

Exploring Category-level Articulated Object Pose Tracking on SE(3) Manifolds

Xianhui Meng^{1*}, Yukang Huo^{2*}, Li Zhang^{1,5†}, Liu Liu³, Haonan Jiang⁷, Yan Zhong⁴,
Pingrui Zhang^{6,8}, Cewu Lu⁹, Jun Liu^{1†}

¹Department of Electronic Engineering and Information Science, University of Science and Technology of China, China

²China Agricultural University, China. ³Hefei University of Technology, Hefei, China. ⁴Peking University, China.

⁵Hefei Institute of Physical Science, Chinese Academy of Sciences, China. ⁶Fudan university, China.

⁷Zhejiang University of Technology, Zhejiang, China. ⁸Shanghai AI Lab, China. ⁹Shanghai Jiao Tong University, China
{mengxh, zanly20}@mail.ustc.edu.cn, yukanghuo.ai@gmail.com, junliu@ustc.edu.cn.

Abstract

Articulated objects are prevalent in daily life and robotic manipulation tasks. However, compared to rigid objects, pose tracking for articulated objects remains an underexplored problem due to their inherent kinematic constraints. To address these challenges, this work proposes a novel point-pair-based pose tracking framework, termed **PPF-Tracker**. The proposed framework first performs quasi-canonicalization of point clouds in the SE(3) Lie group space, and then models articulated objects using Point Pair Features (PPF) to predict pose voting parameters by leveraging the invariance properties of SE(3). Finally, semantic information of joint axes is incorporated to impose unified kinematic constraints across all parts of the articulated object. PPF-Tracker is systematically evaluated on both synthetic datasets and real-world scenarios, demonstrating strong generalization across diverse and challenging environments. Experimental results highlight the effectiveness and robustness of PPF-Tracker in multi-frame pose tracking of articulated objects. We believe this work can foster advances in robotics, embodied intelligence, and augmented reality. Codes are available at <https://github.com/mengxh20/PPFTracker>.

Introduction

Domestic service robots must perceive and understand diverse 3D objects in complex human-centric environments (Mo et al. 2021). Articulated objects (e.g., cabinets with doors and drawers) pose particular challenges due to structural complexity and semantic richness. Accurate six-degree-of-freedom (6-DoF) pose estimation (Wen et al. 2023) is fundamental to downstream applications in embodied intelligence, VR/AR (Davis and Aslam 2024; Zhao et al. 2024a), human-computer interaction (Clark, Newman, and Dutta 2022; Yang et al. 2022), and robotic manipulation (Jain et al. 2021; Zhao et al. 2025b). Reliable 6-DoF pose enables robust interpretation and interaction, supporting sophisticated tasks in dynamic, unstructured environments.

*Xianhui Meng and Yukang Huo contributed equally.

Corresponding authors[†]: Li Zhang and Jun Liu.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

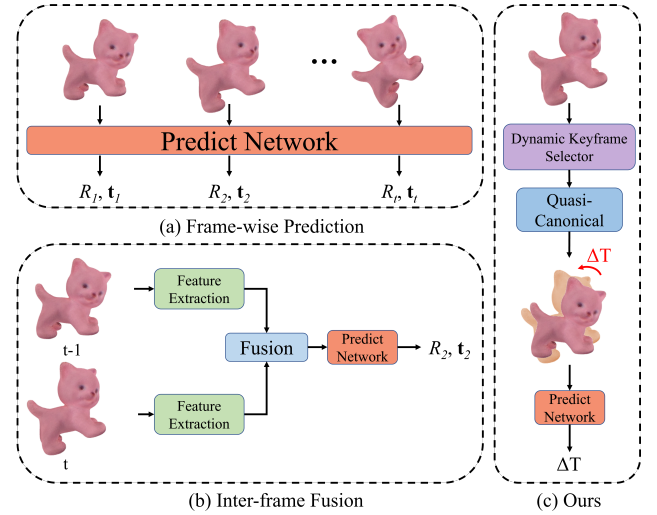


Figure 1: The Categorization of Tracking Methods.

Compared to instance-level tasks (Mao et al. 2021; Wen and Bekris 2021; Wang et al. 2018), category-level 6D object pose estimation presents significantly greater challenges (Heppert et al. 2022; Fu and Wang 2022; Zhao et al. 2025a, 2024b). In category-level scenarios, models must estimate the 3D rotation and translation of previously unseen objects based on partial observations, without access to specific CAD models or instance-level priors. While recent methods (Zhang, Wu, and Dong 2024; Zou et al. 2024) have made encouraging progress in this domain, several critical challenges remain:

(1) **Pose Invalidity and Singularity.** Many existing methods optimize SE(3)-related parameters (e.g., Euler angles or translations) in Euclidean space, which can lead to invalid rotation matrices that violate orthogonality constraints. Moreover, Euler angles suffer from gimbal lock, while quaternions have sign ambiguity, both of which can cause singularities and unstable pose predictions.

(2) **Tracking Methods.** Traditional approaches often depend on dense predictions or frame-by-frame pose updates

(Fig. 1 (a,b)) (You, Yao, and Xu 2020; Wang et al. 2012), which are computationally costly and unsuitable for real-time point cloud streams. These limitations hinder robust and continuous tracking in dynamic environments.

To address the first challenge, we represent object poses using the Lie group $SE(3)$ and perform optimization in its tangent space $\mathfrak{se}(3)$, ensuring geometric consistency. Based on this formulation, we design a framework that leverages $SE(3)$ -invariance for robust pose tracking. We use **Point Pair Features (PPF)** to encode relative pose information, and introduce a **kinematics-aware optimization module** to enforce structural constraints among parts, improving overall accuracy and consistency.

To address the second challenge, we incorporate temporal priors from adjacent frames to guide pose prediction (Fig. 1(c)). At each time step, features are extracted and object states updated using information from previous frames. This inter-frame fusion improves tracking stability and reduces computational overhead, enabling efficient real-time performance on continuous point cloud streams.

We conduct a comprehensive evaluation of the proposed **PPF-Tracker** across multiple datasets. The results demonstrate that PPF-Tracker achieves strong performance on both synthetic dataset (PM-Videos) and semi-synthetic dataset (ReArt-Videos). Additionally, its consistent success in real-world scenarios (RobotArm-Videos) further highlights the method’s robustness and high generalization capability across diverse environments, validating its effectiveness in practical applications. The key contributions of this work are summarized as follows:

- We propose a customized algorithm for category-level articulated object pose estimation, aiming to advance the tracking performance on the $SE(3)$ manifolds.
- We propose a weighted point pair algorithm to predict $SE(3)$ -invariant parameters and estimate pose increments in the associated Lie algebra $\mathfrak{se}(3)$. To respect the kinematic constraints of articulated objects, an axis-based part-level refinement is further introduced for improved accuracy and consistency.
- Extensive experiments on both point cloud and RGB-D datasets demonstrate the superior performance and efficiency of our method. Moreover, real-world evaluations confirm its strong generalization capability across diverse and challenging environments.

RELATED WORK

★ **Category-Level Object Pose Estimation.** Category-level object pose estimation aims to predict the 6D pose (rotation and translation) and scale of previously unseen instances within known categories (Zhang et al. 2025b; Liu et al. 2022b; Zhang et al. 2025a). Early works like NOCS (Wang et al. 2019) proposed normalized coordinate spaces to enhance generalization. DualPoseNet (Lin et al. 2021) introduced dual decoders for joint implicit and explicit modeling. Recent methods (Lin et al. 2022; Irshad et al. 2022; You et al. 2022b; Fernandez-Labrador et al. 2020) leverage shape priors and keypoint voting to handle

intra-class variation, but often rely on extensive real or synthetic training data (Liu et al. 2022a; You et al. 2022a), limiting scalability in complex real-world environments (Di et al. 2022; Zhang, Wu, and Dong 2024). While self-supervised approaches (Deng et al. 2020; Manhardt et al. 2020; Chen et al. 2024) reduce annotation dependency, they still suffer from domain shifts (Sankaranarayanan et al. 2018) and background assumptions (Hurlburt and Schwitzgebel 2011).

★ **Articulated Object Pose Tracking.** Articulated pose tracking extends the estimation task to dynamic, multi-frame settings (Zhang et al. 2024, 2021). Instance-level methods such as BundleTrack (Wen and Bekris 2021) and PA-Pose (Liu et al. 2024) rely on segmentation and alignment optimization, while keypoint-based methods (Maji et al. 2022; Li et al. 2021b; Prakhyia et al. 2015; Jau et al. 2020) struggle with occlusion and articulation variability. RP-MART (Wang et al. 2024) combines deep segmentation with robust features for category-level tracking, but high computational cost limits real-time applicability. A major drawback of existing methods lies in their frame-by-frame processing, which ignores inter-frame motion continuity and structural consistency in $SE(3)$.

Problem Statement

This work addresses the category-level articulated object pose tracking problem on the $SE(3)$ manifolds. Given a sequence of frames $\{P_t\}_{t \geq 0}$ and the initial pose T_0^k as input, our goal is to estimate the per-part pose and scale for each frame. Assuming the articulated object consists of K rigid parts and J joints, we use the superscript k to denote the k -th part and j for the j -th joint. The output of our method is the part-wise pose $T_t^k \in SE(3)$ and the corresponding scale $s_t^k \in \mathbb{R}^3$. The transformation matrix T_t^k is defined as:

$$T_t^k = \begin{bmatrix} R_t^k & \mathbf{t}_t^k \\ \mathbf{0}^T & 1 \end{bmatrix}, \quad R_t^k \in SO(3), \mathbf{t}_t^k \in \mathbb{R}^3 \quad (1)$$

where R_t^k represents the rotation and $\mathbf{t}_t^k$ denotes the translation of the k -th part at time t .

The tracking task is formulated as an inter-frame pose increment estimation problem. Specifically, given the pose in the previous frame, denoted as T_{t-1}^k , the pose in the current frame T_t^k can be expressed as:

$$T_t^k = \Delta T_t^k \cdot T_{t-1}^k \quad (2)$$

where $T \in SE(3)$. ΔT_t^k represents the pose increment from the previous frame to the current frame. Furthermore, we adopt Lie algebra to model the pose, which allows us to reformulate Eq. 2 as:

$$T_t^k = \exp(\xi_t^k) = \exp(\Delta \xi_t^k + \xi_{t-1}^k) \quad (3)$$

Here, $\xi \in \mathfrak{se}(3)$ represents the pose in the Lie algebra. The exponential map $\exp(\cdot)$ projects elements from the Lie algebra to the Lie group. As such, both ξ and T provide valid representations of a 6D pose.

METHOD

Quasi-Canonicalization

To leverage temporal priors across consecutive frames, we introduce a **Quasi-Canonicalization** strategy. Specifically,

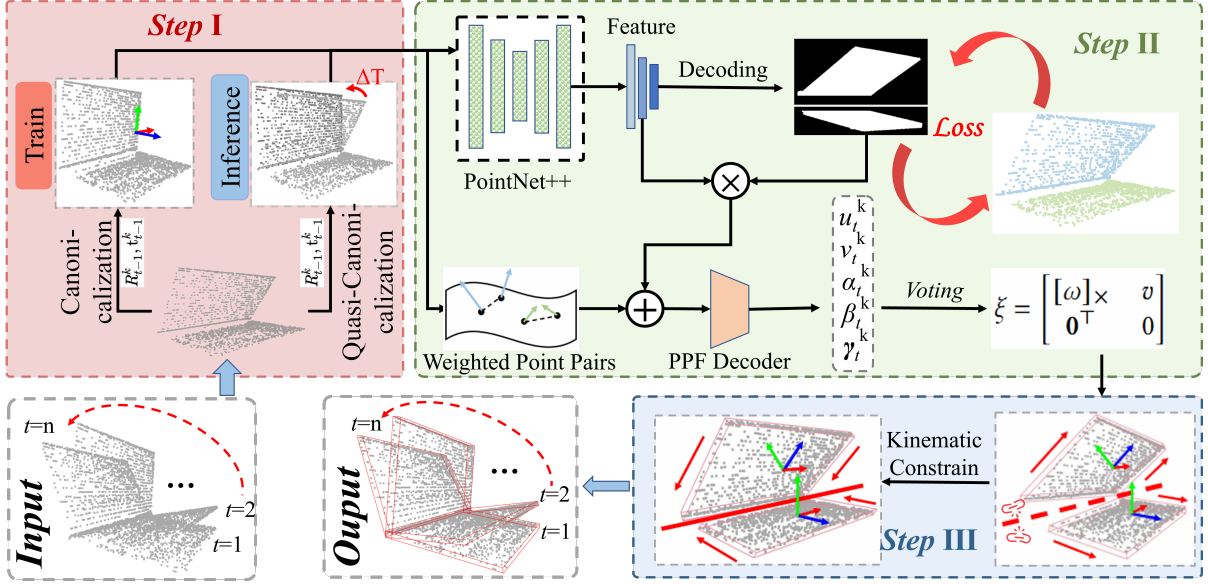


Figure 2: The Overview of Our PPF-Tracker.

the point cloud sequence is divided into temporal segments, each bounded by two consecutively selected keyframes, as illustrated in Fig. 3. For the i -th segment, we designate its first frame as the keyframe and define the associated transformation as $\mathcal{K}_i^k = (T_n^k)^{-1}$, where n denotes the keyframe’s original position in the sequence. This definition establishes a one-to-one mapping between the keyframe index i and its temporal index n .

We formally define the operation of applying the inverse transformation $(T_t^k)^{-1}$ to the t -th frame as *Canonicalization*. In parallel, the application of $\mathcal{K}_i^k$ to all frames within the i -th segment constitutes our *Quasi-Canonicalization* process. As shown in Fig. 4, the resulting point cloud in the quasi-canonical space, denoted as $\bar{\mathcal{P}}_t^k$, is formulated as:

$$\bar{\mathcal{P}}_t^k = \Delta T_t^k \mathcal{P}_c^k \quad (4)$$

where $\mathcal{P}_c^k$ is the point cloud in the canonical space, T_t^k is the transformation of the t -th frame, representing its absolute pose defined in the camera space, and $\mathcal{K}_i^k$ is the pose of the i -th keyframe.

The resulting quasi-canonical point cloud $\bar{\mathcal{P}}_t^k$ is then used as input for predicting the relative transformation ΔT_t^k . The corresponding ground truth is defined as:

$$\Delta T_t^{k(*)} = T_t^k (\mathcal{K}_i^k)^{-1} \quad (5)$$

where $(*)$ represents the ground truth. It is worth noting that our method does not directly regress ΔT_t^k . Instead, a more accurate prediction strategy is employed, as detailed in Sec. .

Recall Eq. 5, T_t^k is with the i -th keyframe as the beginning. We extend this formulation to the initial frame with absolute pose T_0^k . In pose tracking task, the first absolute pose is always T_0^k as shown in Fig. 3. So the pose T_t^k can be formulated as:

$$T_t^k = \Delta T_{[i]}^k \Delta T_{[i-1]}^k \cdots \Delta T_{[1]}^k T_0^k = \left(\prod_i \Delta T_{[i]}^k \right) T_0^k \quad (6)$$

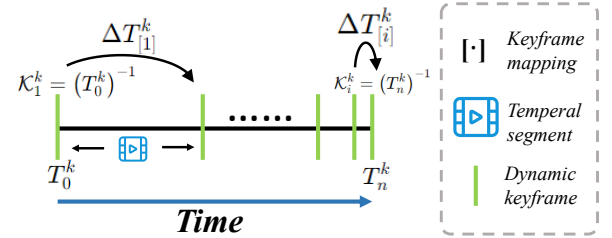


Figure 3: Illustration of Temporal Segment and Dynamic Keyframe. The symbols $\mathcal{K}_i^k$ and T_n^k are the poses of the i -th keyframe and the n -th frame, respectively. The subscript symbol $[\cdot]$ denotes the mapping of the keyframe to the frame stream. For instance, $[1] = 0$ and $[i] = n$.

Despite the keyframe method being able to mitigate cumulative errors, we find that the fixed keyframe strategy lacks flexibility, which leads to inaccuracy, as seen in Section . In this work, we propose a **Dynamic Keyframe Selection (DKS)** strategy. Our DKS strategy intelligently incorporates an energy function, where the keyframe is updated whenever the energy value falls below a predefined threshold $\phi = 0.01$. Specifically, we compute both the chamfer distance D_C and the hausdorff distance D_H between the predicted point cloud $\hat{\mathcal{P}}$ and the actually observed $\mathcal{P}$. Then, the energy function can be formulated as:

$$\mathfrak{E}_t = \frac{1}{|\mathcal{P}|} (D_C + D_H) \quad (7)$$

where $|\mathcal{P}|$ is the the number of points in $\mathcal{P}$. If $\mathfrak{E}_t < \phi$, it indicates a high similarity between the predicted frame and the observed frame, demonstrating the frame’s reliability. Therefore, it is selected as the keyframe for the next tem-

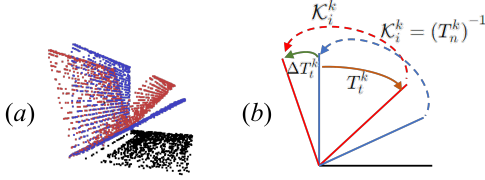


Figure 4: The Illustration of Quasi-Canonicalization within *Temporal Segment*. For clearer expression, we abstract (a) as (b). The **blue** part represents the keyframe transformation (canonicalization). The **red** part depicts the quasi-canonicalization of the t -th frame where the frame follows its associated keyframe’s transformation.

poral segment.

SE(3)-Invariance based Increment Learning

To ensure robust and accurate incremental estimation, we do not directly predict T_t^k . Instead, we adopt a PPF voting strategy to infer SE(3)-invariant parameters, from which a Lie algebra representation is obtained through voting, and the final transformation T_t^k is then derived via the exponential map. The overall process of our method consists of two key steps: **SE(3)-invariant parameters prediction** and **Lie algebra transformation**, detailed as follows:

(1) **SE(3)-invariant parameters prediction:** The PPF describes 3D shape characteristics by computing the relative geometric relationships between neighboring point pairs in a local coordinate system, thus inherently possessing SE(3) invariance. However, conventional point pair descriptors assign equal importance to all point pairs (Fig. 5 (a)) (You et al. 2022c). Intuitively, different point pairs play distinct roles in describing the three-dimensional features of objects (Fig. 5 (b)). Therefore, we use the **weighted PPF**.

Specifically, given the observed point cloud $\mathcal{P}_t$, we uniformly sample N point pairs $(\mathbf{p}_i, \mathbf{p}_j) \in (\mathcal{P}_t, \mathcal{P}_t)$, where all the sampled point $p \in \mathcal{P}_t$. For each pair, we compute an SE(3)-invariant feature vector $\mathcal{F}_{ij}$ using the positions $(\mathbf{p}_i, \mathbf{p}_j)$ and the surface normals $(\mathbf{n}_i, \mathbf{n}_j)$ of the point pairs. This feature ensures invariance to rigid transformations, enabling generalization across unseen object instances (Lin, Li, and Wang 2022). The weight v_{ij} for each point pair $(\mathbf{p}_i, \mathbf{p}_j)$ is determined based on the angle θ_{ij} between their surface normals:

$$v_{ij} = 1 - \lambda |\cos \theta_{ij}| \quad (8)$$

where $\lambda = 0.5$ is a tunable parameter. Eq. 8 assigns lower weights to point pairs with near-parallel normals (where $\theta_{ij} \approx 0^\circ$ or 180°) and higher weights to pairs with near-perpendicular normals (where $\theta_{ij} \approx 90^\circ$).

Then PointNet++ (Qi et al. 2017) is used to process $\mathcal{F}_{ij}$ to predict per part SE(3)-invariant parameters $(\mu_t^k, \nu_t^k, \alpha_t^k, \beta_t^k, \gamma_t^k)$ where (μ_t^k, ν_t^k) , (α_t^k, β_t^k) and γ_t^k related to the translation, rotation and scale parameters respectively. We denote the object center as $\mathbf{o}$, up orientation as $\vec{\mathbf{e}}_1$, right orientation as $\vec{\mathbf{e}}_2$ and unit vector $\vec{\mathbf{d}} = \frac{\mathbf{p}_i \mathbf{p}_j^\top}{\|\mathbf{p}_i \mathbf{p}_j^\top\|_2}$. These

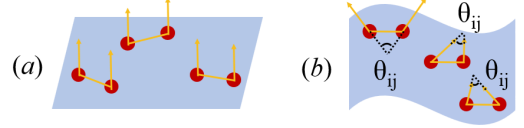


Figure 5: Traditional (a) and Weighted (b) Point Pairs.

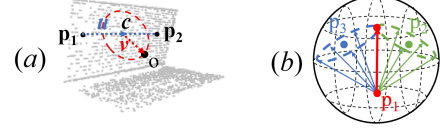


Figure 6: Illustration of Voting Scheme. (a) For orientation voting scheme, the center of each point pair is determined by a circle which is divided into bins. (b) For orientation voting scheme, a Fibonacci sphere with uniformly distributed bins is used for voting accumulation.

SE(3)-invariant parameters can be formulated as:

$$\begin{cases} \mu_t^k = \vec{\mathbf{p}}_i \vec{\mathbf{o}} \cdot \vec{\mathbf{d}}, \\ \nu_t^k = \|\vec{\mathbf{p}}_i \vec{\mathbf{o}} - \mu_t^k \vec{\mathbf{d}}\|_2, \\ \alpha_t^k = \vec{\mathbf{e}}_1 \cdot \vec{\mathbf{d}}, \\ \beta_t^k = \vec{\mathbf{e}}_2 \cdot \vec{\mathbf{d}}. \end{cases} \quad (9)$$

As depicted in Fig. 6, the per-part object center is constrained to lie on the circle defined by the translation parameters (μ_t^k, ν_t^k) , where μ_t^k determines the center c of the circle, and ν_t^k defines its radius. The orientation of the per-part object is determined by conics, which are derived from votes generated by multiple point pairs.

The remaining pose parameter γ_t^k represents the scaling factor which can be predicted directly.

(2) **Lie algebra transformation:** To avoid numerical divergence or loss of orthogonality that may arise from directly voting for ΔT_t^k , we first vote for elements in the Lie algebra tangle space where $\Delta \xi \in \mathfrak{se}(3)$, and then the multiplication of transformation matrices can be converted into the addition of their corresponding Lie algebras.

With the SE(3)-invariant parameters, we can transform them into Lie algebra according to (Eade 2013). We define ξ as the Lie algebra element by constructing a 4×4 matrix:

$$\xi = \begin{bmatrix} [\omega]_\times & v \\ \mathbf{0}^\top & 0 \end{bmatrix} \in \mathfrak{se}(3), \quad (10)$$

where $\omega \in \mathbb{R}^3$ represents rotation parameter and $v \in \mathbb{R}^3$ represents translation parameter. The symbol $[\cdot]_\times$ means the skew-symmetric matrix.

Utilizing Lie Algebra increments for pose tracking inherently offers more stable performance (Li et al. 2021a). Recall Eq. 6, we can re-formulate it as:

$$\xi_t^k = \sum_i \Delta \xi_{[i]}^k + \xi_0^k \quad (11)$$

where ξ^k is the Lie Algebra of T^k , and $\Delta \xi^k$ is the Lie Algebra of ΔT^k . Through the exponential map $T_t^k = \exp(\xi_t^k)$,

this linearized variation can be remapped to the SE(3) manifolds. This mapping inherently satisfies the orthogonality constraint of the rotation matrix, effectively preventing the output of non-orthogonal matrices. In our model, the angular velocity vector $w \in \mathbb{R}^3$ is free from the gimbal lock problem and maintains a one-to-one correspondence with the rotation matrix within the range $\|w\| < \pi$, thereby ensuring accuracy and consistency in pose estimation.

Optimization via Kinematic Constraints

Since our SE(3)-invariant framework models each part as an independent rigid body, this assumption may result in discontinuities across kinematic chains, leading to physically inconsistent motion. To mitigate this issue, we introduce a **Kinematic-Constrained Optimization** strategy that enforces rigid coupling between connected parts along their articulated axes.

Specifically, our method begins with coarse per-part pose predictions obtained through PPF voting, which provides a robust initialization for subsequent refinement. We then propose a comprehensive energy function that simultaneously minimizes geometric alignment errors while enforcing kinematic joint constraints. For per-part of the observed point cloud $\mathcal{P}_t^k$ and the canonical point cloud $\mathcal{P}_c^k$, the geometric alignment term $\mathfrak{E}_{geo}$ is defined as:

$$\mathfrak{E}_{geo} = \sum_{k=1}^K \left\| \frac{1}{|N^k|} (T^k)^{-1} \mathcal{P}^k - \mathcal{P}_c^k \right\|^2, \quad (12)$$

For the kinematic term $\mathfrak{E}_{kin}$, we take the j -th axis point q^j as an example, homogeneous normalization is applied to the axis point $(q^j)' = [q^j, 1]$. The kinematic term can be formulated as:

$$\mathfrak{E}_{kin} = \sum_{j=1}^{J-1} \|T^j(q_j')^T - T^{j+1}(q_j')^T\|^2 \quad (13)$$

where T^j and T^{j+1} represent the transformations of two parts connected by the same axial joint. Finally, our comprehensive energy function is given by $\mathfrak{E}_{comp} = \mathfrak{E}_{geo} + \mathfrak{E}_{kin}$.

With the comprehensive energy function, we can optimize $\hat{T}_t^k$ as $(\hat{T}_t^k)_{optim}$. Our optimization reduces system degrees of freedom and eliminates pose ambiguity. By incorporating kinematic constraints that reflect the geometric configuration of articulated structures, we restrict the motion space, thereby enhancing physical plausibility and ensuring more consistent pose tracking.

★ The overall articulation tracking procedure for the frame stream is summarized in Algorithm 1.

Loss Functions

We adopt the KL divergence as the loss function to quantify the discrepancy between probability distributions, with particular emphasis on deviations in low-probability regions. By applying the KL divergence to point-wise feature voting outcomes, we promote both efficient network convergence and computational efficiency. For scale prediction, we employ the MSE loss to quantify the difference between the

Algorithm 1: PPF-Tracker: Category-level Articulated Object Pose Tracking on SE(3) Manifolds.

```

1: Input: The frame stream  $\{\mathcal{P}_t^k\}_{t \geq 0}$  and initial pose  $\xi_0^k$ 
2: Output: Per-part 6D pose  $(\hat{T}_t^k)_{optim}$  and scale  $\hat{s}_t^k$  for all the  $t > 0$  frames.
3: Initialize keyframe as  $\mathcal{P}_0^k$ .
4: for each rigid part  $\mathcal{P}_t^k \in \{\mathcal{P}_t^k\}_{t > 0}$  do
5:   Sample  $N$  weighted point pairs  $(\mathbf{p}_i, \mathbf{p}_j) \in (\mathcal{P}_t, \mathcal{P}_t)$ .
6:   Predict the SE(3)-invariant parameters.
7:   Vote for the Lie algebra element  $\Delta \hat{\xi}_t^k$ .
8:   Accumulate increments  $\hat{\xi}_t^k = \hat{\xi}_{t-1}^k + \Delta \hat{\xi}_t^k$ .
9:   Compute coarse per-part pose  $\hat{T}_t^k$  and scale  $\hat{s}_t^k$ .
10:  Kinematic constraints:  $\hat{T}_t^k \rightarrow (\hat{T}_t^k)_{optim}$ .
11:  if  $\mathfrak{E}_t < \phi$  then
12:    Update keyframe.
13:  end if
14: end for

```

predicted and ground-truth scales. For mask learning, we adopt the BCE loss to supervise the predicted binary masks against the ground truth.

The detailed computation of the loss functions is provided in the Supplementary Material. Here, the total loss is formulated as a weighted sum of the aforementioned components:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{trans} + \lambda_2 \mathcal{L}_{orient} + \lambda_3 \mathcal{L}_{scale} + \lambda_4 \mathcal{L}_{mask} \quad (14)$$

where $\lambda_1 = 0.3$, $\lambda_2 = 0.3$, $\lambda_3 = 0.2$, and $\lambda_4 = 0.2$ are hyperparameters that balance the contributions of each loss term. This comprehensive loss function ensures that the model effectively learns the geometric relationships and SE(3)-invariant features, leading to robust and accurate pose tracking results.

EXPERIMENT

Experimental Setup

Implementation. During data pre-processing, input point cloud is downsampled to 3,072 points, and objects in RGB-D images are cropped and projected into the point cloud to serve as network inputs. The initial learning rate is set to 0.001, decreasing by a factor of 0.1 every 10 epochs. The number of total training epochs is 200. All the experiments are implemented on an NVIDIA GeForce RTX 4090 GPU with 24GB of memory.

Datasets and metrics. Following the setup in (Weng et al. 2021), we construct a synthetic dataset for category-level articulated object pose tracking based on PartNet-Mobility (Xiang et al. 2020). As the original data generation scripts and configuration details were not released, direct reproduction was not feasible. To ensure scientific rigor and comparability, we rebuild the synthetic tracking dataset using a similar strategy. Concretely, we generate a synthetic tracking dataset named PM-Videos from PartNet-Mobility (Xiang et al. 2020), semi-synthetic tracking dataset named ReArt-Videos from ReArt-48 repository (Liu et al. 2022a), and Real-world Scenario tracking dataset named

Category	Method	Per-part 6D Pose			Inference Time (s) ↓
		Rotation Error (°) ↓	Translation Error (m) ↓	3D IOU (%) ↑	
Laptop	A-NCSH (Li et al. 2020)	8.5, 9.2	0.084, 0.103	40.8, 28.3	1.67
	CAPTRA (Weng et al. 2021)	5.9, 5.3	0.080, 0.063	70.1 , 43.5	0.10
	ContactArt (Zhu et al. 2024)	8.8, 9.2	0.101, 0.154	50.2, 36.5	0.71
	GAPS (Yu et al. 2025)	8.7, 8.9	0.092, 0.096	54.3, 39.3	0.34
	PPF-Tracker (Ours)	3.7, 4.5	0.043, 0.055	68.5, 49.6	0.07
Eyeglasses	A-NCSH (Li et al. 2020)	7.6, 24.8, 26.6	0.079, 0.324, 0.319	40.5, 29.3, 28.4	2.59
	CAPTRA (Weng et al. 2021)	4.5, 12.6, 13.1	0.054, 0.097, 0.084	53.1, 41.2, 39.8	0.14
	ContactArt (Zhu et al. 2024)	11.3, 16.5, 13.2	0.134, 0.158, 0.137	47.3, 47.4, 43.2	1.00
	GAPS (Yu et al. 2025)	8.5, 9.3, 9.6	0.105, 0.123, 0.118	48.2, 36.7, 35.6	0.84
	PPF-Tracker (Ours)	2.5, 3.6, 3.8	0.031, 0.038, 0.041	58.6, 55.7, 59.2	0.12
Dishwasher	A-NCSH (Li et al. 2020)	5.0, 5.7	0.074, 0.119	64.5, 43.8	1.70
	CAPTRA (Weng et al. 2021)	4.6, 5.4	0.055, 0.089	85.3, 61.2	0.11
	ContactArt (Zhu et al. 2024)	7.8, 6.9	0.106, 0.145	78.3, 65.2	0.67
	GAPS (Yu et al. 2025)	6.2, 7.0	0.126, 0.207	80.3, 54.3	0.36
	PPF-Tracker (Ours)	3.2, 3.4	0.038, 0.045	87.2, 76.1	0.06
Scissors	A-NCSH (Li et al. 2020)	5.0, 5.7	0.041, 0.057	32.3, 32.8	1.21
	CAPTRA (Weng et al. 2021)	4.1, 4.7	0.032, 0.039	43.2, 42.8	0.12
	ContactArt (Zhu et al. 2024)	6.8, 7.5	0.085, 0.067	39.6, 38.0	0.5
	GAPS (Yu et al. 2025)	6.1, 6.6	0.055, 0.069	41.3, 40.2	0.29
	PPF-Tracker (Ours)	3.4, 4.1	0.027, 0.033	45.2, 44.6	0.09
Drawer	A-NCSH (Li et al. 2020)	8.6, 9.8, 11.5, 8.5	0.088 , 0.255, 0.257, 0.175	66.5, 62.3, 58.6, 61.3	3.64
	CAPTRA (Weng et al. 2021)	4.8 , 6.5, 6.3, 6.0	0.112, 0.185, 0.177, 0.156	89.5, 80.1, 76.7, 78.2	0.25
	ContactArt (Zhu et al. 2024)	6.5, 7.8, 7.6, 8.1	0.132, 0.172, 0.188, 0.207	71.2, 70.6, 68.3, 69.4	1.00
	GAPS (Yu et al. 2025)	6.5, 6.5, 6.5, 6.5	0.168, 0.242, 0.243, 0.239	86.3, 76.5, 75.6, 77.6	0.62
	PPF-Tracker (Ours)	5.2, 5.2, 5.2, 5.2	0.095, 0.172, 0.162, 0.153	90.2, 85.1, 78.6, 80.9	0.16

Table 1: Comparison with SOTA Methods on PM-Videos Dataset.

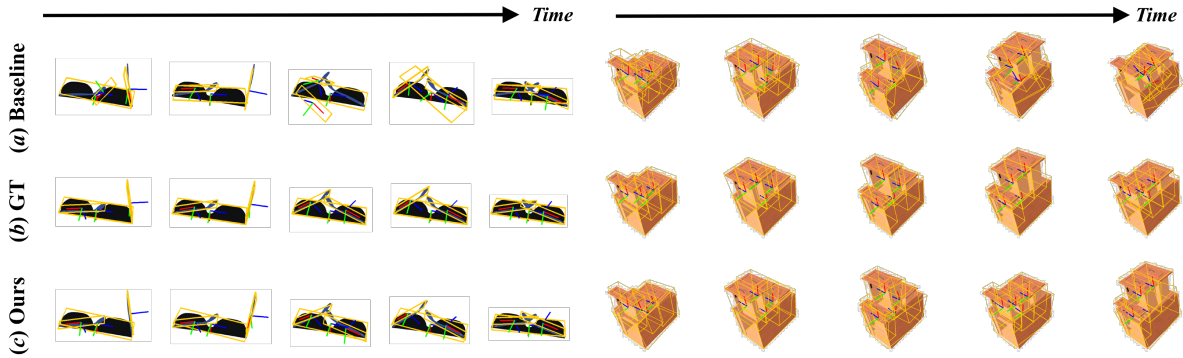


Figure 7: Qualitative Results on PM-Videos Dataset.

RobotArm-Videos from RobotArm dataset. To comprehensively evaluate the performance of PPF-Tracker, multiple metrics are adopted: degree error (°) for rotation, distance error (m) for translation, 3D IOU (%) for scale, and tracking speed (s) for real-time performance. We also measure cumulative tracking error across entire video to better quantify long-term tracking ability.

Comparison with the SOTA Methods

In this section, we evaluate the proposed PPF-Tracker on the synthetic articulated object dataset PM-Videos, which are generated from PartNet-Mobility (Xiang et al. 2020), with quantitative results summarized in Table 1. Compared to previous methods such as A-NCSH (Li et al. 2020) and ContactArt (Zhu et al. 2024), PPF-Tracker consistently achieves superior tracking performance across all object parts, as evidenced by significantly lower rotation and translation errors.

Specifically, for the *Eyeglasses* category, our method

achieves an average rotation error of **3.3°** and a translation error of **0.036m**, representing a relative reduction of approximately **60%** compared to the second-best method, GAPS. Additionally, the 3D IoU shows a notable improvement, with an average increase of **17.6%**. In terms of runtime performance, PPF-Tracker demonstrates strong real-time capabilities, with the inference time **0.12s** per frame.

Qualitative tracking results are illustrated in Fig. 7. We hypothesize that PPF-Tracker’s superior performance arises from the combination of voting-based strategies and Lie algebra-based transformation. Together, these provide a robust trade-off between tracking accuracy and efficiency.

Ablation Study

Kinematic Constraints. To evaluate the effectiveness of the proposed Kinematic Constraints, we conduct ablation experiments, with results summarized in Table 2. The application of kinematic constraints results in substantial error reduc-

Kinematic Constraints		Per-part 6D Pose	
		Rotation Error ($^{\circ}$)	Translation Error (m)
I	$\times$	8.6, 9.1	0.109, 0.122
II (Ours)	$\checkmark$	3.2, 3.4	0.038, 0.045

Keyframe Selection		Per-part 6D Pose	
		Rotation Error ($^{\circ}$)	Translation Error (m)
III	No Keyframe	13.6, 15.0	0.125, 0.153
IV	Fixed Keyframe	5.3, 6.2	0.061, 0.087
V (Ours)	Dynamic Keyframe	3.2, 3.4	0.038, 0.045

Table 2: Ablation Experiments with the PPF-Tracker.

tion—rotation errors decrease by 5.4° and 5.7° , while translation errors are reduced by $0.071m$ and $0.077m$, amounting to over **50%** improvement compared to the unconstrained setting. *These results underscore the necessity of incorporating kinematic constraints when performing part-level pose tracking, as they effectively enforce structural consistency across articulated components.*

Keyframe Selection. To validate the effectiveness of the dynamic keyframe strategy, we designed the following ablation experiment to evaluate three different settings: 1) No Keyframe, 2) Fixed Keyframe, and 3) Dynamic Keyframe. The experimental results are shown in Table 2, and all experiments were conducted on the same dataset. The errors are largest when no keyframes are used. *This indicates that without keyframes, accumulated errors quickly grow, leading to a significant decline in pose tracking accuracy. The fixed keyframe strategy somewhat alleviates the issue of accumulated errors but still underperforms compared to the dynamic keyframe approach.*

Generalization Capacity

Experiments on Semi-Synthetic Scenarios. We evaluate the effect of our PPF-Tracker on the semi-synthetic dataset ReArt-Videos. Results are shown in Fig. 8 (Top) and Table 3. The tracking results show that our method can perform well in the semi-synthetic scenarios.

Experiments on Real-world Scenarios. To investigate the tracking performance in real-world scenarios, we also evaluate the proposed PPF-Tracker on the 7-part RobotArm-Videos. Fig. 8 (Bottom) and Table 4 show the qualitative and quantitative results, respectively. It is evident that our approach achieves acceptable 6D pose tracking performance in real-world scenarios.

Category	Per-part 6D Pose	
	Rotation Error ($^{\circ}$)	Translation Error (m)
Box	4.8, 4.9	0.008, 0.008
Stapler	5.4, 5.8	0.010, 0.009
Cutter	3.3, 3.3	0.008, 0.010
Scissors	7.9, 8.7	0.007, 0.009
Drawer	7.9, 7.9	0.025, 0.021

Table 3: Results on ReArt-Video Dataset.

Per-part Rotation Error ($^{\circ}$)						
0.08	0.75	2.26	9.73	13.55	17.80	19.43
Per-part Translation Error (m)						
0.002	0.016	0.022	0.068	0.070	0.127	0.155

Table 4: Results on RobotArm-Videos Dataset.

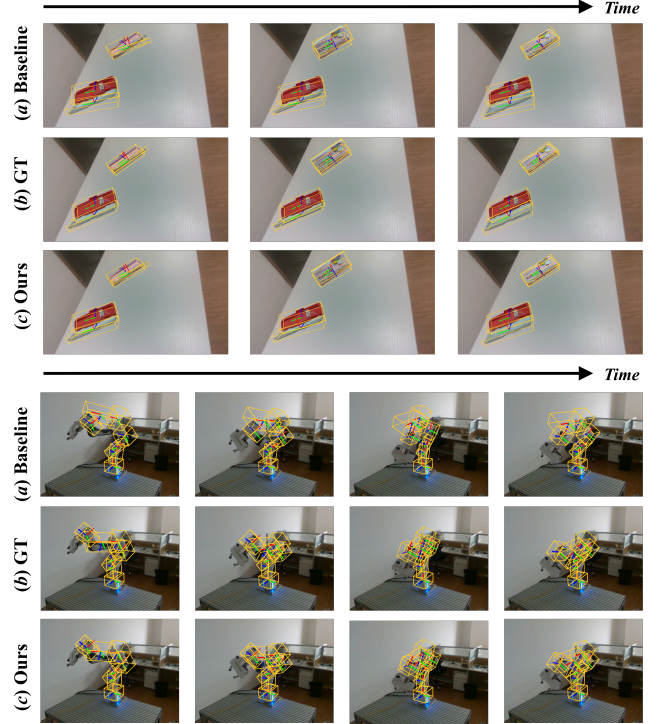


Figure 8: Qualitative Results on ReArt-Videos (Top) and RobotArm-Videos (Bottom).

CONCLUSION

This paper introduces PPF-Tracker, a framework for category-level articulated object pose tracking on the SE(3) manifold. Through a Quasi-Canonicalization strategy, the tracking task is reformulated as pose increment learning in Lie algebra space, ensuring stable and SE(3)-invariant motion modeling. To enhance accuracy and reduce drift, we adopt a dynamic keyframe selection mechanism based on geometric similarity and enforce structural consistency via kinematic constraints. Comprehensive evaluations on synthetic, semi-synthetic, and real-world datasets show that PPF-Tracker delivers state-of-the-art performance in accuracy, robustness, and generalization. We believe this work provides a principled and extensible foundation for articulated object tracking, with promising implications for robotics, embodied AI, and AR/VR applications.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Contract 62471450.

References

- Chen, Z.; Jing, L.; Li, Y.; and Li, B. 2024. Bridging the domain gap: Self-supervised 3d scene understanding with foundation models. *Advances in Neural Information Processing Systems*, 36.
- Clark, M.; Newman, M. W.; and Dutta, P. 2022. ARtulate: One-Shot Interactions with Intelligent Assistants in Unfamiliar Smart Spaces Using Augmented Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(1): 1–24.
- Davis, L.; and Aslam, U. 2024. Analyzing consumer expectations and experiences of Augmented Reality (AR) apps in the fashion retail sector. *Journal of Retailing and Consumer Services*, 76: 103577.
- Deng, X.; Xiang, Y.; Mousavian, A.; Eppner, C.; Bretl, T.; and Fox, D. 2020. Self-supervised 6d object pose estimation for robot manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 3665–3671. IEEE.
- Di, Y.; Zhang, R.; Lou, Z.; Manhardt, F.; Ji, X.; Navab, N.; and Tombari, F. 2022. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6781–6791.
- Eade, E. 2013. Lie groups for 2d and 3d transformations. URL <http://ethaneade.com/lie.pdf>, revised Dec, 117: 118.
- Fernandez-Labrador, C.; Chhatkuli, A.; Paudel, D. P.; Guerrero, J. J.; Demonceaux, C.; and Gool, L. V. 2020. Unsupervised learning of category-specific symmetric 3d keypoints from point sets. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16*, 546–563. Springer.
- Fu, Y.; and Wang, X. 2022. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *Advances in Neural Information Processing Systems*, 35: 27469–27483.
- Heppert, N.; Migimatsu, T.; Yi, B.; Chen, C.; and Bohg, J. 2022. Category-independent articulated object tracking with factor graphs. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3800–3807. IEEE.
- Hurlburt, R. T.; and Schwitzgebel, E. 2011. Presuppositions and background assumptions.
- Irshad, M. Z.; Zakharov, S.; Ambrus, R.; Kollar, T.; Kira, Z.; and Gaidon, A. 2022. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In *European Conference on Computer Vision*, 275–292. Springer.
- Jain, A.; Lioutikov, R.; Chuck, C.; and Niekum, S. 2021. Screwnet: Category-independent articulation model estimation from depth images using screw theory. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 13670–13677. IEEE.
- Jau, Y.-Y.; Zhu, R.; Su, H.; and Chandraker, M. 2020. Deep keypoint-based camera pose estimation with geometric constraints. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4950–4957. IEEE.
- Li, X.; Wang, H.; Yi, L.; Guibas, L. J.; Abbott, A. L.; and Song, S. 2020. Category-level articulated object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3706–3715.
- Li, X.; Weng, Y.; Yi, L.; Guibas, L. J.; Abbott, A.; Song, S.; and Wang, H. 2021a. Leveraging SE (3) equivariance for self-supervised category-level object pose estimation from point clouds. *Advances in neural information processing systems*, 34: 15370–15381.
- Li, Y.; Zhang, S.; Wang, Z.; Yang, S.; Yang, W.; Xia, S.-T.; and Zhou, E. 2021b. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International conference on computer vision*, 11313–11322.
- Lin, J.; Wei, Z.; Li, Z.; Xu, S.; Jia, K.; and Li, Y. 2021. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3560–3569.
- Lin, S.; Li, W.; and Wang, Y. 2022. Pose estimation of 3D objects based on point pair feature and weighted voting. In *International Conference on Intelligent Robotics and Applications*, 383–394. Springer.
- Lin, Y.; Tremblay, J.; Tyree, S.; Vela, P. A.; and Birchfield, S. 2022. Keypoint-based category-level object pose tracking from an RGB sequence with uncertainty estimation. In *2022 International Conference on Robotics and Automation (ICRA)*, 1258–1264. IEEE.
- Liu, L.; Xu, W.; Fu, H.; Qian, S.; Yu, Q.; Han, Y.; and Lu, C. 2022a. AKB-48: a real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14809–14818.
- Liu, L.; Xue, H.; Xu, W.; Fu, H.; and Lu, C. 2022b. Toward real-world category-level articulation pose estimation. *IEEE Transactions on Image Processing*, 31: 1072–1083.
- Liu, Z.; Wang, Q.; Liu, D.; and Tan, J. 2024. PA-Pose: Partial point cloud fusion based on reliable alignment for 6D pose tracking. *Pattern Recognition*, 148: 110151.
- Maji, D.; Nagori, S.; Mathew, M.; and Poddar, D. 2022. Yolo-pose: Enhancing yolo for multi person pose estimation using object keypoint similarity loss. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2637–2646.
- Manhardt, F.; Wang, G.; Busam, B.; Nickel, M.; Meier, S.; Minciullo, L.; Ji, X.; and Navab, N. 2020. CPS++: Improving class-level 6D pose and shape estimation from monocular images with self-supervised learning. *arXiv preprint arXiv:2003.05848*.
- Mao, W.; Tian, Z.; Wang, X.; and Shen, C. 2021. Fc-pose: Fully convolutional multi-person pose estimation with dynamic instance-aware convolutions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9034–9043.

- Mo, K.; Guibas, L. J.; Mukadam, M.; Gupta, A.; and Tulsiani, S. 2021. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6813–6823.
- Prakhya, S. M.; Bingbing, L.; Weisi, L.; and Qayyum, U. 2015. Sparse depth odometry: 3D keypoint based pose estimation from dense depth data. In *2015 IEEE international conference on robotics and automation (ICRA)*, 4216–4223. IEEE.
- Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.
- Sankaranarayanan, S.; Balaji, Y.; Jain, A.; Lim, S. N.; and Chellappa, R. 2018. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3752–3761.
- Wang, H.; Davoine, F.; Lepetit, V.; Chaillou, C.; and Pan, C. 2012. 3-D head tracking via invariant keypoint learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 22(8): 1113–1126.
- Wang, H.; Sridhar, S.; Huang, J.; Valentin, J.; Song, S.; and Guibas, L. J. 2019. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2642–2651.
- Wang, J.; Liu, W.; Yu, Q.; You, Y.; Liu, L.; Wang, W.; and Lu, C. 2024. RPMArt: Towards Robust Perception and Manipulation for Articulated Objects. *arXiv preprint arXiv:2403.16023*.
- Wang, W.; Yu, R.; Huang, Q.; and Neumann, U. 2018. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2569–2578.
- Wen, B.; and Bekris, K. 2021. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 8067–8074. IEEE.
- Wen, B.; Tremblay, J.; Blukis, V.; Tyree, S.; Müller, T.; Evans, A.; Fox, D.; Kautz, J.; and Birchfield, S. 2023. BundleSDF: Neural 6-DoF Tracking and 3D Reconstruction of Unknown Objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 606–617.
- Weng, Y.; Wang, H.; Zhou, Q.; Qin, Y.; Duan, Y.; Fan, Q.; Chen, B.; Su, H.; and Guibas, L. J. 2021. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 13209–13218.
- Xiang, F.; Qin, Y.; Mo, K.; Xia, Y.; Zhu, H.; Liu, F.; Liu, M.; Jiang, H.; Yuan, Y.; Wang, H.; et al. 2020. Sapien: A simulated part-based interactive environment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11097–11107.
- Yang, L.; Li, K.; Zhan, X.; Wu, F.; Xu, A.; Liu, L.; and Lu, C. 2022. OakInk: A Large-Scale Knowledge Repository for Understanding Hand-Object Interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20953–20962.
- You, S.; Yao, H.; and Xu, C. 2020. Multi-target multi-camera tracking with optical-based pose association. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(8): 3105–3117.
- You, Y.; He, W.; Liu, M. X.; Wang, W.; and Lu, C. 2022a. Go Beyond Point Pairs: A General and Accurate Sim2Real Object Pose Voting Method with Efficient Online Synthetic Training. *CoRR*.
- You, Y.; Liu, W.; Ze, Y.; Li, Y.-L.; Wang, W.; and Lu, C. 2022b. UkpGAN: A general self-supervised keypoint detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 17042–17051.
- You, Y.; Shi, R.; Wang, W.; and Lu, C. 2022c. Cppf: Towards robust category-level 9d pose estimation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6866–6875.
- Yu, Q.; Hao, C.; Yuan, X.; Zhang, L.; Liu, L.; Huo, Y.; Agarwal, R.; and Lu, C. 2025. Generalizable Articulated Object Perception with Superpoints. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5. IEEE.
- Zhang, J.; Wu, M.; and Dong, H. 2024. Generative Category-level Object Pose Estimation via Diffusion Models. *Advances in Neural Information Processing Systems*, 36.
- Zhang, L.; Han, Z.; Zhong, Y.; Yu, Q.; Wu, X.; Wang, X.; and Wang, R. 2024. Vocaber: Voting-based pose tracking for category-level articulated object via inter-frame priors. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 8942–8951.
- Zhang, L.; Jiang, H.; Huo, Y.; Zhong, Y.; Wang, J.; Wang, X.; Wang, R.; and Liu, L. 2025a. R²-Art: Category-Level Articulation Pose Estimation from Single RGB Image via Cascade Render Strategy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 9985–9993.
- Zhang, L.; Meng, W.; Zhong, Y.; Kong, B.; Xu, M.; Du, J.; Wang, X.; Wang, R.; and Liu, L. 2025b. U-COPE: Taking a Further Step to Universal 9D Category-Level Object Pose Estimation. In *European Conference on Computer Vision*, 254–270. Springer.
- Zhang, S.; Zhao, W.; Guan, Z.; Peng, X.; and Peng, J. 2021. Keypoint-graph-driven learning framework for object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1065–1073.
- Zhao, C.; Cai, W.; Dong, C.; and Hu, C. 2024a. Wavelet-based fourier information interaction with frequency diffusion adjustment for underwater image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8281–8291.
- Zhao, C.; Cai, W.; Hu, C.; and Yuan, Z. 2024b. Cycle contrastive adversarial learning with structural consistency for unsupervised high-quality image deraining transformer. *Neural Networks*, 178: 106428.

Zhao, C.; Chen, Z.; Xu, Y.; Gu, E.; Li, J.; Yi, Z.; Wang, Q.; Yang, J.; and Tai, Y. 2025a. From Zero to Detail: Deconstructing Ultra-High-Definition Image Restoration from Progressive Spectral Perspective. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 17935–17946.

Zhao, C.; Ci, E.; Xu, Y.; Fan, T.; Guan, S.; Ge, Y.; Yang, J.; and Tai, Y. 2025b. UltraHR-100K: Enhancing UHR Image Synthesis with A Large-Scale High-Quality Dataset. *arXiv preprint arXiv:2510.20661*.

Zhu, Z.; Wang, J.; Qin, Y.; Sun, D.; Jampani, V.; and Wang, X. 2024. ContactArt: Learning 3d interaction priors for category-level articulated object and hand poses estimation. In *2024 International Conference on 3D Vision (3DV)*, 201–212. IEEE.

Zou, L.; Huang, Z.; Gu, N.; and Wang, G. 2024. Learning geometric consistency and discrepancy for category-level 6D object pose estimation from point clouds. *Pattern Recognition*, 145: 109896.