using pose-induced geometric consistency, enabling stable metric pose updates from RGB alone. We propose a dynamic depth outlier removal module based on metric consistency and infer object pose from dense 2D correspondences. Both components are embedded into a recurrent optimization loop, enabling iterative depth correction and pose refinement. Experiments on seven BOP datasets demonstrate state-of-the-art performance among RGB-only methods, without requiring real depth input.

Keywords: 6D Object Pose Estimation · RGB-Only Pose Refinement · Recurrent Pose-Depth Refinement

1 Introduction

Estimating accurate 6D object poses from visual observations is fundamental to various applications, including robotic manipulation [36], augmented reality [28], and autonomous perception. Despite recent progress in deep learning-based pose estimation [23, 24, 33, 37], achieving instance-level 6D pose estimation from a single RGB image remains a longstanding challenge. The absence of true depth information makes monocular observations inherently ambiguous in scale and geometry, leading to unstable optimization and limited accuracy.

Existing RGB-based refinement methods typically rely on photometric alignment or feature flow consistency between rendered and observed images [14–16, 22, 26, 29]. However, such approaches often fail in complex real-world scenarios with textureless regions, occlusions, or illumination changes, where purely appearance-based cues are insufficient for robust geometric correction. The lack of reliable depth constraints prevents these methods from converging to accurate metric geometry, as shown in Fig. 1.

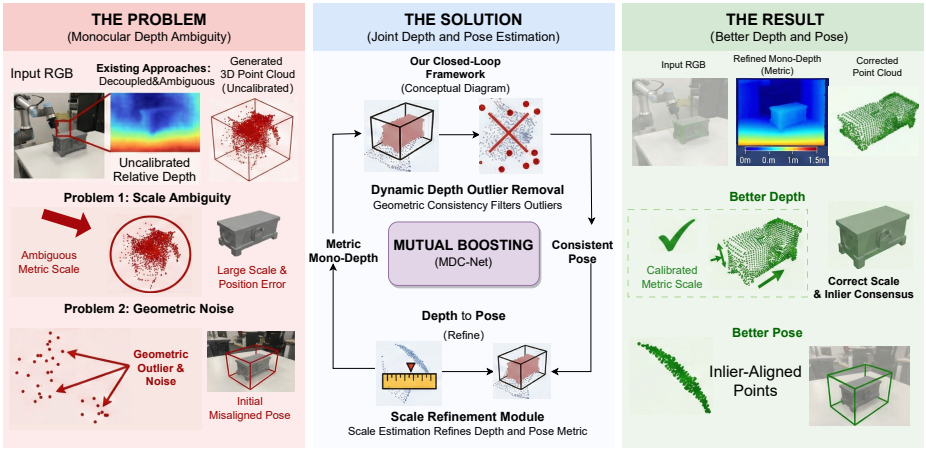

Fig. 1: Overview. RGB-based 6D object pose estimation is challenging due to the lack of depth information. On the other hand, despite recent advances in monocular depth estimation, most predictions remain noisy and metric-inconsistent because of the problem's ill-posed nature. Nevertheless, they provide useful relative depth cues, which we exploit for RGB-only 6D pose estimation in this work. We propose a recurrent refinement framework that jointly optimizes pose and depth, iteratively improving both depth consistency and pose accuracy.

To address these issues, we introduce MDC-Net, a framework that capitalizes on monocular depth cues to inject dense 3D geometry into the RGB-based pose refinement process. Starting from an initial 6D pose, MDC-Net iteratively refines the object's pose by enforcing both geometric and photometric alignment within a unified optimization loop.

Specifically, we design a Scale Refinement Module (SRM) that decouples relative and absolute geometry by predicting a global scale factor from image-level features, converting relative monocular depth into metric depth. In parallel, an Inlier-Guided Depth Refinement (IGDR) module identifies and removes inconsistent regions based on point cloud registration residuals, enabling iterative monocular depth updates and preventing error accumulation. Together, these two modules form a closed-loop refinement system that progressively corrects pose and depth estimation toward accurate metric geometry, as shown in Fig. 1.

Extensive experiments on the LINEMOD [3] and YCB-Video [43] benchmarks demonstrate that MDC-Net significantly boosts both pose accuracy and depth consistency using only RGB input. Our method surpasses recent RGB-based state-of-the-art techniques [29–31] across key metrics, establishing new state-of-the-art performance while maintaining strong robustness and generalization across diverse scenes. By integrating scale-aware depth prediction with outlier-resistant geometric reasoning, MDC-Net provides a new perspective for low-cost, depth-free 6D pose refinement.

Our main contributions are summarized as follows:

- First, we propose a recurrent joint pose-depth refinement formulation for RGB-only 6D object pose estimation. Given a coarse initial pose, MDC-Net iteratively updates both monocular depth and object pose, rather than using monocular depth as a fixed auxiliary prior.
- Second, we propose SRM, which enforces metric consistency by aligning global scale/shift and introducing geometry-aware constraints to transform relative monocular depth into an absolute depth proxy suitable for iterative refinement.
- Third, we propose IGDR, a geometric-consistency-driven refinement strategy that filters and updates unreliable depth regions using pose-induced residuals and pairwise compatibility, effectively mitigating error propagation over iterations.
- Extensive experiments on LINEMOD and YCB-Video benchmarks demonstrate that MDC-Net consistently outperforms existing RGB-based refinement methods. The proposed framework offers a scalable, low-cost, and geometry-consistent solution for monocular 6D pose refinement.

2 Related Work

Monocular depth estimation (MDE) has recently evolved from supervised [10, 27] or self-supervised [12,48] structural constraints toward generalizable geometric modeling with strong cross-domain and zero-shot transfer capability [13,34]. Methods such as Depth Anything V2 [46], ZoeDepth [1], Metric3D V2 [21], DepthPro [2], and MoGe-2 [41] enhance depth quality through unified encoders, scale-transfer mechanisms, semantic priors, and improved data synthesis, collectively advancing monocular-depth quality toward universal geometric representation learning. Despite this progress, two fundamental challenges remain unresolved: the inherent scale ambiguity of monocular prediction (where models produce only relative or affine-invariant depth) and the persistence of outliers in reflective, thin-structure, or low-texture regions. These issues limit the reliability of monocular-depth in downstream geometric tasks, and current global regularization or post-processing strategies still lack adaptive mechanisms to address scale inconsistency and error propagation [2, 41, 46]. Resolving both scale determinacy and depth-outlier suppression remains a key bottleneck for applying monocular-depth in geometry-intensive applications.

Monocular-depth-based 6D pose estimation. Recent methods have explored monocular depth as a geometric cue for 6D pose estimation. OLD-Net [11] reconstructs object-level depth from a monocular RGB image by deforming category-level shape priors and aligns the predicted depth with canonical object representations for category-level pose recovery. SingRef6D [38] further improves monocular novel object pose estimation by fine-tuning Depth Anything V2 and introducing depth-aware matching from a single RGB reference. These works demonstrate the value of monocular depth for recovering object geometry when real depth is unavailable. However, they primarily use monocular depth as an estimation or matching cue. In contrast, MDC-Net targets the refinement problem given an initial pose and explicitly performs iterative bidirectional updates:

the current pose induces geometric consistency constraints for depth correction, and the corrected depth is then used to refine the next pose. This recurrent mutual update differentiates our method from prior monocular-depth-driven pose estimation pipelines.

Pose refinement has increasingly relied on end-to-end correspondence [31, 32] and flow-based networks [20, 35] that follow a "render-matching-iterative refinement" paradigm [26, 42], as exemplified by SCFlow [15], GenFlow [29], and their successors. By embedding 3D shape priors and multi-scale correlations, these methods achieve efficient refinement and strong generalization, while maintaining full differentiability [15]. Recent advances also show that incorporating depth or 3D field constraints can further stabilize the optimization process [39]. However, most models still lack explicit outlier rejection or robust weighting during pose computation, relying instead on implicit suppression learned by the network. This design makes them vulnerable to mismatches caused by occlusions, specularities, or appearance inconsistencies, and their robustness often deteriorates under noisy or out-of-distribution conditions. Emerging research trends thus aim to integrate explicit residual-based filtering [19], geometric consistency checks [47], or depth-aware reweighting [6] into end-to-end frameworks to recover the robustness traditionally offered by classical geometric registration.

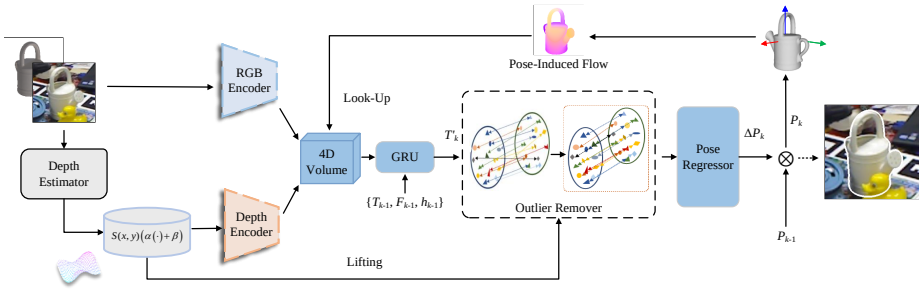

Fig. 2: Overview of the proposed framework. Given an RGB image, an initial pose, and the object model, our method iteratively refines the pose through a closed-loop pipeline. We first predict monocular depth from the RGB input and then extract features to construct a 4D cost volume. Using a recurrent matching structure, we incorporate a dynamic outlier removal module to explicitly handle depth noise when predicting the pose, which is subsequently used to compute pose-induced flow for the next iteration. By explicitly integrating monocular depth, geometric consistency, and dense matching, our design enables robust pose refinement in RGB-only settings.

3 Method

Given a single RGB image, an initial pose estimate P_{k-1}, camera intrinsics, and the CAD model of the target object M, our goal is to iteratively refine the

object's 6D pose under RGB-only input. Unlike prior RGB-based refinement methods that rely solely on photometric consistency, our framework leverages monocular depth predicted from a monocular depth estimator to inject geometric supervision into the refinement process. Specifically, we first provide an overview of the MDC-Net pipeline in Section 3.1, which integrates monocular depth estimation with iterative pose optimization. We then describe the design of the two key components: the Scale Refinement Module (SRM) in Section 3.2, which converts relative depth into metric depth through global scale prediction, and the Inlier-Guided Depth Refinement (IGDR) component in Section 3.3, which suppresses depth outliers and iteratively updates monocular depth to enhance geometric consistency. Finally, we present implementation details and training strategies in Section 3.4.

3.1 Overview

The overall architecture of MDC-Net is illustrated in Fig. 2. Given a single RGB image, a known 3D object model, and an initial pose estimate, our goal is to iteratively refine the 6D pose of the object without relying on ground-truth depth. The core idea of MDC-Net is to leverage monocular depth as a geometric bridge between image appearance and object pose, progressively improving both the depth quality and the pose accuracy in a closed-loop manner. Specifically, we first employ a pre-trained monocular depth estimator [41] to generate a relative depth map from the input RGB image. This relative depth lacks metric scale and may contain noise or inconsistent regions.

To address these issues, MDC-Net introduces two lightweight but complementary modules: (1) a Scale Refinement Module (SRM) that predicts a global scale factor and local residuals from image features to transform D_{rel} into a metric monocular depth D_{abs}; (2) an Inlier-Guided Depth Refinement (IGDR) module that detects and removes mismatched points based on registration residuals, iteratively updating the monocular depth to improve geometric consistency.

At each iteration, the mesh is rendered under the current pose to obtain rendered RGB/depth, object masks, and visible surface points, which provide shape constraints for dense matching and pose-induced flow. The rendered and observed features are used to build a 4D cost volume, while SRM converts monocular depth into a metric-consistent depth proxy. IGDR removes geometrically inconsistent correspondences based on the mesh-constrained 3D relations. In particular, P_{k-1} and P_k denote the previous and updated pose estimates, ΔP_k denotes the predicted residual pose update, F_k denotes the pose-induced flow or matching feature used at iteration (k), and h_k denotes the hidden state of the GRU-based recurrent pose regressor. The GRU takes the current matching features, corrected monocular depth features, and previous hidden state as input, and predicts ΔP_k, which is applied to P_{k-1} to obtain P_k.

Through this design, MDC-Net effectively integrates monocular depth reasoning and pose refinement into a unified iterative framework, enabling accurate, scale-aware, and geometry-consistent 6D pose estimation from RGB-only input.

3.2 Scale Refinement Module

Monocular depth estimation typically recovers only relative depth, lacking metric scale and often exhibiting affine bias or scale drift across scenes and viewpoints. To address these limitations, we introduce a robust Scale Refinement Module (SRM) that combines global affine correction with region-aware scale adaptation, enabling the network to jointly model global depth scaling and local scale variations for producing stable metric monocular-depth.

Given a predicted relative depth map, the scale refinement module first extracts global image features and regresses a set of global affine parameters to linearly correct the overall depth range. To further compensate for spatially varying scale inconsistencies commonly observed in complex scenes, a lightweight convolutional branch predicts a pixel-wise adaptive scale field. The final metric monocular depth is computed as

$$D_{abs}(x,y) = S(x,y) \cdot (\alpha \cdot D_{rel}(x,y) + \beta) + \Delta D(x,y) \tag{1}$$

where α and β denote scalar global scale and shift parameters, respectively, $S(x,y) \in \mathbb{R}^{H \times W}$ represents a spatially varying local scale field, and $\Delta D(x,y) \in \mathbb{R}^{H \times W}$ denotes a local depth residual term used for fine-grained detail refinement. By combining the stability of global affine normalization with the flexibility of local adaptation, the scale refinement module maintains consistent metric depth relationships across diverse viewing conditions, illumination changes, and reflective regions. During training, we enhance the original depth supervision with a geometric consistency loss to ensure that the refined metric depth is coherent in 3D space. Specifically, we introduce normal-consistency and reprojection terms formulated as:

$$L_{geo} = \|\nabla n(D_{abs}) - \nabla n(D_{gt})\|_1 + \|P(D_{abs}) - P(D_{gt})\|_1 \tag{2}$$

where $n(\cdot)$ denotes surface normals derived from depth, and $P(\cdot)$ represents the reprojection of 3D points back to the image plane. The overall loss function is defined as:

$$L_{SRM} = L_{scale} + \lambda_1 L_{depth} + \lambda_2 L_{grad} + \lambda_3 L_{geo} \tag{3}$$

where L_{scale} supervises global scale corrections, L_{grad} penalizes depth-gradient discrepancies to preserve edges. L_{depth}constrains pixel-level accuracy, and L_{geo} enforces geometric consistency. By decoupling affine scale and enabling region-level adaptation, the scale refinement module effectively suppresses global bias and local drift in monocular depth predictions while remaining lightweight, providing a significantly more reliable geometric input for subsequent outlier removal and pose refinement.

3.3 Inlier-Guided Depth Refinement Module

In RGB-only 6D pose refinement, correspondence errors are a major source of instability because incorrect matches can directly corrupt the iterative optimization. Existing end-to-end refinement networks typically regress pose updates

from dense optical flow or feature fields without explicit outlier rejection, allowing inconsistent matches to propagate through the optimization pipeline. To address this limitation, we introduce the Inlier-Guided Depth Refinement (IGDR) module, which establishes correspondences between rendered and real images via optical flow and explicitly removes mismatches using 3D geometric consistency. This process yields more reliable monocular depth and correspondence constraints, thereby improving stability and accuracy in pose refinement.

Given the current pose estimate P_t, the 3D model M, and the input RGB imageI_{real}, this component renders a synthetic view I_{rend}. A dense flow field $F_{r\to t}$ is then computed (e.g., using RAFT) to obtain 2D correspondences:

$$C = \left\{ \left(u_i^{\text{rend}}, u_i^{\text{real}}\right) | u_i^{\text{real}} = u_i^{\text{rend}} + F_{r\to t}\left(u_i^{\text{rend}}\right) \right\} \tag{4}$$

Both I_{real} and I_{rend} are passed through a monocular depth estimation network to generate monocular depth maps D_{real} and D_{rend}. Using camera intrinsics K, each correspondence pair is lifted to 3D:

$$p_i^{\text{rend}} = D_{\text{rend}}\left(u_i^{\text{rend}}\right) \cdot K^{-1}\tilde{u}_i^{\text{rend}} \tag{5}$$

$$p_i^{\text{real}} = D_{\text{real}}\left(u_i^{\text{real}}\right) \cdot K^{-1}\tilde{u}_i^{\text{real}} \tag{6}$$

Each pair $\left(p_i^{real}, p_i^{rend}\right)$ lies in different coordinate frames related by an unknown rigid transform $[R, t]$. Inspired by TurboClique [44], we evaluate pairwise geometric consistency in this component. For any two correspondence pairs (i, j), We define the geometric inconsistency as:

$$e_{ij} = \left| \left\| p_i^{\text{real}} - p_j^{\text{real}} \right\|_2 - \left\| p_i^{\text{rend}} - p_j^{\text{rend}} \right\|_2 \right| \tag{7}$$

Pairs satisfying $e_{ij} < \tau_{geo}$ are considered compatible. We construct a compatibility graph $G = (V, E)$, where nodes V denote correspondences and edges$E$ denote geometric agreement. Through heuristic clique sampling and consistency scoring, we extract the high-consistency subset as inliers.

To improve robustness, we incorporate feature similarity and flow confidence into a reweighting scheme:

$$w_i = \frac{1}{|N(i)|} \sum_{j \in N(i)} \exp\left(-\frac{e_{ij}^2}{2\sigma^2}\right) \cdot s_i \tag{8}$$

where $N(i)$ is the geometrically compatible neighborhood. s_i is the matching confidence score, and σ controls the sensitivity to geometric inconsistency. High weights correspond to inliers. The refined monocular depth is obtained via weighted fusion:

$$D_{\text{refined}}(x, y) = \frac{\sum_i w_i \cdot D_{\text{real}}\left(u_i^{\text{real}}\right)}{\sum_i w_i} \tag{9}$$

The updated depth is fed into the next pose refinement iteration, forming a closed-loop procedure consisting of flow matching, geometric consistency evaluation, outlier rejection, depth refinement, and pose refinement.

By explicitly modeling multi-point geometric relations, this inlier-guided refinement overcomes the limitations of end-to-end refinement networks that rely on single-point residuals or implicit robustness. Without requiring real depth, the module provides principled outlier detection and yields more stable geometric constraints. Experiments show that this module consistently improves pose and monocular depth accuracy under large initialization errors, illumination changes, and occlusions.

3.4 Implementation Details

For the input RGB images I, we first crop the target out based on the initial pose, and then resize it to a fixed resolution of 256×256. Given the ground-truth pose during training, we render the object to obtain supervisory signals for optical flow and depth (used only for training losses), while no ground-truth depth is used at inference. We use the exponentially weighted strategy [15,39] to calculate the loss in each iteration for flow loss, pose loss, and depth matching loss.

4 Experiments

Table 1: RGB-only 6D pose estimation results on 7 BOP datasets. We report the official BOP metric, Average Recall, on the unseen track. All methods operate without real depth at inference; our method achieves the best results with only one hypothesis.

Method	LM-O	T-LESS	TUD-L	IC-BIN	ITODD	HB	YCB-V	Avg.
MegaPose [26]	62.0	48.5	**84.6**	46.2	46.0	72.5	**76.4**	62.3
SCFlow [15]	62.9	57.9	75.9	47.9	34.2	74.5	70.7	60.5
FoundPose [32]	61.0	57.0	69.4	47.9	40.7	72.3	69.0	59.6
GenFlow [29]	56.3	52.3	68.4	45.3	39.5	73.9	63.3	57.0
GigaPose [31]	63.1	58.2	66.4	49.8	45.3	75.6	65.2	60.5
Co-op(5 hypo) [30]	65.5	**64.8**	72.9	**54.2**	**49.0**	**84.9**	68.9	65.7
Ours	**66.8**	**65.2**	75.2	**54.8**	48.2	84.2	74.1	**66.9**

4.1 Experimental Settings

The MDC-Net was trained on a combination of large-scale object-mesh datasets, including ShapeNet-Objects [4], Google-Scanned-Objects [8], and Objaverse [5],

covering about 90K unique 3D models. Since these datasets only provide object meshes, we reuse the synthetic renderings provided by [42], which are generated from these meshes with approximately 3 million images in total. All models are trained under identical settings without using real depth or ground-truth pose supervision. For evaluation, we follow the BOP benchmark protocol [18] and test on seven datasets: LM-O [3], T-LESS [17], TUD-L [18], IC-BIN [7], ITODD [9], HB [25] and YCB-V [43]. Together, these datasets contain 108 objects with diverse scales, textures, and symmetry properties, providing a comprehensive generalization and robustness evaluation.

4.2 Comparison with the State of the Art

Table 1 compares MDC-Net with recent RGB-only 6D pose estimation and refinement approaches, including MegaPose [26], SCFlow [15], FoundPose [32], GenFlow [29], GigaPose [31], and Co-op [30]. We report the official BOP metrics, including Average Recall (AR), Visible Surface Discrepancy (VSD), Maximum Symmetry Surface Distance (MSSD), and Maximum Symmetry Projection Distance (MSPD), all under the RGB-only setting. Our MDC-Net achieves the highest overall Average Recall (AR) of 66.9%, outperforming the previous best Co-op (65.7%) and other flow-based refinement networks. Notably, our method ranks first on LM-O (66.8%) ,T-LESS(65.2%),and IC-BIN (54.8%), demonstrating strong generalization to both synthetic-to-real transfer and texture-rich scenes. This improvement verifies that monocular depth guided optimization can effectively refine poses without requiring real depth input.

Table 2: Ablation of SRM and IGDR on LM-O and YCB-V under the RGB-only setting. "✓" indicates the component is enabled and "–" disabled.

		LM-O				YCB-V			
SRM	IGDR	AR ↑	VSD↑	MSSD↑	MSPD↑	AR ↑	VSD ↑	MSSD ↑	MSPD↑
-	-	59.9	53.8	55.6	70.3	66.3	56.4	70.3	72.2
✓	-	63	57.2	58.9	72.9	70.8	64.4	72.4	75.6
-	✓	64.5	59.1	60.4	74.1	72.3	63.5	75.1	78.3
✓	✓	**66.8**	**61.6**	**62.8**	**76.0**	**74.1**	**61.6**	**78.6**	**82.1**

Further analysis shows that MDC-Net remains competitive even on textureless or symmetry-dominant datasets such as T-LESS and ITODD, where depth ambiguity often causes failure in previous RGB methods. The Scale Refinement Module (SRM) enhances the geometric consistency of monocular depth, while the Inlier-Guided Depth Refinement (IGDR) removes inconsistent 3D correspondences during optimization. Together, these modules contribute to robust performance across different types of objects and illumination conditions.

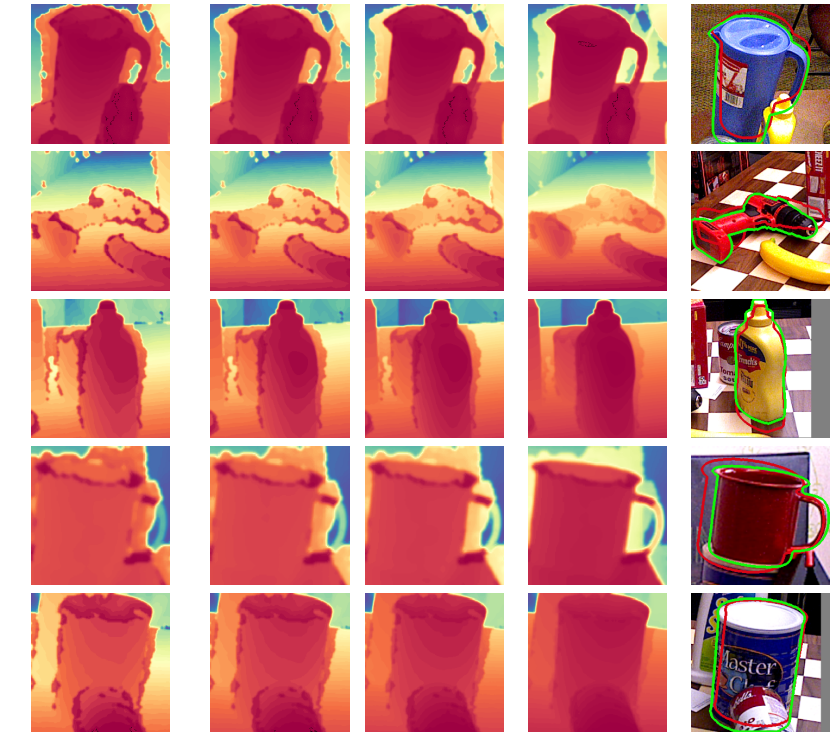

Fig. 3: Iterative refinement of pose and depth map. Our method co-optimizes the 6D object pose and the monocular depth map across iterations. The monocular depth becomes progressively cleaner and more geometrically consistent, leading to highly accurate final pose estimation.

As shown in Figure 3, this synergistic process results in clear visual improvements: the monocular depth progressively becomes more accurate and consistent with the object's geometry, so the estimated pose (visualized by the outline projection) converges rapidly towards the ground truth. This illustrates that MDC-Net does not merely use monocular depth as a static cue, but actively refines it into a more reliable form, which is fundamental to achieving RGB-D level accuracy from RGB-only input.

4.3 Ablation Study

To analyze the contribution of each component in MDC-Net, we performed an ablation study on the Scale Refinement Module (SRM) and the proposed Inlier-Guided Depth Refinement (IGDR) module on the LM-O dataset. As summarized in Table 2, the removal of both SRM and IGDR results in a noticeable degradation across all metrics (59.9% AR), confirming that both modules are essential

for stabilizing pose refinement.When enabling only the SRM, the model achieves 63.0% AR, indicating that global scale recovery and local depth correction improve the geometric consistency between monocular depth and image features. Activating only the IGDR further increases the performance to 64.5% AR, as its inlier-driven matching strategy effectively removes inconsistent correspondences that may lead to unstable optimization. When both modules are applied together, MDC-Net achieves its best overall performance, achieving 66.8% AR, (62.8% VSD, 61.6% MSSD and 76.0% MSPD), outperforming all single-module variants. The results highlight the complementarity of SRM and IGDR, the former provides precise depth adjustments and refinement, the latter ensures robust internal point selection guided by geometric consistency. Their joint optimization allows MDC-Net to maintain both postural accuracy and convergent stability.

To evaluate the robustness of MDC-Net across different monocular-depth estimation (MDE) models, we compare its performance using both relative and metric depth inputs, as summarized in Table 3. All experiments use the same initialization and training configuration, with all depth estimation networks kept frozen during inference. For each type of depth input—whether relative or metric—we apply our proposed Scale Refinement Module (SRM) to recover absolute scale and correct local depth biases, thereby ensuring consistent geometric representations across diverse monocular depth sources.

Table 3: Impact of different monocular depth sources on MDC-Net. We evaluate relative-depth and metric-depth estimators as inputs while keeping all depth networks frozen. Results are reported in AR on YCB-V and LM-O, showing MDC-Net remains effective across diverse depth qualities and scales.

Monocular Depth (input)	Depth Type	AR↑	
		YCB-V	LM-O
ZoeDepth [1]	Relative	65.8	68.7
DepthAnything v1 [45]	Relative	66.1	69.2
DepthAnything v2 [46]	Relative	66.2	69.4
MoGe [40]	Relative	66.5	70.5
Metric3D v2 [21]	Metric	66.2	70.8
MoGe-2 [41]	Metric	**66.8**	**74.1**
Depth-GT	-	79.1	85.5

As shown in Table 3, MDC-Net achieves consistent pose refinement improvements across different level of inputs, even when the monocular depth quality varies notably. For fair comparison, we use the same initialization and training configuration, and apply SRM to normalize scale and correct local biases for every depth source. Compared with the RGB-only refiner, all monocular depth models yield substantial gains on both YCB-V and LM-O datasets. Among the relative-depth models, MoGe-Rel reaches the highest accuracy (66.5% on YCB-

V, 70.5% on LM-O), while the metric-depth version MoGe-2 achieves the best overall performance (66.8% / 74.1%), approaching the upper bound established by real depth-GT supervision (79.1% / 85.5%). This trend demonstrates that MDC-Net can effectively leverage geometric cues from monocular depth maps regardless of their original scale or noise level.

In particular, the performance difference between ZoeDepth and MoGe-2 remains within 1–2%, indicating that MDC-Net is robust to depth estimation quality and can extract meaningful 3D signals even from noisy or biased predictions. The SRM module provides consistent scale normalization, while the outlier removal module filters inconsistent correspondences during refinement, jointly contributing to stable optimization across diverse MDE qualities.

To further evaluate the versatility and robustness of MDC-Net, we investigated its optimization performance under different baseline framework initialization poses. All baseline methods follow an initialization–refinement paradigm, where the initial pose is estimated from RGB input and then iteratively refined using their own pose refinement network. We replace the original refinement stage of each baseline (SCFlow, FoundPose, GenFlow, Co-op) with MDC-Net while keeping their initialization stage unchanged. The qualitative results on the YCB-V and LM-O datasets, shown in Figure 4, illustrate how MDC-Net achieves high-fidelity pose estimation through improved geometric consistency, even under challenging conditions like occlusion and low texture.

Table 4: Robustness to different initialization poses. "Depth" indicates whether the baseline's original refinement uses real depth supervision/input. We report AR on LM-O and YCB-V. MDC-Net consistently improves all initializations using RGB-only input

Method	Depth	LM-O	YCB-V
FoundPose(init)	-	39.6	45.2
FoundPose(final)	✓	61	69
FoundPose(final)+Ours	-	64.2	70.9
GenFlow(init)	-	25	27.7
GenFlow(final)	✓	63.2	70.9
GenFlow(final)+Ours	-	63.5	73.3
Co-op(init)	-	59.7	62.6
Co-op(final)	-	59.7	62.6
Co-op(final)+Ours	-	66.8	74.1
SCFlow(init)	-	41.3	50.2
SCFlow(final)	✓	64.9	68.6
SCFlow(final)+Ours	-	**66.8**	**74.1**

As shown in Table 4, MDC-Net consistently improves the performance of all baselines, regardless of the initial pose quality. When applied to the coarse initial poses from SCFlow, FoundPose, GenFlow, and Co-op, our method yields an average gain of +20–40% AR, significantly surpassing the refinement accuracy

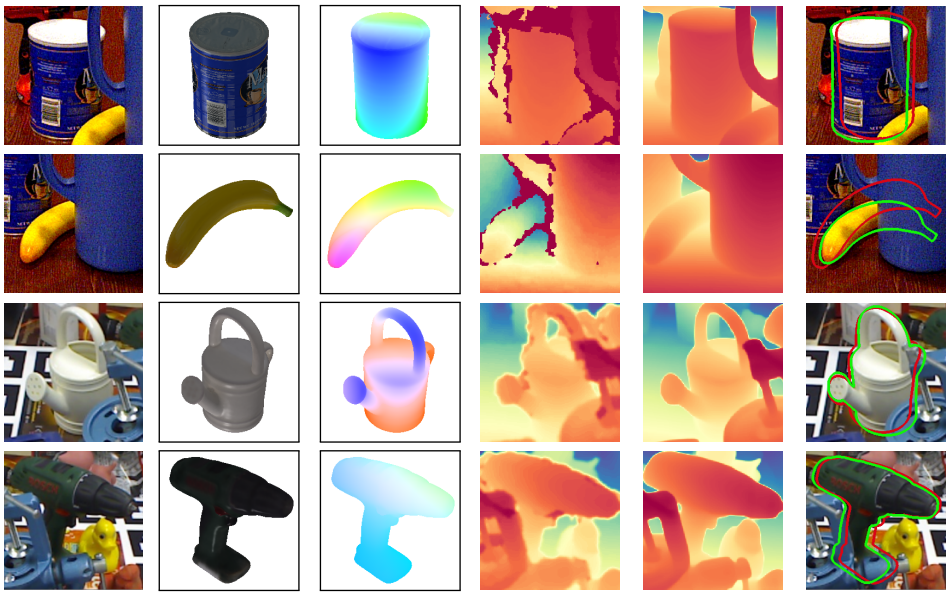

Fig. 4: Qualitative results. We show the example results on YCB-V and LM-O datasets before and after refinement. MDC-Net significantly improves object alignment in challenging scenarios including occlusions, reflective surfaces, and large initial pose deviations.

of their original models. For example, when refining SCFlow(init) results, MDC-Net boosts the AR from 41.3% to 62.5% on LM-O and from 50.2% to 61.2% on YCB-V. A similar improvement trend is observed for FoundPose and GenFlow, where our framework not only surpasses their initialization results but also outperforms their own final refinement stages that use true depth input. Even when applied to strong multi-hypothesis methods like Co-op, MDC-Net achieves the best results—66.8% on LM-O and 74.1% on YCB-V—demonstrating superior refinement capability purely from RGB input. These results highlight that MDC-Net is both model-agnostic and initialization-robust, capable of refining noisy or biased initial poses into high-precision estimates without relying on explicit depth supervision.

Comparison with RGB-D methods. FoundationPose achieves its performance using RGB-D inputs with real depth inputs. In contrast, our method operates in a purely RGB setting by extracting geometric information from pseudo-depth images. We compare our method against two approaches that support both RGB and RGB-D settings under a *single-hypothesis* setup, as shown in Table 5. While real depth improves the performance, our method substantially outperforms their RGB variants and, more importantly, produces higher-quality denoised pseudo-depth, as shown in the main paper.

Table 5: Comparison with RGB-D settings under a single-hypothesis setup.

Method	Modality	YCBV	LM-O	**Avg.**
Co-op (1 hypo.)	RGB	67.0	64.2	64.0
	RGB-D	**87.4**	<u>71.5</u>	<u>73.6</u>
SCFlow2	RGB	72.7	58.9	63.8
	RGB-D	84.9	69.6	68.6
Ours	RGB	74.1	66.8	66.9
	RGB-D (Oracle)	<u>85.5</u>	**79.1**	**75.3**

IGDR complexity and sensitivity. Designed for real-time refinement, IGDR reuses dense flow from fixed-resolution (256^2) inputs and selects top-$2k$ ($N = 2048$) keypoints to avoid re-computation. By relaxing the NP-hard clique problem to a greedy formulation (similar to TurboReg [44]), we reduce the solver complexity to $O(N \log N)$. While graph construction remains $O(N^2)$, it is highly parallelized via GPU operations. IGDR demonstrates high robustness because its hyperparameters are physically grounded rather than arbitrary—yielding $<$ 0.6% performance variance across reasonable ranges (e.g., $\tau_{geo} \in [0.05m, 0.2m]$).

Runtime analysis. We report the runtime of each stage in our framework in Table 6. The measured stages include monocular depth estimation, RGB/depth feature encoding, SRM-based depth calibration, IGDR-based geometric outlier removal, and three recurrent pose-regression iterations. All timings are measured on an RTX 3090 GPU. The refinement stage takes 164.3 ms in total, with monocular depth estimation and IGDR accounting for the main computational cost. When FoundPose is used to provide the coarse initial pose, the initialization stage takes 1753.6 ms, resulting in a full pipeline latency of 1.92s. Under the same RTX 3090 setting, this runtime is lower than those of recent RGB-based baselines such as Co-op (5 hyp.) and FoundationPose, which require 4.19s and 9.42s per image on average, respectively.

Table 6: Runtime breakdown of the proposed pose refinement method. We report the time consumption of each component in MDC-Net after a coarse initial pose is provided, including monocular depth prediction, SRM, IGDR, and recurrent pose regression. All timings are measured on an RTX 3090 GPU.

Component	Time (ms)
Depth Estimation (MoGe-2 [41])	62.0
RGB Encoder & 4D Volume	15.0
SRM + Depth Encoding	7.8
IGDR (3 iters, 21.2 ms each)	63.6
Pose Regressor (3 iters)	15.9
Total Refinement	**164.3**

5 Conclusion

In this work, we presented MDC-Net, a robust RGB-only 6D pose refinement framework that leverages monocular depth as an intermediate geometric representation. To address the inherent scale ambiguity and local inconsistency of monocular depth, we introduced the Scale Refinement Module (SRM), which decouples global affine correction from spatially adaptive scale adjustment to produce stable metric-like depth. Furthermore, we proposed the Inlier-Guided Depth Refinement (IGDR) module, which performs explicit outlier removal in 3D via pairwise geometric consistency, yielding reliable correspondence constraints for iterative pose refinement. Through the integration of these components into a differentiable refinement pipeline, MDC-Net progressively enhances both depth quality and pose accuracy across iterations.

Extensive experiments on seven BOP datasets demonstrate that MDC-Net sets a new state of the art among RGB-only refinement methods and significantly boosts the performance of existing pose estimation frameworks when used as a plug-and-play refinement module. Ablation studies further validate the effectiveness of SRM and IGDR, particularly in challenging scenarios involving noise, occlusion, and inaccurate initialization.

In the future, we plan to extend our framework toward unified RGB-only geometric reasoning, explore its integration with generative reconstruction models, and investigate real-time deployment on resource-limited platforms. We believe that leveraging robust monocular depth will continue to fuel progress toward accurate, scalable, and cost-efficient 6D pose estimation.

Acknowledgments. This work was supported in part by the Youth Innovation Team of Shaanxi Universities, the "Scientist + Engineer" team of the Qin Chuang Yuan in Shaanxi Province, the Xi'an Science and Technology Program, the "Leading the Charge" initiative for the industrialization of core technologies in key industrial chains in Shaanxi Province, the National Nature Science Foundation of China under Grant 62371359, the Key Research and Development Program of Shaanxi, the Key Research Program of the Chinese Academy of Sciences.

References

1. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: ZoeDepth: Zero-Shot Transfer By Combining Relative And Metric Depth. arXiv preprint arXiv:2302.12288 (2023)
2. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth Pro: Sharp Monocular Metric Depth In Less Than A Second. arXiv preprint arXiv:2410.02073 (2024)
3. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D Object Pose Estimation Using 3D Object Coordinates. In: European conference on computer vision. pp. 536–551. Springer (2014)

4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: ShapeNet: An Information-Rich 3D Model Repository. arXiv preprint arXiv:1512.03012 (2015)
5. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A Universe Of Annotated 3D Objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13142–13153 (2023)
6. Ding, Y., Kocur, V., Vávra, V., Haladová, Z.B., Yang, J., Sattler, T., Kukelova, Z.: RePoseD: Efficient Relative Pose Estimation With Known Depth Information. arXiv preprint arXiv:2501.07742 (2025)
7. Doumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.K.: Recovering 6D Object Pose And Predicting Next-Best-View In The Crowd. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3583–3592 (2016)
8. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google Scanned Objects: A High-Quality Dataset Of 3D Scanned Household Items. In: 2022 International Conference on Robotics and Automation (ICRA). pp. 2553–2560. IEEE (2022)
9. Drost, B., Ulrich, M., Bergmann, P., Hartinger, P., Steger, C.: Introducing MVTec ITODD-A Dataset For 3D Object Recognition In Industry. In: Proceedings of the IEEE international conference on computer vision workshops. pp. 2200–2208 (2017)
10. Eigen, D., Puhrsch, C., Fergus, R.: Depth Map Prediction From A Single Image Using A Multi-Scale Deep Network. Advances in neural information processing systems **27** (2014)
11. Fan, Z., Song, Z., Xu, J., Wang, Z., Wu, K., Liu, H., He, J.: 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)
12. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised Monocular Depth Estimation With Left-Right Consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
13. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards Zero-Shot Scale-Aware Monocular Depth Estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9233–9243 (2023)
14. Hai, Y., Song, R., Li, J., Ferstl, D., Hu, Y.: Pseudo Flow Consistency For Self-Supervised 6D Object Pose Estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14075–14085 (2023)
15. Hai, Y., Song, R., Li, J., Hu, Y.: Shape-Constraint Recurrent Flow For 6D Object Pose Estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4831–4840 (2023)
16. Haugaard, R.L., Buch, A.G.: SurfEmb: Dense And Continuous Correspondence Distributions For Object Pose Estimation With Learnt Surface Embeddings. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6749–6758 (2022)
17. Hodan, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D Dataset For 6D Pose Estimation Of Texture-Less Objects. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 880–888. IEEE (2017)

18. Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J.: BOP Challenge 2023 On Detection, Segmentation And Pose Estimation Of Seen And Unseen Rigid Objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5610–5619 (2024)
19. Hong, Z.W., Hung, Y.Y., Chen, C.S.: RDPN6D: Residual-Based Dense Point-Wise Network For 6DoF Object Pose Estimation Based On RGB-D Images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5251–5260 (2024)
20. Horn, B.K., Schunck, B.G.: Determining Optical Flow. Artificial intelligence **17**(1-3), 185–203 (1981)
21. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3D V2: A Versatile Monocular Geometric Foundation Model For Zero-Shot Metric Depth And Surface Normal Estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
22. Hu, Y., Fua, P., Salzmann, M.: Perspective Flow Aggregation For Data-Limited 6D Object Pose Estimation. In: European Conference on Computer Vision. pp. 89–106. Springer (2022)
23. Hu, Y., Fua, P., Wang, W., Salzmann, M.: Single-Stage 6D Object Pose Estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2930–2939 (2020)
24. Hu, Y., Speierer, S., Jakob, W., Fua, P., Salzmann, M.: Wide-Depth-Range 6D Object Pose Estimation In Space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15870–15879 (2021)
25. Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: HomebrewedDB: RGB-D Dataset For 6D Pose Estimation Of 3D Objects. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019)
26. Labbé, Y., Manuelli, L., Mousavian, A., Tyree, S., Birchfield, S., Tremblay, J., Carpentier, J., Aubry, M., Fox, D., Sivic, J.: MegaPose: 6D Pose Estimation Of Novel Objects Via Render & Compare. arXiv preprint arXiv:2212.06870 (2022)
27. Liu, F., Shen, C., Lin, G.: Deep Convolutional Neural Fields For Depth Estimation From A Single Image. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5162–5170 (2015)
28. Marchand, E., Uchiyama, H., Spindler, F.: Pose Estimation For Augmented Reality: A Hands-On Survey. IEEE transactions on visualization and computer graphics **22**(12), 2633–2651 (2015)
29. Moon, S., Son, H., Hur, D., Kim, S.: GenFlow: Generalizable Recurrent Flow For 6D Pose Refinement Of Novel Objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10039–10049 (2024)
30. Moon, S., Son, H., Hur, D., Kim, S.: Co-op: Correspondence-Based Novel Object Pose Estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11622–11632 (2025)
31. Nguyen, V.N., Groueix, T., Salzmann, M., Lepetit, V.: GigaPose: Fast And Robust Novel Object Pose Estimation Via One Correspondence. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9903–9913 (2024)
32. Örnek, E.P., Labbé, Y., Tekin, B., Ma, L., Keskin, C., Forster, C., Hodan, T.: FoundPose: Unseen Object Pose Estimation With Foundation Features. In: European Conference on Computer Vision. pp. 163–182. Springer (2024)

33. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: PVNet: Pixel-Wise Voting Network For 6DoF Pose Estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4561–4570 (2019)
34. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular Depth Estimation Using Diffusion Models. arXiv preprint arXiv:2302.14816 (2023)
35. Teed, Z., Deng, J.: RAFT: Recurrent All-Pairs Field Transforms For Optical Flow. In: European Conference on Computer Vision (2020)
36. Wang, C., Xu, D., Zhu, Y., Martín-Martín, 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. pp. 3343–3352 (2019)
37. 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. pp. 16611–16621 (2021)
38. Wang, J., Zhu, H., Guo, H., Al Mamun, A., Xiang, C., Lee, T.H.: SingRef6D: Monocular Novel Object Pose Estimation With A Single RGB Reference. Advances in Neural Information Processing Systems **38**, 25400–25437 (2026)
39. Wang, Q., Song, R., Li, J., Cheng, K., Ferstl, D., Hu, Y.: SCFlow2: Plug-And-Play Object Pose Refiner With Shape-Constraint Scene Flow. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22045–22054 (2025)
40. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: MoGe: Unlocking Accurate Monocular Geometry Estimation For Open-Domain Images With Optimal Training Supervision. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5261–5271 (2025)
41. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: MoGe-2: Accurate Monocular Geometry With Metric Scale And Sharp Details. arXiv preprint arXiv:2507.02546 (2025)
42. Wen, B., Yang, W., Kautz, J., Birchfield, S.: FoundationPose: Unified 6D Pose Estimation And Tracking Of Novel Objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17868–17879 (2024)
43. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network For 6D Object Pose Estimation In Cluttered Scenes. arXiv preprint arXiv:1711.00199 (2017)
44. Yan, S., Shi, P., Zhao, Z., Wang, K., Cao, K., Wu, J., Li, J.: TurboReg: TurboClique For Robust And Efficient Point Cloud Registration. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 26371–26381 (2025)
45. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: 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)
46. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
47. Zhao, H., Wei, S., Shi, D., Tan, W., Li, Z., Ren, Y., Wei, X., Yang, Y., Pu, S.: Learning Symmetry-Aware Geometry Correspondences For 6D Object Pose Estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14045–14054 (2023)
48. Zhou, T., Brown, M., Snavely, N., Lowe, D.G.: Unsupervised Learning Of Depth And Ego-Motion From Video. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1851–1858 (2017)