Omni6DPose: A Benchmark and Model for Universal 6D Object Pose Estimation and Tracking

Jiyao Zhang[1,2,3*], Weiyao Huang[1*], Bo Peng[1*], Mingdong Wu[1,2,3], Fei Hu[1,2,3], Zijian Chen[4], Bo Zhao[5,6], and Hao Dong[1,2,3†]

[1] Center on Frontiers of Computing Studies, School of Computer Science, Peking University
[2] PKU-Agibot Lab, School of Computer Science, Peking University
[3] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University [4] Waseda University [5] Shanghai Jiao Tong University [6] BAAI
{jiyaozhang, sshwy, bo.peng, hofee}@stu.pku.edu.cn
zijianchenwaseda@akane.waseda.jp bo.zhao@sjtu.edu.cn {wmingd, hao.dong}@pku.edu.cn

https://jiyao06.github.io/Omni6DPose/

Abstract. 6D object pose estimation is crucial in the field of computer vision. However, it suffers from a significant lack of large-scale and diverse datasets, impeding comprehensive model evaluation and curtailing downstream applications. To address these issues, this paper introduces **Omni6DPose**, a substantial benchmark featured by its diversity in object categories, large scale, and variety in object materials. Omni6DPose is divided into three main components: **ROPE** (Real 6D Object Pose Estimation Dataset), which includes 332K images annotated with over 1.5M annotations across 581 instances in 149 categories; **SOPE**(Simulated 6D Object Pose Estimation Dataset), a simulated training set created by mixed reality and physics-based depth simulation; and **PAM**(Pose Aligned 3D Models), the manually aligned real scanned objects used in ROPE and SOPE. Omni6DPose is inherently challenging due to the substantial variations and ambiguities. To address this issue, we introduce **GenPose++**, an enhanced version of the SOTA category-level 6D object pose estimation framework, incorporating two pivotal improvements: Semantic-aware feature extraction and Clustering-based aggregation. Moreover, we provide a comprehensive benchmarking analysis to evaluate the performance of previous methods on this new large-scale dataset in the realms of 6D object pose estimation and pose tracking.

Keywords: Benchmark · object pose estimation · object pose tracking

1 Introduction

6D object pose estimation [16,28,45] and pose tracking [15,21] from single images is an essential task in computer vision, holding immense potential for applications in robotics [1] and augmented reality/virtual reality (AR/VR) [19]. Over recent decades, the domain has experienced significant advancements, primarily dominated by data-driven learning approaches. Analogous to the pivotal role of data in learning-based 2D foundation tasks, high-quality, comprehensive datasets are paramount in the context of 6D object pose estimation and tracking.

*: equal contributions, †: corresponding author

Fig. 1: We introduce **Omni6DPose** for universal 6D object pose estimation and tracking, which comprises three key components: (1) **ROPE** (right), a large-scale real-world dataset for evaluation. (2) **SOPE** (left), a vast synthetic dataset generated through mixed reality and physics-based simulation for training. (3) **PAM** (center), a collection of pose-aligned 3D object models. Omni6DPose is featured by its diversity in object categories, large scale, and variety in object materials.

Today 6D object pose estimation is studied under two lenses: instance-level and category-level. In instance-level settings, datasets such as Linemod [10], YCB-Video [36], and T-LESS [11] have gained widespread acceptance as benchmarks. These datasets are distinguished by their focus on detailed, individual object instances, thereby enabling algorithms to precisely learn and predict the poses of specific items. On the other hand, category-level pose estimation emphasizes generalization across different items within a particular object category. The NOCS [28] dataset stands out as the most widely used in the category-level object pose estimation field, providing a simulated dataset for training and a small-scale real-world dataset for evaluation. Despite their contributions to advancing the field, these datasets present limitations due to their small scale in terms of instances or categories. This results in two significant challenges:

1. It hampers comprehensive evaluation of different models' performance, limiting the development of research in this field.
2. It restricts the applicability of research findings across diverse domains, due to the limited variety of object instances or categories represented.

To address the aforementioned challenges and drive advancements in this field, this paper introduces **Omni6DPose**, a universal 6D object pose estimation dataset featured by its diversity in object categories, large scale, and variety in object materials. Omni6DPose is segmented into three principal components: (1) **ROPE** (Real 6D Object Pose Estimation Dataset), which encompasses 332K images annotated with over 1.5M annotations across 581 instances in 149 categories; (2) **SOPE** (Simulated 6D Object Pose Estimation Dataset), comprising 475K images generated in a mixed reality setting with depth simulation, furnished with over 5M annotations spanning 4162 instances in the same 149

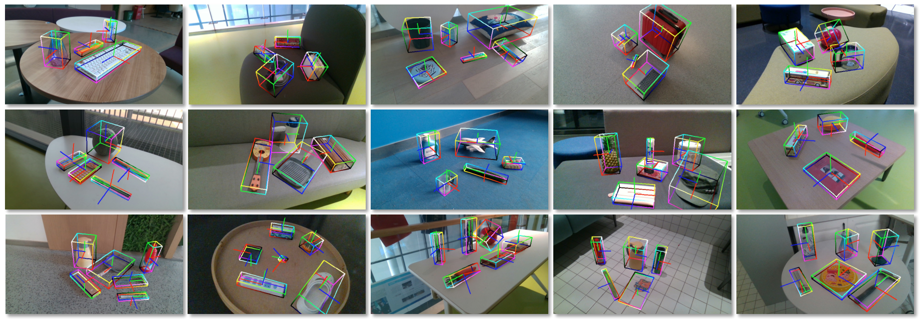

Fig. 2: ROPE dataset visualization. In the figure, bounding boxes are colored according to the coordinates in the object's coordinate system.

categories. The mixed reality bridges the semantic sim2real gap, while the depth sensor simulation closes the geometric sim2real gap; (3) **PAM**(Pose Aligned 3D Models), the manually aligned, real scanned objects utilized in both ROPE and SOPE, enabling the generation of diverse downstream task data.

Omni6DPose poses inherent challenges due to its considerable variations, diverse materials, and inherent ambiguities, which reflect the complexities encountered in real-world applications. Figure 2 illustrates some examples from ROPE. To tackle these issues, we introduce **GenPose++**, which incorporates GenPose [43] with two pivotal improvements: Semantic-aware feature extraction and Clustering-based aggregation, tailored specifically to the nuances of the Omni6DPose in question. Furthermore, as a Universal 6D object pose estimation dataset, this paper also offers a comprehensive benchmarking analysis to assess the performance of existing methods on category-level 6D object pose estimation and pose tracking. We summarize our contributions as follows:

1. We present **Omni6DPose**, a comprehensive 6D object pose estimation dataset with extensive categories, instance diversity, and material variety.
2. We propose a real data collection pipeline and a simulation framework for generating synthetic data with low semantic and geometry sim2real gaps.
3. We introduce **GenPose++** for category-level 6D object pose estimation and tracking, demonstrating SOTA performance on Omni6DPose.

2 Related work

For 6D object pose estimation, there are two main branches: instance-level and category-level. Instance-level estimation is tested on seen objects or on unseen objects with reference, while category-level estimation is tested on unseen instances of known categories without CAD models. In this section, we review and compare existing datasets to our large-scale category-level dataset and review relevant algorithms for category-level pose estimation and tracking.

2.1 6D Object Pose Estimation Datasets

Following the outlined branches of 6D object pose estimation, we have reviewed datasets corresponding to both instance-level and category-level 6D object pose estimation. A comparative analysis of these datasets is provided in Table 1.

Table 1: This table compares datasets for 6D object pose estimation, focusing on object category count, reality of the data, data modalities (RGB, Depth, IR), and object attributes such as quantity, CAD model availability, and inclusion of transparent and specular objects. It also details video characteristics by number and marker presence, along with image and annotation counts. 'Wild6D*' refers specifically to the test split of the Wild6D dataset, as the training data does not provide annotations. The symbol '-' indicates the absence of a particular feature within the dataset.

Dataset	Cat.	Real	Modality			Object				Marker-free	Vid.	Img.	Anno.
			RGB	Depth	IR	Num.	CAD	Trans.	Spec.				
CAMERA [28]	6	✗	✓	✓	✗	1085	✓	✗	✗	✓	-	300K	4M
SOPE(Ours)	149	✗	✓	✓	✓	4162	✓	✓	✓	✓	-	475K	5M
YCB-Video [36]	-	✓	✓	✓	✗	21	✓	✗	✗	✓	92	133K	598K
T-LESS [11]	-	✓	✓	✓	✗	30	✓	✗	✗	✗	20	48K	48K
Linemod [10]	-	✓	✓	✓	✗	15	✓	✗	✗	✗	-	18K	15K
StereoOBJ-1M [17]	-	✓	✓	✗	✗	18	✓	✓	✓	✗	182	393K	1.5M
REAL275 [28]	6	✓	✓	✓	✗	42	✓	✗	✗	✗	18	8K	35K
PhoCaL [30]	8	✓	✓	✓	✗	60	✓	✓	✓	✓	24	3.9K	91K
HouseCat6D [12]	10	✓	✓	✓	✗	194	✓	✓	✓	✓	41	23.5K	160K
Wild6D* [9]	5	✓	✓	✓	✗	162	✗	(✓)	✗	✓	486	10K	10K
PACE [41]	44	✓	✓	✓	✗	576	✓	✗	✗	✓	300	55K	258K
ROPE(Ours)	149	✓	✓	✓	✓	581	✓	✓	✓	✓	363	332K	1.5M

Instance-level 6D object pose estimation dataset. LineMod [10] is one of the most used datasets, providing non-temporal RGB-D images and ground truth pose annotations. YCB-Video [36] provides RGB-D videos and annotations, enabling pose-tracking approaches. T-LESS [11] features texture-less objects with symmetries and mutual similarities. StereoOBJ-1M [17] achieves a leap in dataset scale and features transparent and reflective objects. While these datasets are extensive in terms of image and annotation count, they are limited in the diversity of instances they cover. For example, StereoOBJ-1M comprises 339K frames and 1.5M annotations, yet it includes only 18 unique object instances.

Category-level 6D object pose estimation dataset. NOCS [28] provides the first benchmark in category-level pose estimation, while Wild6D [9] addresses the scalability issue of datasets by leveraging unlabeled and synthetic data. PhoCal [30] focuses on photometrically challenging objects and HouseCat6D [12] offers diverse scenes, viewpoints, and grasping annotations. However, these datasets cover only a limited number of categories, even the most extensive dataset, HouseCat6D, includes merely ten categories. PACE [41], a concurrent work, focusing on object pose estimation in complex scenes. In contrast, our datasets, SOPE and ROPE, set new benchmarks by offering the widest range of categories and featuring objects with diverse materials, thereby enhancing dataset diversity and realism for pose estimation research.

2.2 6D Object Pose Estimation and Tracking Algorithm

Category-leval 6D object pose estimation. Category-level 6D object pose estimation [28,39,40] aims to estimate unseen instance poses within the same category. NOCS [28] introduces a normalized object coordinate space for pose prediction without CAD models, while SPD [29] and SGPA [3] utilize category-level priors for enhanced estimation. HS-Pose [44] extends 3D-GC [14] for improved feature extraction from point clouds. IST-Net [16] transforms features between camera and world spaces implicitly, without relying on 3D priors, surpassing previous methods. However, these techniques, mainly regression-based, require ad-hoc designs for symmetric objects. GenPose [43] addresses it by generatively modeling the pose distribution, eliminating the need for symmetry considerations. Yet, GenPose neglects RGB semantic information, which is increasingly vital as object category scales grow. Additionally, its energy-based aggregation algorithm fails with discontinuous multimodal distributions. We introduce GenPose++, which incorporates a 2D foundation model to leverage RGB semantics, improving generalization, and introduces an aggregation module to handle discrete multimodal distributions, addressing the limitations of GenPose.

6D object pose tracking. This paper is situated within the domain of category-level 6D object pose tracking and model-free object tracking. BundleTrack [31] pioneers model-free tracking by leveraging multi-view feature detection for tracking unseen objects without 3D models. CAPTRA [32] enhances articulated pose tracking through recursive updates for better temporal consistency. CATRE [18] aligns partially observed point clouds to abstract shape priors for relative transformations and pose estimation. GenPose, adopting a generative approach, effectively addresses the challenge of pose ambiguities in symmetric objects. Together, these approaches underscore the evolving landscape of object pose tracking, highlighting both the progress and the diversity of strategies being explored.

3 Omni6DPose Dataset

This paper introduces a rich variety of object categories, a large-scale, and materially diverse dataset for real 6D object pose estimation, named ROPE. And, a simulated dataset, SOPE, synthesizing with mixed reality and featuring depth simulation, is provided for training. Section 3.1 will discuss the collection and alignment of 3D objects. Section 3.2 will cover the acquisition and labeling of the ROPE dataset. Section 3.3 will explain the generation of the SOPE dataset.

3.1 3D Object Collection and Alignment

Universal 6D object pose estimation relies on a comprehensive set of objects. We selected 149 categories of everyday objects, all reconstructed with high-precision scanners, and categorized them into two sets: SOPE for simulated data and ROPE for real-world scenes. SOPE primarily includes objects from sources like OmniObject3D [35], PhoCal [30], and GoogleScan [7], alongside a subset from our scans, totaling 5,000 instances. ROPE consists of 580 instances we reconstructed using industrial scanners. Importantly, while most SOPE objects are from public datasets, manual category-level pose alignment is necessary. For object reconstruction, as shown in Figure 3, we use two professional scanners, EinScan H2[7] and Revopoint POP 3 [8] for objects in different scales. The scan-

[7] `https://www.einscan.com/`
[8] `https://www.revopoint3d.com/`

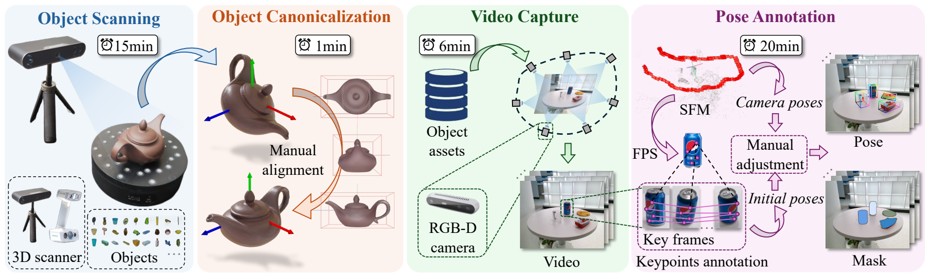

Fig. 3: ROPE dataset collection and annotation. (1) Object scanning, where high-precision industrial scanners are used to acquire the CAD models of objects; (2) Object canonicalization, involving the alignment of each object category to the canonical space; (3) Video capture, capturing video sequences in varied scenarios with a depth camera; and (4) Pose annotation, calculating camera poses through Structure from Motion (**SFM**), further utilizing Farthest Point Sampling (**FPS**) to select keyframes for keypoint annotation, and performing bundle adjustment to derive initial object pose values, which are then manually refined to obtain more precise annotations.

ning time depends on object features: it took about 15 minutes for small, simple, Lambertian items like a mouse, and up to an hour for complex, large, or non-Lambertian items like a transparent mug. Finally, we constructed a specialized annotation tool for manually aligning objects in the same category to the category-level canonical space, with each alignment taking roughly one minute.

3.2 ROPE Acquisition and Annotation

The ROPE dataset was systematically acquired utilizing the RealSense D415 imaging device, encompassing scenarios with 2 to 6 distinct objects and video sequences extending from 762 to 1,349 frames. The integrity and utility of the dataset are underpinned by the precision of object pose annotations, which present notable challenges, chiefly:

1. Derivation of relative camera poses, denoted as $T_c = \{(R_i, t_i)\}_{i=1}^{n}$, where each pair $(R_i, t_i) \in \mathrm{SE}(3)$, signifying the transformation from the camera space to the world space for the i^{th} frame, with n symbolizing the aggregate frame count within the video sequence.
2. Procurement of high-accuracy object poses, represented as $T_o = \{(R, t)\}$, where the pair $(R, t) \in \mathrm{SE}(3)$, delineating the transformation from the object space the camera space for any selected frame.

With T_c and T_o known, it is possible to automate the generation of all annotations within the dataset. Addressing these challenges, we propose a marker-free annotation system. Previous datasets for calculating relative camera pose rely on markers, like NOCS [28], or external robot arms for indirect calculation, such as PhoCal [30], limiting scene diversity. As shown in Figure 3, to enable marker-free annotation in open scenes, we consider it a structure-from-motion (SfM) problem, utilizing intrinsic scene features to optimize camera poses T_c using bundle adjustment techniques. This approach aims to solve the optimization problem:

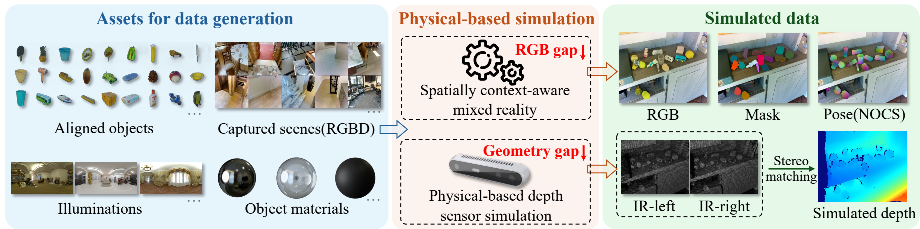

Fig. 4: SOPE synthesis, utilizing mixed reality to bridge the RGB sim2real gap and physical-based depth sensor simulation to minimize the geometric sim2real gap.

$$\min_{T_c} \sum_{i=1}^{n} \sum_{j \in \mathcal{P}_i} \|\pi(R_i X_j + t_i) - x_{ij}\|^2 \tag{1}$$

where π denotes the camera projection function, X_j represents the 3D points in the world space, and x_{ij} corresponds to the 2D projection of X_j in the i^{th} camera frame, with $\mathcal{P}_i$ being the set of point correspondences in frame i.

For object pose T_o annotations, previous methods consider only a single frame, leading to inaccuracies due to the lack of multi-viewpoint constraints. To overcome this, we introduce a two-stage object pose annotation process: in the first stage, keyframes are sampled from SfM results using Farthest Point Sampling (FPS), and 2D-3D keypoint pairs on these keyframes are annotated, providing initial object pose by minimizing the reprojection error:

$$\min_{R,t} \sum_{k \in \mathcal{K}} \|\pi(RX_k + t) - x_k\|^2 \tag{2}$$

where X_k are 3D keypoints of object, x_k are corresponding 2D annotations in the image, and $\mathcal{K}$ is the set of all keypoint correspondences. The initial poses are manually fine-tuned based on the object's projection across all keyframes.

3.3 SOPE Synthesis

ROPE represents a comprehensive benchmark in category-level object pose estimation by scaling up the diversity and number of object categories to unprecedented levels, encompassing a wide range of materials. This diversity presents new challenges for network training data due to the higher demands on the dataset's scale and diversity. Collecting a larger real-world dataset would be prohibitively expensive and unlikely to ensure sufficient diversity.

To bridge the sim2real gap, which is pronounced when using synthetic data, either in RGB or geometry, this paper proposes a novel method based on mixed reality with depth simulation for synthetic data generation. Specifically, as demonstrated in Figure 4, we employ mixed reality [28] techniques to generate RGB data, thereby reducing the RGB sim2real gap. In parallel, we simulate the mechanism of structured light depth sensors within blender [6]. This involves rendering infrared (IR) images and applying stereo matching to produce synthetic depth maps, effectively narrowing the geometry's sim2real gap.

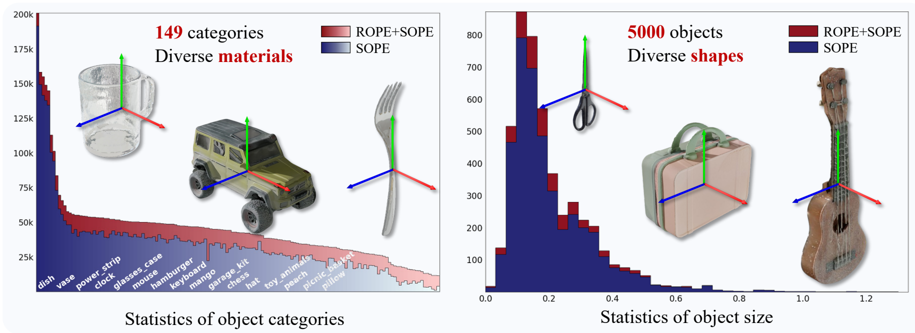

Fig. 5: Omni6DPose statistics, showcasing the dataset distribution. Left: Category distribution, highlighting 149 categories and diverse materials. Right: Object size distribution across 5000 objects, illustrating diversity in shapes.

During data generation, we implement domain randomization for illumination and object materials to further enhance the dataset's diversity. All the background images are sourced from public datasets, including 19,658 images from MatterPort3D [2], 2,572 from ScanNet++ [38], and 540 from IKEA [28]. To the best of our knowledge, this is the first simulated dataset that uses a context-aware mixed reality approach combined with physical-based depth sensor simulation for object pose estimation tasks.

3.4 Dataset Statistics

Object Category Statistics The comprehensive distribution of object category and size are both demonstrated in Figure 5. Most of the categories possess $\geq$ 25K pose annotations in the SOPE dataset, providing sufficient training opportunities. Categories containing objects with diverse and challenging material options (e.g., transparent or specular materials) are equipped with apparently more data generation, such as dishes, cups, bottles, bowls, mugs, etc.

Object Size Statistics As shown in Figure 5, the objects in our dataset span a wide range of sizes. The majority of the objects are approximately 0.1 meters in length along the diagonal of their bounding boxes, with the largest objects exceeding 1 meter.

4 Category-level 6D Pose Estimation Method

Given the Omni6DPose, one naturally ponders the optimal technical approach for large-scale category-level pose estimation. The recently introduced state-of-the-art category-level 6D pose estimation technique, GenPose [43], offers a promising avenue by employing a diffusion-based probabilistic method [23, 24]. Moreover, the diffusion model has demonstrated remarkable efficacy across various high-dimensional domains with extensive training data [5, 25, 34, 42]. Expanding on this groundwork, our study delves further into the probabilistic approach, presenting an enhanced iteration of GenPose, named GenPose++.

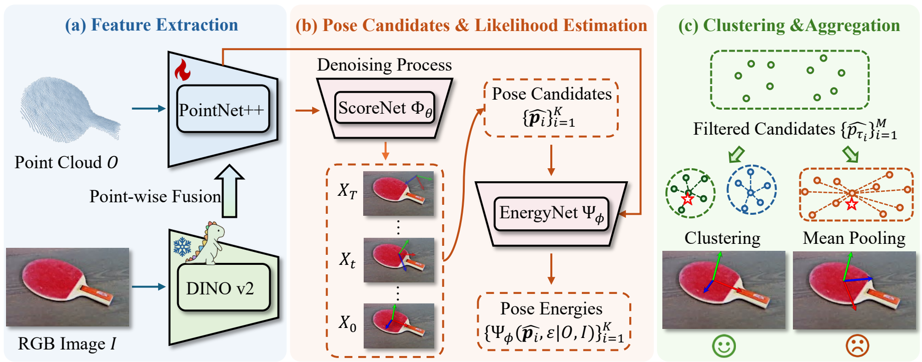

Fig. 6: Overview of GenPose++. We employ segmented point clouds and cropped RGB images as inputs. PointNet++ extracts geometric features, while DINO v2, a pre-trained 2D foundation model, extracts semantic features. These features are fused to condition a diffusion model, which generates object pose candidates and their corresponding energy. Finally, clustering is applied to address aggregation issues in the multimodal distribution of poses for objects with non-continuous symmetry, such as boxes, effectively resolving the pose estimation challenge.

GenPose++ integrates two crucial enhancements: Semantic-aware feature extraction (see Fig.6 (a)) and Clustering-based aggregation (as shown in Fig.6 (c)). The subsequent sections will detail the three primary stages of the GenPose++ pipeline. Moreover, given the estimated 6D pose, GenPose++ provide a additional regression network to predict the 3D scale of the object.

4.1 Training Semantic-aware Score and Energy Networks

The learning agent is trained on our paired dataset $\mathcal{D} = \{(\boldsymbol{p}_k, O_k, I_k)\}_{k=1}^n$, where $\boldsymbol{p}_k \in \mathrm{SE}(3)$, $O_k \in \mathbb{R}^{3\times N}$, and $I_k \in \mathbb{R}^{3\times H\times W}$ denote a 6D pose, a partially observed 3D point cloud with N points, and a cropped RGB image with $H \times W$ resolution, respectively. Given an unseen object with point cloud O^* and RGB image I^*, the goal is to recover the corresponding ground-truth pose $\boldsymbol{p}^*$.

Following GenPose, we initially train a score network $\boldsymbol{\Phi}_\theta : \mathbb{R}^{|\mathcal{P}|} \times \mathbb{R}^1 \times \mathbb{R}^{3\times N} \times \mathbb{R}^{3\times H\times W} \rightarrow \mathbb{R}^{|\mathcal{P}|}$ and an energy network $\boldsymbol{\Psi}_\phi : \mathbb{R}^{|\mathcal{P}|} \times \mathbb{R}^1 \times \mathbb{R}^{3\times N} \times \mathbb{R}^{3\times H\times W} \rightarrow \mathbb{R}^1$ from the dataset $\mathcal{D}$ using the denoising score-matching objective [26]:

$$\mathbb{E}_{t\sim\mathcal{U}(\epsilon,1)} \left\{ \lambda(t) \mathbb{E}_{\substack{\boldsymbol{p}(0)\sim p_{\mathrm{data}}(\boldsymbol{p}(0)|O,I),\\ \boldsymbol{p}(t)\sim\mathcal{N}(\boldsymbol{p}(t);\boldsymbol{p}(0),\sigma^2(t)\mathbf{I})}} \left[\left\| \boldsymbol{s}(\boldsymbol{p}(t), t|O, I) - \frac{\boldsymbol{p}(0) - \boldsymbol{p}(t)}{\sigma(t)^2} \right\|_2^2 \right] \right\} \tag{3}$$

The training loss of the score and energy network can be obtained by replacing $\boldsymbol{s}(\boldsymbol{p}(t), t|O, I)$ in Eq. 3 with $\boldsymbol{\Phi}_\theta(\boldsymbol{p}(t), t|O, I)$ and $\nabla_{\boldsymbol{p}} \boldsymbol{\Psi}^*_\phi(\boldsymbol{p}(t), t|O, I)$, respectively.

Unlike GenPose, our score and energy network are semantic-aware, as both networks are conditioned on an RGB image to incorporate semantic cues for pose estimation. To fuse the features extracted from the image and point cloud, we encode the RGB image img and point cloud obj using the pre-trained feature extractors from DINOv2 [20] and PointNet++ [22], respectively. Then, we concatenate these features together in a pointwise manner similar to [27].

4.2 Candidates Generation and Outlier Removal

Following GenPose, we subsequently sample pose candidates $\{\hat{\boldsymbol{p}}_i\}_{i=1}^K$ by solving the following *Probability Flow* (PF) ODE [24] constructed by the score network $\boldsymbol{\Phi}_\theta$ from $t = 1$ to $t = \epsilon$:

$$\frac{d\boldsymbol{p}}{dt} = -\sigma(t)\dot{\sigma}(t)\boldsymbol{\Phi}_\theta(\boldsymbol{p}(t), t|O, I) \tag{4}$$

where $\boldsymbol{p}(1) \sim \mathcal{N}(\mathbf{0}, \sigma_{\max}^2\mathbf{I})$, $\sigma(t) = \sigma_{\min}(\frac{\sigma_{\max}}{\sigma_{\min}})^t$, $\sigma_{\min} = 0.01$ and $\sigma_{\max} = 50$.

To remove the outliers in candidates, we sort the candidates into a sequence $\hat{\boldsymbol{p}}_{\tau_1} \succ \hat{\boldsymbol{p}}_{\tau_2} ... \succ \hat{\boldsymbol{p}}_{\tau_K}$ where:

$$\hat{\boldsymbol{p}}_{\tau_i} \succ \hat{\boldsymbol{p}}_{\tau_j} \iff \boldsymbol{\Psi}_\phi(\hat{\boldsymbol{p}}_{\tau_i}, \epsilon|O) > \boldsymbol{\Psi}_\phi(\hat{\boldsymbol{p}}_{\tau_j}, \epsilon|O) \tag{5}$$

Then, we filter out the last $1 - \delta\%$ candidates and obtain $\hat{\boldsymbol{p}}_{\tau_1} \succ \hat{\boldsymbol{p}}_{\tau_2} ... \succ \hat{\boldsymbol{p}}_{\tau_M}$ where $\delta = 40\%$ is a hyper parameter and $M = \lfloor \delta \cdot K \rfloor$.

4.3 Clustering-based Aggregation

In this section, we aggregate the remaining candidates $\{\hat{\boldsymbol{p}}_{\tau_i} = (\hat{T}_{\tau_i}, \hat{R}_{\tau_i})\}_{i=1}^M$ to obtain the final results. GenPose achieves this by simply mean-pooling the filtered candidates. However, this strategy will encounter a severe *mean-mode issue* when the object possesses discrete symmetrical properties. As illustrated in Fig. 6, a ping pong paddle has two symmetric ground truth poses (modes). Since the score network has encountered both modes during training, an optimally trained score network will likely output candidates around both modes. Simply mean-pooling these candidates will yield the average of the two modes, known as the 'mean mode', which will deviate from both modes.

To mitigate this issue, we introduce a clustering-based aggregation mechanism. We employ DBSCAN [8] to cluster the candidates. It identify dense regions in the data space, forming clusters based on the density of data points and effectively separating noise from meaningful patterns, without the need to specify the number of clusters. This is achieved through dynamic determination of cluster quantities based on distance threshold (ε) and density threshold (MinPts). For instance, in our empirical setting, we set $\varepsilon \approx 0.45$rad and MinPts $= 5$. After clustering the candidates, we select the cluster with the largest number of objects and get the mean-pooling result as the final estimation following GenPose.

5 Experiments

5.1 Category-Level 6D Object Pose Estimation

Metric In prior studies, metrics such as the mean average precision (mAP) for 3D bounding box IoU and the mean average precision (mAP) for objects with translation errors less than m cm and rotation errors less than $n°$ have been commonly used [28]. This evaluation typically involves two steps: instance-level object segmentation and pose estimation from these detections. However, these metrics are influenced by the detection model's performance. To focus solely on evaluating the precision of pose estimation, we assume ground truth instance segmentation is known and propose the following two metrics:

Table 2: Quantitative comparison of category-level pose estimation on ROPE. ↑ represents a higher value indicating better performance, while ↓ represents a lower value indicating better performance. **Prior-free** indicates whether the method requires category prior. '-' indicates that GenPose does not predict the object scale.

Method		Input	Prior-free	AUC ↑			VUS ↑			
				IoU_{25}	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
Deterministic	NOCS [28]	RGBD	✓	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SGPA [3]	RGBD	✗	10.5	2.0	0.0	4.3	6.7	9.3	15.0
	IST-Net [16]	RGBD	✓	28.7	10.6	0.5	2.0	3.4	5.3	8.8
	HS-Pose [44]	D	✓	31.6	13.6	1.1	3.5	5.3	8.4	12.7
Probabilistic	GenPose [43]	D	✓	-	-	-	6.6	9.6	13.1	19.3
	GenPose++(Ours)	RGBD	✓	**39.0**	**19.1**	**2.0**	**10.0**	**15.1**	**19.5**	**29.4**

- **AUC@IoU**$_n$: This metric assesses the accuracy of predicted 3D bounding boxes, calculated via the Area Under the Curve (AUC) from various Intersection over Union (IoU) thresholds starting at n. In our study, we utilize AUC@IoU$_{25}$, AUC@IoU$_{50}$, and AUC@IoU$_{75}$ as the benchmarks.
- **VUS@**$n°m$**cm**: This metric offers a detailed analysis of 6D pose estimation accuracy, derived from the Volume Under Surface (VUS) across ranges of rotational (up to $n°$) and translational (up to $m$cm) error thresholds. VUS aggregates the accuracy of pose predictions within set boundaries. In this paper, we apply VUS@5°2cm, VUS@5°5cm, VUS@10°2cm, and VUS@10°5cm for comprehensive performance assessment.

Additionally, to demonstrate the robustness of our method to segmentation masks, we present the pose estimation performance of GenPose++ using SAM [13] segmentation results in the supplementary material.

Baselines We evaluate five category-level pose estimation methods: NOCS [28], SGPA [3], HS-Pose [44], IST-Net [16] and GenPose [43]. Except for NOCS, which conducts both object detection and pose estimation as a whole, all methods are equipped with ground truth detection results. For SGPA, the prior point cloud of each category is constructed by randomly selecting an object from the training dataset. Considering that previous methods' augmentations for symmetry properties are only applicable to specific object categories within the NOCS dataset and not suitable for all object categories in Omni6DPose, data augmentation for object symmetry is disabled during training for all baseline methods. All methods are trained on SOPE and directly tested on ROPE.

Results and Analysis. In Table 2, we present the quantitative evaluation results of previous methods compared to GenPose++ on the ROPE dataset. Overall, generative methods continue to dominate in the performance evaluation on ROPE. The VUS surface depicted in Figure 7 provides a more detailed reflection of the performance of each model. Unlike deterministic approaches, generative methods can handle ambiguity without any specialized design requirements. Moreover, these methods directly generate the distribution of object poses, eliminating the need for depth map-based pose fitting. This approach is particularly

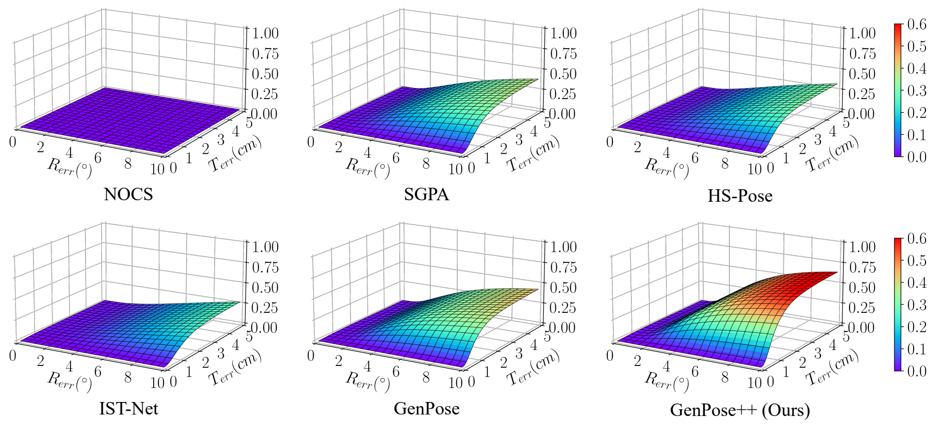

Fig. 7: Qualitative comparison with baselines on ROPE dataset.

advantageous for challenging material types, such as transparent or reflective objects, where structured-light depth cameras tend to introduce significant noise, severely impacting pose fitting accuracy. Furthermore, the NOCS method does not demonstrate effective performance on the ROPE dataset, leading to the supposition that methods relying solely on RGB information to predict the shape of an object in the canonical space become less robust as the scale of category diversity increases. Compared to GenPose, GenPose++ achieves a significant lead by leveraging the powerful perception capabilities of the 2D foundation model, along with the robustness of clustering towards discrete symmetric properties. You can find qualitative visualizations in Figure 8.

5.2 Ablation Study

In order to validate the design decisions of our approach, we performed a series of ablation experiments on our method:

– **w/o clustering.** Without clustering directly take the average of all remaining pose candidates after outlier removal as the pose estimation output.
– **w/o scale prediction.** Use the estimated pose to transform the observed point cloud into object space, then take the bounding box length as the maximum projection from the point cloud to each axis.
– **w/o simulated depth.** Use perfect depth for training.
– **w/o point-wise feature fusion.** In the feature extraction stage, separately extract the RGB feature and geometric feature, and then concatenation.

Table 3 illustrates the contribution of each component of GenPose++ to its performance. The introduction of the clustering module allows GenPose++ to effectively aggregate the multimodal distributions caused by discrete symmetries, leading to higher performance. The scale prediction in GenPose++ significantly surpasses direct calculations from the object's point cloud due to ambiguities from partial observations and errors from point cloud noise, particularly in transparent and reflective objects. Training with simulated depth data results in

Table 3: Ablation study on category-level 6D object pose estimation

Ablation	AUC ↑			VUS ↑			
	IoU_{25}	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
Full	**39.0**	**19.1**	**2.0**	**10.0**	**15.1**	**19.5**	**29.4**
w/o scale prediction	13.8	3.4	0.2	**10.0**	**15.1**	**19.5**	**29.4**
w/o clustering	38.6	18.7	1.9	9.4	14.1	18.4	27.8
w/o simulated depth	35.3	16.5	1.7	7.3	11.9	14.6	24.3
w/o point-wise fusion	34.2	15.4	1.4	7.1	10.6	14.6	21.9

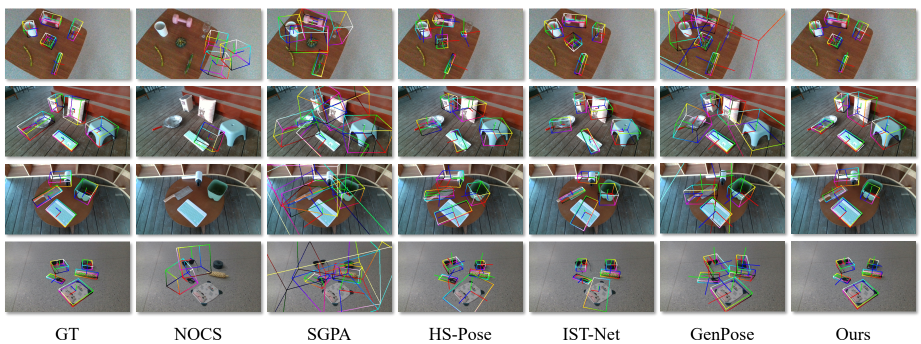

Fig. 8: Qualitative comparison with baselines on ROPE dataset.

better performance than training with perfect point clouds, as the physics-based depth camera simulation substantially reduces the sim2real gap for depth data. Point-wise fusion outperforms global fusion as it retains more of the object's local geometric features, which are crucial for accurate object pose prediction.

5.3 Category-Level 6D Object Pose Tracking

Metric. We report the following metrics for object pose tracking evaluation:

- **FPS**: Frames Per Second, which indicates the speed of pose tracking.
- **VUS@5°5cm**: Volume Under Surface, assessing pose estimation accuracy for rotation errors within 0 to 5° and translation errors within 0 to 5cm.
- **mIoU**: Mean Intersection over Union, representing the average 3D overlap between the ground truth and the predicted bounding boxes.
- **Rerr(°)**: Average rotation error in degrees.
- **Terr(cm)**: Average translation error in centimeters.

Baselines. This paper employs BundleTrack, CATRE, and GenPose as baselines for object pose tracking. BundleTrack is a training-free approach that utilizes multi-view feature point detection and matching for tracking the pose of unseen objects. CATRE aligns partially observed point clouds to abstract shape priors to estimate relative transformations, enhancing pose accuracy. Conversely, GenPose utilizes a generative approach to effectively resolve pose ambiguities, notably in symmetric objects. Following GenPose, we have adapted GenPose++ to object pose tracking with minor modifications.

Results and Analysis. Table 4 presents the results of category-level object pose tracking algorithms CATRE, GenPose, and our method, as well as the results for the unseen object pose tracking algorithm BundleTrack. The category-level pose estimation methods appear to achieve relatively better outcomes since they benefit from learning the category-level canonical space of objects within the SOPE dataset. However, for the training-free BundleTrack, reliance on RGB information for keypoint detection and matching poses challenges, often failing on objects with weak textures. Additionally, its dependency on depth values for global optimization renders it less effective in handling instances with substantial depth noise, such as transparent or specular objects. Our method, without undergoing specialized design, has achieved results comparable to state-of-the-art approaches. Although the inference speed of our method is lower than that of CATRE and GenPose, the achieved 17.8 FPS is sufficient for certain downstream tasks, such as robotic manipulation. Furthermore, recent rapid developments in research on fast samplers are beneficial to our approach, potentially enhancing its performance and applicability in real-time scenarios.

Table 4: Results of category-level object pose tracking on ROPE. The results are averaged over all 149 categories. *GT. Pert.* denotes that a perturbed ground truth pose is utilized as the initial object pose.

Methods	Input	Init.	Speed(FPS)↑	5°5cm↑	mIoU↑	R_{err}(°)↓	T_{err}(cm)↓
BundleTrack [31]	RGBD	GT.	12.4	1.3	3.9	46.9	23.5
CATRE [18]	D	GT. Pert.	**38.5**	**15.9**	**55.4**	21.3	2.6
GenPose [43]	D	GT. Pert.	26.3	13.3	-	19.3	**1.2**
GenPose++(Ours)	RGBD	GT. Pert.	17.8	**15.9**	53.4	**17.6**	**1.2**

6 Conclusions and Discussion

In this study, we introduce Omni6DPose, a comprehensive dataset for 6D object pose estimation, featuring extensive scale, diversity, and material variety. Through thorough experimentation, our findings suggest that the probabilistic framework holds promise for category-level 6D object estimation, leveraging semantic information provided by RGB images to address large-scale pose estimation challenges. However, the performance of GenPose++ on Omni6DPose reveals significant room for improvement, with the model still hampered by slow inference speeds resulting from the iterative refinement nature inherent in the diffusion model. Future works could focus on addressing these challenges and integrating the universal 6D pose estimation module trained on Omni6DPose into a broader range of downstream tasks [4, 33, 37, 42].

7 Acknowledgment

This work is supported by the National Natural Science Foundation of China - General Program (Project ID: 62376006), the National Youth Talent Support Program (Project ID: 8200800081), and NSFC-62306046.

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 (2023)
2. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017)
3. 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. pp. 2773–2782 (2021)
4. Cheng, J., Wu, M., Zhang, R., Zhan, G., Wu, C., Dong, H.: Score-pa: Score-based 3d part assembly. arXiv preprint arXiv:2309.04220 (2023)
5. Ci, H., Wu, M., Zhu, W., Ma, X., Dong, H., Zhong, F., Wang, Y.: Gfpose: Learning 3d human pose prior with gradient fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4800–4810 (2023)
6. Dai, Q., Zhang, J., Li, Q., Wu, T., Dong, H., Liu, Z., Tan, P., Wang, H.: Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In: European Conference on Computer Vision. pp. 374–391. Springer (2022)
7. 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)
8. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: kdd. vol. 96, pp. 226–231 (1996)
9. Fu, Y., Wang, X.: Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. In: Advances in Neural Information Processing Systems (2022)
10. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. vol. 7724 (10 2012). https://doi.org/10.1007/978-3-642-33885-4_60
11. Hodaň, 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. IEEE Winter Conference on Applications of Computer Vision (WACV) (2017)
12. Jung, H., Zhai, G., Wu, S.C., Ruhkamp, P., Schieber, H., Rizzoli, G., Wang, P., Zhao, H., Garattoni, L., Meier, S., Roth, D., Navab, N., Busam, B.: Housecat6d – a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios (2023)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)
14. Lin, Z.H., Huang, S.Y., Wang, Y.C.F.: Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
15. Liu, J., Sun, W., Liu, C., Zhang, X., Fan, S., Wu, W.: Hff6d: Hierarchical feature fusion network for robust 6d object pose tracking. IEEE Transactions on Circuits

and Systems for Video Technology pp. 1–13 (06 2022). https://doi.org/10.1109/TCSVT.2022.3181597
16. Liu, J., Chen, Y., Ye, X., Qi, X.: Prior-free category-level pose estimation with implicit space transformation. arXiv preprint arXiv:2303.13479 (2023)
17. Liu, X., Iwase, S., Kitani, K.M.: Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In: ICCV (2021)
18. Liu, X., Wang, G., Li, Y., Ji, X.: CATRE: iterative point clouds alignment for category-level object pose refinement. In: European Conference on Computer Vision (ECCV) (October 2022)
19. 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 (2016). https://doi.org/10.1109/TVCG.2015.2513408
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
21. Piga, N.A., Onyshchuk, Y., Pasquale, G., Pattacini, U., Natale, L.: Roft: Real-time optical flow-aided 6d object pose and velocity tracking. IEEE Robotics and Automation Letters **7**(1), 159–166 (2022). https://doi.org/10.1109/LRA.2021.3119379
22. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. Advances in neural information processing systems **30** (2017)
23. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. Advances in Neural Information Processing Systems **32** (2019)
24. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
25. Tian, Y., Zhang, J., Huang, G., Wang, B., Wang, P., Pang, J., Dong, H.: Robokeygen: Robot pose and joint angles estimation via diffusion-based 3d keypoint generation. arXiv preprint arXiv:2403.18259 (2024)
26. Vincent, P.: A connection between score matching and denoising autoencoders. Neural computation **23**(7), 1661–1674 (2011)
27. 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)
28. 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: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
29. 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. IEEE (2021)
30. Wang, P., Jung, H., Li, Y., Shen, S., Srikanth, R.P., Garattoni, L., Meier, S., Navab, N., Busam, B.: Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21222–21231 (2022)
31. Wen, B., Bekris, K.E.: Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2021)

32. Weng, Y., Wang, H., Zhou, Q., Qin, Y., Duan, Y., Fan, Q., Chen, B., Su, H., Guibas, L.J.: Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 13209–13218 (October 2021)
33. Wu, M., Zhong, F., Xia, Y., Dong, H.: Targf: Learning target gradient field to rearrange objects without explicit goal specification. Advances in Neural Information Processing Systems **35**, 31986–31999 (2022)
34. Wu, T., Wu, M., Zhang, J., Gan, Y., Dong, H.: Learning score-based grasping primitive for human-assisting dexterous grasping. Advances in Neural Information Processing Systems **36** (2024)
35. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 803–814 (2023)
36. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes (2018)
37. Xue, T., Wu, M., Lu, L., Wang, H., Dong, H., Chen, B.: Learning gradient fields for scalable and generalizable irregular packing. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–11 (2023)
38. 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. pp. 12–22 (2023)
39. You, Y., He, W., Liu, J., Xiong, H., Wang, W., Lu, C.: Cppf++: Uncertainty-aware sim2real object pose estimation by vote aggregation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
40. 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. pp. 6866–6875 (2022)
41. You, Y., Xiong, K., Yang, Z., Huang, Z., Zhou, J., Shi, R., Fang, Z., Harley, A.W., Lu, C.: Pace: Pose annotations in cluttered environments. arXiv preprint arXiv:2312.15130 (2023)
42. Zeng, Y., Wu, M., Yang, L., Zhang, J., Ding, H., Cheng, H., Dong, H.: Distilling functional rearrangement priors from large models. arXiv preprint arXiv:2312.01474 (2023)
43. Zhang, J., Wu, M., Dong, H.: Generative category-level object pose estimation via diffusion models. Advances in Neural Information Processing Systems **36** (2024)
44. Zheng, L., Wang, C., Sun, Y., Dasgupta, E., Chen, H., Leonardis, A., Zhang, W., Chang, H.J.: Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 17163–17173 (June 2023)
45. Zhou, L., Liu, Z., Gan, R., Wang, H., au2, M.H.A.J.: Dr-pose: A two-stage deformation-and-registration pipeline for category-level 6d object pose estimation (2023)