

Category-Level Articulated Object Pose Estimation via Pose-Shape Hypothesis Generation and Verification

Kaifeng Tang^{1,2,3}, Chi Xu^{1,2,3*}, Xin Ao^{1,2,3}, Yuting Ge^{1,2,3}, Tingrui Guo^{1,2,3}, and Jun Zhou^{1,2,3}

¹ School of Artificial Intelligence and Automation, China University of Geosciences, Wuhan 430074, China

xuchi@cug.edu.cn

² Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China

³ Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, China

Abstract. Category-level 6D pose estimation of articulated objects from a single-view input is a fundamental yet challenging task. Its core difficulty lies in occlusion-induced pose and shape ambiguities, which are further exacerbated by highly entangled pose-shape representations and real-world depth noise. Existing generative methods often neglect the inherent correlations between pose and shape, resulting in geometrically inconsistent or physically implausible predictions. To address this limitation, we propose a novel generation-and-verification framework for this task: First, a Pose-Shape Hypothesis Generator jointly models the distributions of pose and shape to generate intrinsically coupled hypothesis pairs, providing explicitly paired candidates for subsequent verification. Second, a Part-Level Point Cloud Denoising module recovers clean point clouds from noisy and partial depth observations, ensuring robust inputs for the verification stage. Finally, a Cross-Space Consistency Verification module explicitly verifies the geometric consistency between the observed point clouds and canonical shapes using a continuous SDF-based scoring mechanism, which effectively prunes geometrically implausible hypotheses. Extensive experiments on challenging benchmarks demonstrate that our method significantly improves the robustness and accuracy of pose estimation in complex, heavily occluded scenarios. Project page: <https://github.com/KaifengT/ArtPS>.

Keywords: Pose estimation · Articulated object · Shape estimation

1 Introduction

Many common articulated objects in daily life, such as laptops, cabinet doors, and trash bins, consist of multiple rigid parts connected by joints. For applications like embodied AI [3, 5, 23, 39, 41, 49, 52], robust pose estimation of these

* Corresponding author.

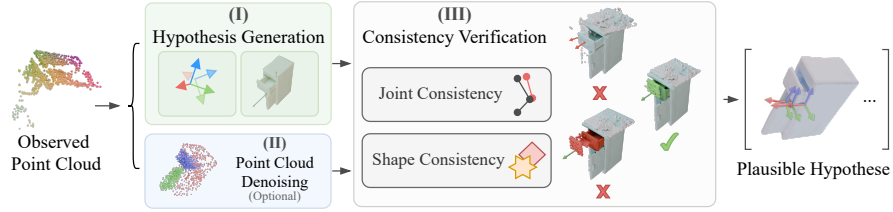


Fig. 1: Overview of the proposed framework. Given a single-view point cloud observation, our framework outputs a set of plausible pose-shape hypothesis pairs.

parts is crucial for safe robot interaction. Under single-view observations, partial visibility caused by self-occlusion or human-robot interactions leads to a loss of pose-related clues [8, 38, 53], resulting in pose ambiguity. Consequently, an observed point cloud often corresponds to multiple plausible poses, making it difficult to obtain a single solution.

Most existing category-level articulated object pose estimation methods are regression-based. They typically regress either dense [22, 54] or sparse [11] representations in a canonical space and solve for the pose via optimization-based correspondence matching, or explicitly regress the pose parameters [48]. Although they achieve high accuracy when parts are largely visible, they are highly sensitive to severe occlusion. Most importantly, regression-based methods fail to model object pose or shape uncertainties.

Recently, generative models have been adopted to produce hypotheses to address pose ambiguity. However, they are typically designed for specific targets, such as rigid objects [16, 50, 53], human hands [4, 8, 43, 44], human bodies [20], or specific instances with known kinematic parameters [35]. Consequently, these methods are not applicable to general category-level articulated objects, which exhibit highly varied kinematic parameters. Jain *et al.* [18] attempt to model the distribution of kinematic parameters from a depth video and use high-probability regions as outputs. However, this approach lacks further verification of the generated hypotheses, making it highly sensitive to implausible ones.

To identify plausible hypotheses, Huo *et al.* [15] and Xu *et al.* [45] build energy-based networks to score the likelihood of each hypothesis. However, part poses and shapes of articulated objects are highly entangled [27]. Existing methods ignore the impact of shape uncertainty on the pose distribution during scoring. Consequently, some high-scoring poses may correspond to geometrically implausible shapes. This limitation degrades model robustness, and real-world depth noise further amplifies this challenge.

To address these issues, we first propose a Pose-Shape Hypothesis Generator to model the joint distribution of pose and shape, effectively tackling their highly entangled uncertainties caused by occlusion. Unlike approaches restricted to modeling pose distributions, our network generates a joint latent code from a shared condition, yielding intrinsically coupled pose-shape hypothesis pairs. Specifically, it decodes part of this code into the observation space for per-part

poses, and maps the remainder into the canonical space to represent the object shape as a Signed Distance Function (SDF).

Second, real-world depth observations inevitably suffer from holes, noise, and distortions caused by low object reflectance. We introduce a Part-Level Point Cloud Denoising module to replace the noisy, partial input with a clean point cloud, thereby providing robust observational conditions for subsequent hypothesis verification.

Finally, we propose a Cross-Space Consistency Verification module. Ideally, when transformed using the ground-truth pose, an observed point cloud should perfectly align with the ground-truth shape in the canonical space. Motivated by this, we employ the SDF as a continuous scoring mechanism to verify hypothesis pairs. Specifically, we transform the observed point cloud into the canonical space using the hypothesized pose and compute its distance to the zero-level set of the hypothesized shape to derive a consistency score. This module effectively captures shape deviations, pruning geometrically implausible hypotheses and outputting explicitly consistent pose-shape predictions to address occlusion-induced ambiguity. In summary, our main contributions are:

- We propose a novel Pose-Shape Hypothesis Generator that models their joint distribution. It tackles highly entangled uncertainties by generating intrinsically coupled pose-shape hypothesis pairs as candidates for subsequent verification.
- We introduce a Part-Level Point Cloud Denoising module to recover clean point clouds from noisy, real-world depth observations, ensuring robust inputs for verification.
- We design a Cross-Space Consistency Verification module utilizing an SDF-based scoring mechanism. It explicitly evaluates geometric consistency to prune implausible hypotheses, significantly boosting global robustness.

Notably, while previous works [2, 27] have utilized cross-space shape consistency for self-supervised representation learning, we formulate it as an explicit verification criterion. By verifying hypothesis pairs, we effectively mitigate the negative impact of shape uncertainty on pose estimation, thereby addressing occlusion-induced pose ambiguity. Furthermore, instead of relying on the widely used Chamfer Distance (CD) [2, 11, 27], we adopt an SDF-based continuous scoring mechanism. This overcomes the sensitivity of CD to non-uniform point densities, resulting in a smoother and more robust consistency evaluation. The overview of the proposed framework is illustrated in Fig. 1.

2 Related Work

2.1 Articulated Object Pose Estimation

Regression-based Methods. Early methods tackled articulated object pose estimation at the instance-level, requiring known CAD models or precise kinematic parameters [6, 30]. However, such information is difficult to obtain in practice, limiting their applicability. To address this, Li *et al.* [22] first extended the

task to the category-level. These methods typically regress dense coordinates of the observed points in the part-level canonical space and solve for the pose using 3D-3D correspondences [13, 14, 26, 40, 54]. Subsequent works [11] simplify these dense coordinates into ordered keypoints or regress the pose directly [7, 48]. While these methods achieve accurate pose estimation when the object is largely visible, they face challenges under occlusion. Specifically, when occlusion leads to pose ambiguity, these regression-based methods fail to capture pose and shape uncertainties.

Generation-based Methods. Recently, generative models have been adopted to produce hypotheses to address pose ambiguity for specific targets. Human hands and bodies are specific articulated objects with fixed topologies, and their poses or shapes can be parameterized by skinning models [32] or Lie groups [44]. Works [4, 8, 20] attempt to model the parameter distributions of these models, solving occlusion-induced ambiguities for hands or bodies. Furthermore, Zhou *et al.* [53] extend this approach to hand-object interaction scenarios, and Tian *et al.* [35] estimate poses for robotic arms using known CAD models.

Research on general or category-level articulated object pose estimation remains limited. Specifically, Jain *et al.* [18] model pose distributions for general articulated objects using depth videos, while other works [15, 45] estimate category-level pose distributions from single-frame point clouds. However, category-level articulated objects naturally exhibit variations in shape and kinematic parameters, leading to highly entangled poses and shapes [27]. Existing methods ignore the impact of shape uncertainty on the pose distribution during scoring. Consequently, some pose hypotheses may correspond to geometrically implausible shapes. This limitation significantly degrades model robustness. Unlike these methods, which only model the pose distribution, we propose modeling the joint distribution of pose and shape. This ensures they are intrinsically coupled during the generation phase, providing paired hypotheses as candidates for subsequent consistency verification.

2.2 Shape Consistency for Pose Learning

Cross-space shape consistency is widely used as a key cue to evaluate pose accuracy. In terms of representation, some methods render 3D shapes (*e.g.*, point clouds [51], SDFs [1], NeRFs [46], or other implicit representations [17]) into the observation space, using 2D pixel differences to supervise or optimize the pose. However, 2D comparisons inevitably suffer from depth ambiguity. To address this, subsequent studies shift to 3D space. For instance, some works [11, 25] regress ordered keypoints of articulated objects in both the observation and canonical spaces, using the distances between corresponding keypoints to optimize the pose. However, sparse keypoints struggle to capture fine-grained shape details. Alternatively, some works [2, 24, 27] represent shapes as dense point clouds. Since these point clouds lack point-to-point correspondences with the partial observations, Chamfer Distance (CD) provides an effective metric for measuring shape discrepancies.

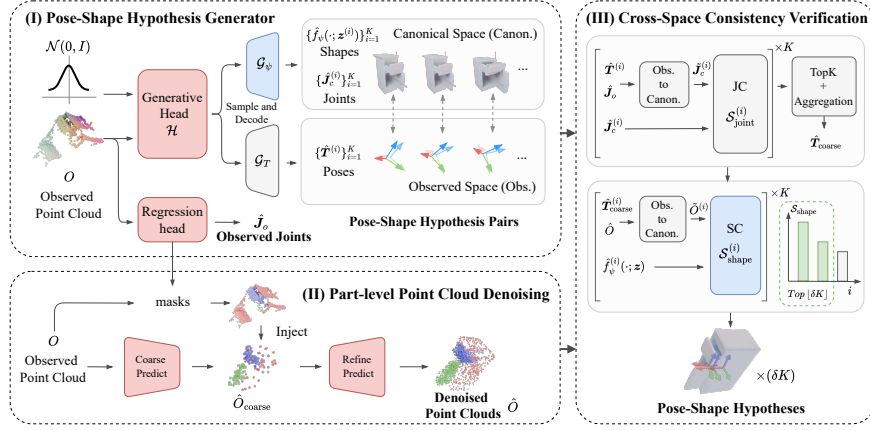


Fig. 2: Given an observation point cloud, the Pose-Shape Hypothesis Generator (I) first samples K pose-shape hypotheses conditioned on the input. The observation is then denoised and completed (II). Finally, the Cross-Space Consistency Verification module (III) evaluates joint and shape consistency to prune implausible hypotheses, yielding the final valid pose-shape pairs.

In practice, most of the aforementioned methods treat consistency merely as a supervision constraint during network training. In contrast, inspired by the analysis-by-synthesis paradigm, our method utilizes cross-space shape consistency as an explicit verification criterion during the inference stage. By verifying pose-shape hypothesis pairs, we output intrinsically coupled and explicitly consistent hypotheses, thereby addressing occlusion-induced pose ambiguity. Furthermore, since CD is sensitive to non-uniform point densities, we adopt an SDF-based continuous scoring mechanism, resulting in a smoother consistency evaluation.

3 Method

3.1 Problem Formulation

Category-level articulated objects consist of P rigid parts connected by $P - 1$ joints $\mathbf{J} = \{\mathbf{J}^{(j)}\}_{j=1}^{P-1}$ in a tree topology. Objects within the same category share this kinematic topology. Joints are either revolute or prismatic. A revolute joint is parameterized as $\mathbf{J}_{\text{rev}} = \{\mathbf{u}, \mathbf{q}\}$, where $\mathbf{u} \in \mathbb{R}^3$ is a pivot point on the joint axis, and $\mathbf{q} \in \mathbb{S}^2$ is the axis direction. Since a prismatic joint translates along a specific direction independent of a pivot point, it is parameterized as $\mathbf{J}_{\text{pris}} = \{\mathbf{q}\}$.

Given a partial observation point cloud O of an articulated object from a known category with known kinematic topology, we aim to jointly model its pose and shape distributions. Specifically, our goal is to infer hypotheses for both the per-part poses $\mathbf{T} = \{\mathbf{T}^{(j)} \in \text{SE}(3)\}_{j=1}^P$ in the observation space and the

corresponding object shape in the canonical space. Our framework is illustrated in Fig. 2.

3.2 Pose-Shape Hypothesis Generator

Canonical Shape Prior Learning Directly predicting high-dimensional 3D representations under occlusion is computationally expensive and prone to topological errors. Inspired by [12, 21, 31], we construct a compact, continuous latent space to parameterize canonical shape variations and define part reference poses. This shape prior provides both a low-dimensional target for joint generation and a native SDF for subsequent hypothesis verification.

To learn the canonical shape prior, we employ a Variational Autoencoder (VAE). A PointNet++ [33] encoder extracts a latent code $\mathbf{z} \in \mathbb{R}^D$ from complete point clouds, while a Triplane-MLP decoder $\mathcal{G}_\psi$ maps $\mathbf{z}$ to an implicit neural SDF $f_\psi(\cdot; \mathbf{z}) : \mathbb{R}^3 \mapsto \mathbb{R}$ alongside canonical joint parameters $\mathbf{J}_c$. This VAE is pre-trained on canonical meshes with the following objective:

$$\mathcal{L}_{\text{prior}} = \mathcal{L}_{\text{sdf}} + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}} + \lambda_{\text{joint}} \mathcal{L}_{\text{joint}} + \lambda_{\text{KL}} \mathcal{L}_{\text{KL}} \quad (1)$$

Here, $\mathcal{L}_{\text{sdf}}$ is the L_1 reconstruction loss between the estimated and ground-truth SDF; $\mathcal{L}_{\text{eik}}$ enforces the Eikonal constraint [9, 37] ($\|\nabla f\| = 1$) to provide a smooth distance field for verification; $\mathcal{L}_{\text{joint}}$ supervises the joint parameters [11]; and $\mathcal{L}_{\text{KL}}$ regularizes the latent space toward $\mathcal{N}(0, \mathbf{I})$ to ensure continuous shape variations. The variables λ_{eik} , λ_{joint} , and λ_{KL} are hyperparameters.

Pose-Shape Distribution Learning Following VAE pre-training, the encoded latent codes serve as generation targets. To eliminate redundant degrees of freedom (DoF) imposed by joint constraints, a pose decoder $\mathbf{T} = \mathcal{G}_T(\mathbf{t}, \mathbf{J}_c)$ deterministically maps low-dimensional pose parameters $\mathbf{t}$ into complete per-part poses. Defining $\boldsymbol{\theta} = \{\mathbf{t}, \mathbf{z}\}$, we aim to model the conditional distribution $p(\boldsymbol{\theta} \mid O)$ given a partial point cloud O .

We use Mamba3D [10] to extract global features from O as conditioning. We then adopt a Glow-based normalizing flow [19], $\mathcal{H}$, to model the target joint distribution. Given noise $\mathbf{z}_0 \sim \mathcal{N}(0, \mathbf{I})$, Glow applies invertible transforms to obtain samples conditioned on these features. The distribution is:

$$\hat{p}(\boldsymbol{\theta} \mid O) = p(\mathbf{z}_0) \left| \det \frac{\partial \mathcal{H}(\mathbf{z}_0, O)}{\partial \mathbf{z}_0} \right|^{-1} \quad (2)$$

During training, given the ground truth $\boldsymbol{\theta}_{\text{GT}}$, we use the inverse transformation of the Glow model $\mathcal{H}^{-1} : \boldsymbol{\theta} \mapsto \mathbf{z}_0$ to maximize the posterior likelihood. Thus, the training objective is to minimize the negative log-likelihood (NLL) loss: $\mathcal{L}_{\text{nll}} = -\log \hat{p}(\boldsymbol{\theta}_{\text{GT}} \mid O)$.

During inference, we randomly draw K samples from the base distribution $\mathbf{z}_0$. We feed these samples, together with the conditional features, into the forward Glow model $\mathcal{H}$ to generate K joint hypotheses $\{\hat{\boldsymbol{\theta}}^{(i)}\}_{i=1}^K$. For each hypothesis $\hat{\boldsymbol{\theta}} =$

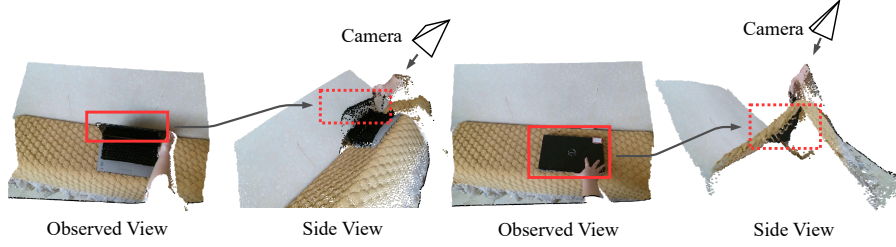


Fig. 3: Illustration of point cloud noise and depth distortions. In real-world scenarios, depth sensors often fail to capture accurate geometries. Here, we highlight two examples of severe depth distortion caused by surface reflectance.

$\{\hat{\mathbf{t}}, \hat{\mathbf{z}}\}$, we directly decode the per-part poses $\hat{\mathbf{T}}$, the implicit shape representation $\hat{f}_\psi(\cdot; \hat{\mathbf{z}})$, and the joint parameters $\hat{\mathbf{J}}_c$ using the pose decoder $\mathcal{G}_T$ and the pre-trained, frozen decoder $\mathcal{G}_\psi$. Please refer to the supplementary material for the detailed designs of the VAE and $\mathcal{G}_T$.

3.3 Part-Level Point Cloud Denoising

In real-world settings, depth sensors typically yield noisy point clouds. As illustrated in Fig. 3, factors such as surface reflectance and occlusions often induce significant depth distortions. Directly aligning these low-quality observations to the generated shape hypotheses often leads to local minima, which corrupts the verification criterion and breaks the subsequent hypothesis verification module. To better handle structurally complex articulated objects, we introduce a Part-Level Point Cloud Denoising module. Because both sensor noise and viewpoint-induced truncation are tightly coupled with the camera projection model, we perform restoration directly in the original observation space.

We formulate this module as a point cloud completion task. Built on a rigid object completion framework [36], it follows a coarse-to-fine, two-stage estimation pipeline. Given a partial observation point cloud O and point-level segmentation masks predicted from O , the network outputs denoised, geometrically complete, and part-wise dense point clouds $\hat{O} = \{\hat{O}^{(j)} \in \mathbb{R}^{N \times 3}\}_{j=1}^P$, where P is the number of rigid parts in the articulated object, N is the number of points.

From an object-level view, the network first predicts coarse yet complete part-wise point clouds $\hat{O}_{\text{coarse}} = \{\hat{O}_{\text{coarse}}^{(j)}\}_{j=1}^P$. Guided by the part assignments, it then injects the partial observation O into these predictions and upsamples them to yield denoised, complete point clouds at the part level. This design provides a geometric reference that balances global structures and local details for subsequent verification: the generated invisible regions recover the missing geometry to help avoid local minima, while the injected visible points preserve fine-grained details, thereby improving verification accuracy and robustness.

The learning objective is defined as:

$$\mathcal{L}_{dn} = \sum_{j=1}^P d_{\text{CD}}(\hat{O}^{(j)}, O_{\text{GT}}^{(j)}) + \sum_{j=1}^P d_{\text{CD}}(\hat{O}_{\text{coarse}}^{(j)}, O_{\text{CGT}}^{(j)}) \quad (3)$$

For each part, the ground truth is uniformly sampled from the complete mesh surface and transformed into the observation space. O_{GT} and O_{CGT} match the point densities of $\hat{O}$ and $\hat{O}_{\text{coarse}}$, respectively. We use the CD d_{CD} to enforce consistency between the predicted and ground-truth shapes.

3.4 Cross-Space Consistency Verification

Although the joint distribution of pose and shape inherently promotes coupling between pose and shape hypotheses, it does not enforce explicit geometric consistency. We therefore propose a Cross-Space Consistency Verification module that uses cross-space geometry to verify pose-shape hypothesis pairs, producing intrinsically coupled and explicitly consistent hypotheses and thereby addressing occlusion-induced ambiguity. The module consists of the following two stages:

Joint Consistency. In practical scenarios, the base part of an articulated object typically occupies a dominant spatial volume, making its observation relatively more robust to severe occlusions than the child parts. Consequently, estimating the connected joint parameters $\hat{\mathbf{J}}_o$ directly from the base part observation provides a more dependable geometric anchor. This approach is intrinsically more reliable than the ill-posed task of regressing absolute per-part poses simultaneously from a highly incomplete global point cloud. Leveraging this insight, we utilize $\hat{\mathbf{J}}_o$ as a foundational reference to initially verify and refine the generated pose-shape hypotheses.

Specifically, we sample K hypotheses $\{\hat{\mathbf{T}}^{(i)}, \hat{f}_\psi(\cdot; \mathbf{z}^{(i)}), \hat{\mathbf{J}}_c^{(i)}\}_{i=1}^K$ from the conditioned distribution $\hat{p}(\boldsymbol{\theta} \mid O)$. For each hypothesis, we use $\hat{\mathbf{T}}^{(i)}$ to transform $\hat{\mathbf{J}}_o$ from the observation space to the canonical space, yielding $\tilde{\mathbf{J}}_c^{(i)}$. We then compute the joint consistency score for the i -th hypothesis as:

$$\mathcal{S}_{\text{joint}}^{(i)} = - \sum_{j=1}^{P-1} \mathcal{D}(\hat{\mathbf{J}}_c^{(j,i)}, \tilde{\mathbf{J}}_c^{(j,i)}) \quad (4)$$

where the function $\mathcal{D}(\cdot, \cdot)$ is used to measure the difference between two sets of joint parameters. It is defined as:

$$\mathcal{D}(\mathbf{J}_1, \mathbf{J}_2) = \begin{cases} \mathcal{D}_a(\mathbf{q}_1, \mathbf{q}_2) + \lambda_d \mathcal{D}_d(\mathbf{u}_1, \mathbf{u}_2, \mathbf{q}_2), & \text{if Revolute Joint} \\ \mathcal{D}_a(\mathbf{q}_1, \mathbf{q}_2), & \text{if Prismatic Joint} \end{cases} \quad (5)$$

In this equation, the angular error $\mathcal{D}_a$ and distance error $\mathcal{D}_d$ are defined as:

$$\mathcal{D}_a(\mathbf{q}_1, \mathbf{q}_2) = \arccos\left(\frac{\mathbf{q}_1 \cdot \mathbf{q}_2}{\|\mathbf{q}_1\| \|\mathbf{q}_2\|}\right) \quad (6)$$

$$\mathcal{D}_d(\mathbf{u}_1, \mathbf{u}_2, \mathbf{q}_2) = \frac{\|(\mathbf{u}_1 - \mathbf{u}_2) \times \mathbf{q}_2\|}{\|\mathbf{q}_2\|} \quad (7)$$

We introduce a hyperparameter λ_d to balance the scale mismatch between the angular and distance terms. We rank the K hypotheses by $\mathcal{S}_{\text{joint}}^{(i)}$ and aggregate the base part poses of the top $\delta_1 K$ samples, where $\delta_1 \in (0, 1]$ is the filtering ratio. We then broadcast the aggregated pose to the remaining hypotheses to obtain $\hat{\mathbf{T}}_{\text{coarse}}^{(i)}$, and pass $\{\hat{\mathbf{T}}_{\text{coarse}}^{(i)}, \hat{f}_\psi(\cdot; \mathbf{z}^{(i)})\}_{i=1}^K$ to the next stage.

Shape Consistency. For the K hypotheses output by the first stage, we verify shape consistency in canonical space using the SDF. Specifically, we apply the poses $\hat{\mathbf{T}}_{\text{coarse}}^{(i)}$ to transform the denoised point clouds $\hat{O}$ into canonical space, yielding $\{\hat{O}^{(i)}\}_{i=1}^K$. Since the base part is anchored by the previous stage, we evaluate shape consistency only for the child parts ($j \geq 2$) by summing point-to-surface projection errors to the zero level set of the hypothesis SDF $\hat{f}_\psi(\cdot; \mathbf{z}^{(i)})$:

$$\mathcal{S}_{\text{shape}}^{(i)} = -\frac{1}{N} \sum_{j=2}^P \sum_{x \in \tilde{O}^{(j,i)}} \left| \hat{f}_\psi(x; \mathbf{z}^{(i)}) \right| \quad (8)$$

Finally, we rank the K hypotheses by $\mathcal{S}_{\text{shape}}^{(i)}$ in descending order and keep the top δK as the final set, where $\delta \in (0, 1]$ is the filtering ratio. If a single solution is required, we output the highest-scoring hypothesis.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate the proposed framework on two public datasets:

- **HOI4D** [28]: Features real-world egocentric hand-object interactions with severe occlusions and sensor noise. Following [54], we evaluate four categories (Laptop, StorageFurniture, Safe, TrashCan), totaling 172 instances across approx. 154k training frames and 19k testing frames.
- **PM-Videos** [29]: A synthetic PartNet-Mobility-based [42] dataset providing high-quality point clouds with less severe occlusion across five categories (Laptop, Eyeglasses, Dishwasher, Drawer, Scissors). It comprises 232 instances with approx. 38k training frames and 4k testing frames.

Evaluation Metrics. Following standard benchmarks, we evaluate pose estimation using mean average precision (mAP) at n° and m cm, as well as mean rotation error (R_{err}) in degrees and mean translation error (T_{err}) in meters. For canonical shape reconstruction, we report the L_2 Chamfer Distance (CD- L_2).

Table 1: Comparison of per-part pose estimation performance across different methods on HOI4D. The best results are highlighted in **bold**.

Category	Method	Metrics					
		5°5cm (%) $\uparrow$		R_{err} (°) $\downarrow$		T_{err} (m) $\downarrow$	
		Per-part	Mean	Per-part	Mean	Per-part	Mean
Laptop	ANCSH(2021)	41.9, 22.3	32.1	42.4, 45.4	43.9	0.464, 0.484	0.474
	ContactArt(2024)	- , -	62.5	- , -	5.20	- , -	0.072
	RPF(2025)	47.3, 14.7	31.0	7.22, 17.6	12.4	0.099, 0.228	0.163
	PPF-Tracker(2025)	38.5, 26.0	32.2	7.25, 12.8	10.0	0.035, 0.044	0.039
	Ours	85.1, 83.5	84.3	3.12, 3.08	3.10	0.018, 0.022	0.020
StorageFurniture	ANCSH	19.8, 37.9, 44.9	34.2	4.40, 8.78, 4.66	5.94	0.099, 0.103, 0.076	0.093
	ContactArt	- , - , -	24.0	- , - , -	5.75	- , - , -	0.060
	RPF	34.5, 10.4, 33.5	26.1	2.62, 20.1, 2.69	8.48	0.063, 0.234, 0.064	0.120
	PPF-Tracker	60.8, 42.3, 38.3	47.1	4.31, 5.80 , 8.02	6.05	0.059, 0.068, 0.063	0.063
	Ours	63.0, 42.5, 53.1	52.9	2.50, 6.67, 2.50	3.89	0.044, 0.049, 0.049	0.047
Safe	ANCSH	32.8, 29.7	31.2	3.46, 4.63	4.05	0.075, 0.065	0.070
	ContactArt	- , -	33.4	- , -	5.11	- , -	0.065
	RPF	64.8, 35.2	50.0	3.21, 10.9	7.09	0.045, 0.119	0.082
	PPF-Tracker	61.2, 51.1	56.1	4.88, 6.08	5.48	0.035, 0.035	0.035
	Ours	80.0, 63.3	71.7	3.11, 4.19	3.65	0.027, 0.030	0.029
TrashCan	ANCSH	33.3, 20.3	26.8	6.21, 10.2	8.22	0.050, 0.052	0.051
	ContactArt	- , -	24.2	- , -	5.98	- , -	0.072
	RPF	24.4, 11.6	18.0	6.37, 13.4	9.92	0.070, 0.168	0.119
	PPF-Tracker	46.0, 22.0	34.0	5.52, 8.35	6.94	0.036, 0.041	0.039
	Ours	50.7, 19.7	35.2	5.59, 9.19	7.39	0.025, 0.038	0.032
Overall	ANCSH		31.0		15.54		0.172
	ContactArt		36.0		5.51		0.067
	RPF		31.2		9.48		0.121
	PPF-Tracker		44.7		6.81		0.038
	Ours		61.0		4.51		0.032

Implementation Details. During inference, we apply the part-level point cloud denoising module (Sec. 3.3) for HOI4D but omit it for PM-Videos, directly using segmented observations for acceleration. We set $K = 1024$. Hyperparameters are $\lambda_{\text{eik}} = 0.1$, $\lambda_{\text{joint}} = 1.0$, $\lambda_{\text{KL}} = 1.0 \times 10^{-4}$, $\lambda_d = 0.1$, and $\delta_1 = 0.1$. Experiments are implemented in PyTorch on a single NVIDIA RTX 4090 GPU.

Baselines. We compare our method with representative state-of-the-art approaches, including single-view pose estimators (ANCSH [22], ContactArt [54], RPF [34], CARTO [12], and GAPS [47]) and sequence-based trackers (PPF-Tracker [29] and CAPTRA [40]).

4.2 Comparison to Baselines

Pose Estimation. Tab. 1 summarizes the quantitative results of all methods on the HOI4D dataset. Averaged across all categories, our framework achieves an mAP of 61.0% at 5° 5 cm, yielding a 16.3 percentage-point improvement over the state-of-the-art method, PPF-Tracker. Furthermore, compared to the best baselines in terms of R_{err} and T_{err} , our framework reduces these metrics by 1.0° and 0.6 cm, respectively. Across most categories, our framework outperforms state-of-the-art baselines on all metrics.

Table 2: Comparison of per-part pose estimation performance across different methods on PM-Videos. The best results are highlighted in **bold**, and the second-best results are underlined. Results of other methods are sourced from [29].

Category	Method	Metrics			
		R_{err} ($^{\circ}$) $\downarrow$		T_{err} (m) $\downarrow$	
		Per-part pose	Mean	Per-part pose	Mean
Laptop	ANCSH	8.5, 9.2	8.85	0.084, 0.103	0.093
	CAPTRA	5.9, 5.3	5.6	0.080, <u>0.063</u>	0.071
	ContactArt	8.8, 9.2	9.0	0.101, 0.154	0.127
	GAPS	8.7, 8.9	8.8	0.092, 0.096	0.094
	PPF-Tracker	<u>3.7</u> , <u>4.5</u>	<u>4.1</u>	0.043 , 0.055	0.049
	Ours	3.0 , 3.0	3.0	<u>0.053</u> , 0.065	<u>0.059</u>
Eyeglasses	ANCSH	7.6, 24.8, 26.6	19.66	0.079, 0.324, 0.319	0.240
	CAPTRA	4.5, 12.6, 13.1	10.06	0.054, 0.097, 0.084	0.078
	ContactArt	11.3, 16.5, 13.2	13.66	0.134, 0.158, 0.137	0.143
	GAPS	8.5, 9.3, 9.6	9.13	0.105, 0.123, 0.118	0.115
	PPF-Tracker	<u>2.5</u> , <u>3.6</u> , <u>3.8</u>	<u>3.3</u>	0.031 , <u>0.038</u> , 0.041	0.036
	Ours	1.9 , 2.5 , 2.4	2.3	<u>0.039</u> , 0.036 , <u>0.048</u>	<u>0.041</u>
Dishwasher	ANCSH	5.0, 5.7	5.35	0.074, 0.119	0.096
	CAPTRA	4.6, 5.4	5.0	<u>0.055</u> , <u>0.089</u>	<u>0.072</u>
	ContactArt	7.8, 6.9	7.35	0.106, 0.145	0.125
	GAPS	6.2, 7.0	6.60	0.126, 0.207	0.166
	PPF-Tracker	<u>3.2</u> , 3.4	<u>3.30</u>	0.038 , 0.045	0.041
	Ours	1.8 , <u>3.5</u>	2.6	0.065, 0.107	0.086
Scissors	ANCSH	5.0, 5.7	5.35	0.041, 0.057	0.049
	CAPTRA	4.1, 4.7	4.40	<u>0.032</u> , <u>0.039</u>	<u>0.035</u>
	ContactArt	6.8, 7.5	7.15	0.085, 0.067	0.076
	GAPS	6.1, 6.6	6.35	0.055, 0.069	0.062
	PPF-Tracker	<u>3.4</u> , <u>4.1</u>	<u>3.75</u>	0.027 , 0.033	0.030
	Ours	2.3 , 2.6	2.5	0.038, 0.041	0.040
Drawer	ANCSH	8.6, 9.8, 11.5, 8.5	9.6	<u>0.088</u> , 0.255, 0.257, 0.175	0.193
	CAPTRA	<u>4.8</u> , 6.5, 6.3, 6.0	5.9	0.112, 0.185, 0.177, 0.156	0.157
	ContactArt	6.5, 7.8, 7.6, 8.1	7.5	0.132, <u>0.172</u> , 0.188, 0.207	0.174
	GAPS	6.5, 6.5, 6.5, 6.5	6.5	0.168, 0.242, 0.243, 0.239	0.223
	PPF-Tracker	5.2, <u>5.2</u> , <u>5.2</u> , <u>5.2</u>	<u>5.2</u>	0.095, 0.172, <u>0.162</u> , <u>0.153</u>	<u>0.145</u>
	Ours	2.3 , 2.3 , 2.3 , 2.3	2.3	0.043 , 0.095 , 0.104 , 0.081	0.081

Tab. 2 summarizes the results on the synthetic PM-Videos dataset. Since this dataset exhibits fewer occlusions and lacks real-world noise, the task is relatively less challenging. Nevertheless, our framework achieves top performance across most metrics, outperforming PPF-Tracker, particularly in R_{err} . While our framework slightly trails PPF-Tracker in T_{err} , the latter is a sequence-based tracking method that inherently leverages inter-frame smoothness constraints. As a single-view estimation framework, our framework is able to attain comparable accuracy, which clearly underscores its robust competitiveness.

Fig. 4 shows qualitative comparisons of per-part pose estimation between our method and baselines on heavily occluded test cases from the HOI4D dataset. For each case, the first row presents the observed view (Obs.), while the second row displays the side view (Side). As illustrated, severe occlusion and high noise significantly reduce part visibility. Nevertheless, our pose estimates remain closely aligned with the ground truth across both viewpoints, whereas other

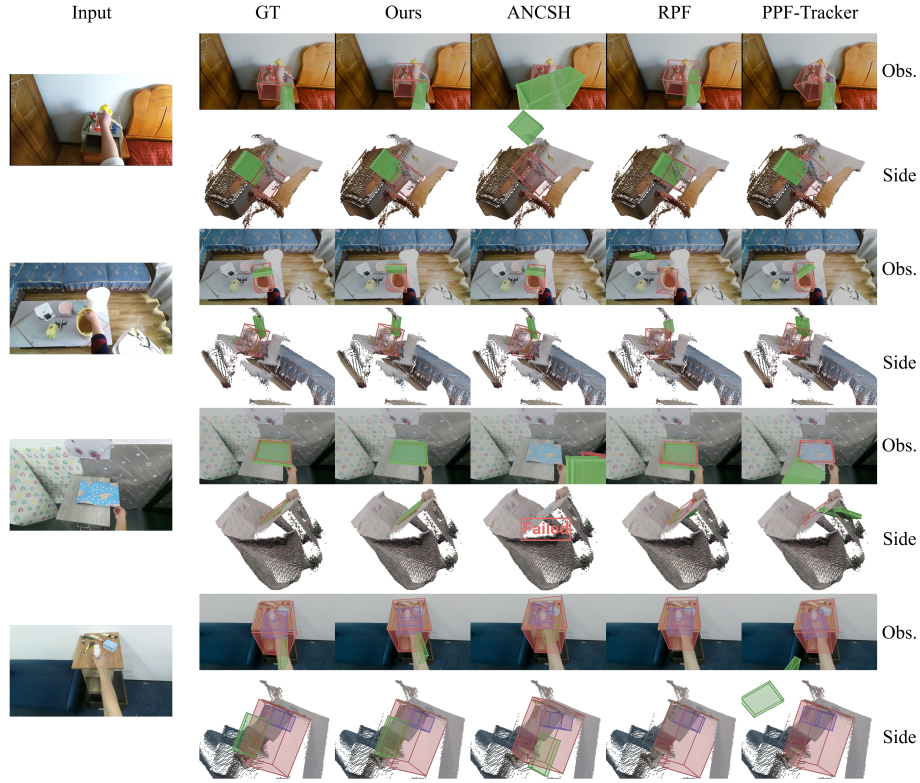


Fig. 4: Visualization for qualitative evaluation. We render the estimated poses from two viewpoints for each input frame: the first row shows the observed view (*Obs.*), and the second row shows a side view (*Side*). Bounding boxes in different colors denote the poses of specific parts. The first column presents the ground truth (*GT*), and the remaining four columns compare our framework with the baselines.

methods exhibit noticeable deviations. This further validates the robustness of our framework against severe occlusions and noisy observations. More qualitative results are provided in the supplementary material.

Shape Reconstruction. We evaluate the category-level shape reconstruction quality and compare it with the shape decoder of CARTO [12]. As shown in Tab. 3, our framework achieves a lower reconstruction error in the $CD-L_2$ metric.

4.3 Further Analysis

We visualize the pose-shape hypothesis pairs generated by our framework, which are categorized into Plausible Pairs (high score) and Implausible Pairs (low score)

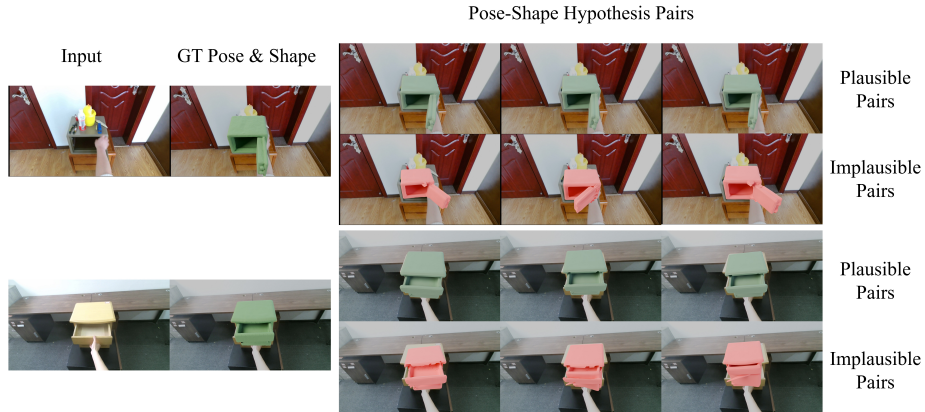


Fig. 5: Visualization of the generated pose-shape hypothesis pairs. Based on the final scores, the predictions are categorized into Plausible Pairs (high score) and Implausible Pairs (low score).



Fig. 6: Qualitative results on in-the-wild images. Predicted shapes and poses are rendered onto unseen object categories. The model is trained solely on the HOI4D dataset.

based on their final verification scores. As illustrated in Fig. 5, the rendered results of Implausible Pairs exhibit significant misalignment with the actual object, whereas Plausible Pairs align closely with the observations. This visually demonstrates our framework’s effectiveness.

To evaluate the real-world applicability of our method, we collect a set of in-the-wild images and test them using the model trained solely on the HOI4D dataset. These test cases include novel environments and unseen object categories absent from the HOI4D training set, yet they share similar underlying kinematic structures. As shown in Fig. 6, our instances include a dry cabinet

Table 3: Shape reconstruction error in $CD-L_2$. Best results are highlighted in **bold**.

Method	Laptop	StorageFurniture	Safe	TrashCan	Overall
CARTO	0.072	0.034	0.060	0.050	0.054
Ours	0.010	0.025	0.016	0.013	0.016

Table 4: Ablation study of the proposed framework. *Pose Dist.* / *Shape Dist.* map to the Pose-Shape Hypothesis Generator. *Part-Level Denoising* maps to the Part-Level Point Cloud Denoising module. *Verify Joint* and *Verify Shape* map to the Cross-Space Consistency Verification module.

Exp.	Pose Dist.	Shape Dist.	Verify Joint	Verify Shape (CD)	Verify Shape (SDF)	Part-Level Denoising	5°5cm (%) $\uparrow$	R_{err} (°) $\downarrow$	T_{err} (m) $\downarrow$
1	✓						51.1	4.96	0.057
2	✓		✓				53.8	4.87	0.044
3	✓	✓					48.3	5.33	0.061
4	✓	✓	✓				59.2	4.86	0.033
5	✓	✓	✓	✓			53.4	5.71	0.034
6	✓	✓	✓	✓		✓	60.2	4.61	0.033
7	✓	✓	✓		✓		54.4	5.82	0.035
8	✓	✓	✓		✓	✓	61.0	4.51	0.032

(top-left), a refrigerator (bottom-left), a laptop (top-right), and a microwave (bottom-right). By rendering the predicted shapes and poses onto the 2D input images, we observe a tight alignment between the overlays and the actual objects. These results indicate promising transfer to unseen categories with similar topology and part-level geometry.

4.4 Ablation Studies

To verify the effectiveness of each component proposed in our framework, we conduct an ablation study on the HOI4D dataset. Tab. 4 reports the average metrics across all tested categories.

Pose-Shape Hypothesis Generator and Verification. Experiment 1 relies on a single pose distribution. When shape hypotheses are not estimated, we use category-averaged joint parameters for verification. Comparing Experiments 1 and 2, adding this joint verification to pose-only hypotheses yields only a moderate gain, indicating that joint consistency alone cannot fully resolve shape-induced ambiguity. Experiment 3 further shows that adding shape distribution without verification enlarges the search space and degrades accuracy. In contrast, Experiment 4 couples pose-shape hypotheses with instance-specific joint verification, leading to clear improvements on all metrics.

Building on Experiment 4, we further add shape consistency verification using either unidirectional CD (Experiment 6) or SDF scoring (Experiment 8) with denoising. Both variants improve over Experiment 4, and Experiment 8 achieves the best overall performance. Overall, these results demonstrate that learning the joint distribution is essential for generating plausible hypotheses, while the verification module with the SDF-based scoring mechanism effectively prunes implausible hypothesis pairs and improves robustness under severe occlusions.

Part-Level Point Cloud Denoising. Building on Experiment 4, we evaluate two variants of shape consistency verification (Experiments 5 and 7). Initially, both variants underperform Experiment 4 because severe noise and occlusions in real-world scans often trap the verification step in local minima. However, integrating the part-level denoising module (Experiments 6 and 8) yields substantial improvements. This demonstrates that the denoising module acts as an essential step to address noisy and incomplete observations, thereby enabling reliable consistency verification downstream.

5 Conclusion

We address single-view category-level 6D pose estimation for articulated objects. To resolve ambiguities caused by severe occlusions, depth noise, and entangled pose-shape representations, we propose a generation-and-verification framework. Our method (I) generates coupled hypotheses via joint pose-shape distribution modeling, (II) recovers clean point clouds through part-level denoising, and (III) prunes implausible hypotheses using an SDF-based cross-space consistency score. Experiments on the HOI4D and PM-Videos datasets demonstrate substantial improvements in pose estimation robustness and accuracy, proving the efficacy of coupling generative modeling with explicit verification.

Limitations. A primary limitation of our framework lies in its visibility assumption for the base part. Because the joint consistency verification leverages the connected joint parameters as a foundational geometric anchor, it implicitly assumes the base part remains largely visible. Consequently, extreme occlusions of this dominant region can compromise the initial joint estimation, leading to unstable or unreliable verification results. Our framework also assumes known category topology and joint types and relies on complete-shape priors, trading scalability for accurate SDF verification; thus, it targets structurally related categories rather than open-world generalization to arbitrary articulated objects with unknown topology or substantially different shapes.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 62273318 and, in part, by the Major Program of Xiangjiang Laboratory under Grant No. 25XJJB01.

References

1. Bruns, L., Jensfelt, P.: SDFEst: Categorical Pose and Shape Estimation of Objects From RGB-D Using Signed Distance Fields. *IEEE Robotics and Automation Letters* **7**(4), 9597–9604 (2022). <https://doi.org/10.1109/LRA.2022.3189792>

2. Che, Y., Furukawa, R., Kanezaki, A.: OP-Align: Object-Level and Part-Level Alignment for Self-supervised Category-Level Articulated Object Pose Estimation. In: *Computer Vision – ECCV 2024*. pp. 72–88. Springer Nature Switzerland, Cham (2025). https://doi.org/10.1007/978-3-031-73226-3_5
3. Cheang, C., Lin, H., Fu, Y., Xue, X.: Learning 6-DoF Object Poses to Grasp Category-Level Objects by Language Instructions. In: *2022 International Conference on Robotics and Automation (ICRA)*. pp. 8476–8482 (2022). <https://doi.org/10.1109/ICRA46639.2022.9811367>
4. Chen, R., Yang, L., Yao, A.: Mhentropy: Entropy meets multiple hypotheses for pose and shape recovery. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 14840–14849 (2023)
5. Chu, R., Liu, Z., Ye, X., Tan, X., Qi, X., Fu, C.W., Jia, J.: Command-driven articulated object understanding and manipulation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8813–8823 (June 2023)
6. Desingh, K., Lu, S., Pipari, A., Jenkins, O.C.: Factored Pose Estimation of Articulated Objects using Efficient Nonparametric Belief Propagation. In: *2019 International Conference on Robotics and Automation (ICRA)*. pp. 7221–7227 (2019). <https://doi.org/10.1109/ICRA.2019.8793973>
7. Fu, L., Ishikawa, R., Sato, Y., Oishi, T.: CAPT: Category-level Articulation Estimation from a Single Point Cloud Using Transformer. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 751–757 (2024). <https://doi.org/10.1109/ICRA57147.2024.10611073>
8. Ge, Y., Xu, C., Cheng, L.: A Coarse-to-Fine Multi-Hypothesis Method for Ambiguous Hand Pose Estimation. *IEEE Transactions on Image Processing* **34**, 4302–4314 (2025). <https://doi.org/10.1109/TIP.2025.3578256>
9. Gropp, A., Yariv, L., Haim, N., Atzmon, M., Lipman, Y.: Implicit geometric regularization for learning shapes. In: *Proceedings of the 37th International Conference on Machine Learning. ICML’20*, vol. 119, pp. 3789–3799. JMLR.org (2020)
10. Han, X., Tang, Y., Wang, Z., Li, X.: Mamba3D: Enhancing Local Features for 3D Point Cloud Analysis via State Space Model. In: *Proceedings of the 32nd ACM International Conference on Multimedia*. pp. 4995–5004. MM ’24, Association for Computing Machinery, New York, NY, USA (2024). <https://doi.org/10.1145/3664647.3681173>
11. Han, X., Liu, L., Wenqiang, X., Haoyuan, F., Cewu, L.: OMAD: Object Model with Articulated Deformations for Pose Estimation and Retrieval. In: *British Machine Vision Conference (BMVC)* (2021). <https://doi.org/10.5244/C.35.155>
12. Heppert, N., Irshad, M.Z., Zakharov, S., Liu, K., Ambrus, R.A., Bohg, J., Valada, A., Kollar, T.: CARTO: Category and Joint Agnostic Reconstruction of ARTiculated Objects. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 21201–21210 (2023)
13. Huang, J., Lin, H., Wang, T., Fu, Y., Jiang, Y.G., Xue, X.: You Only Estimate Once: Unified, One-stage, Real-Time Category-Level Articulated Object 6D Pose Estimation for Robotic Grasping. In: *2025 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 14044–14051 (2025). <https://doi.org/10.1109/ICRA55743.2025.11128513>
14. Huang, J., Lin, H., Wang, T., Fu, Y., Xue, X., Zhu, Y.: CAP-Net: A Unified Network for 6D Pose and Size Estimation of Categorical Articulated Parts from a Single RGB-D Image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 11654–11664 (2025)

15. Huo, Y., Meng, X., Zhang, L., Jiang, H., Zhong, Y., Yao, M., Wang, H.: Diff-Art: Category-level Articulation Pose Estimation via Conditional Diffusion. In: 2025 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2025). <https://doi.org/10.1109/ICME59968.2025.11209596>
16. Ikeda, T., Zakharov, S., Ko, T., Irshad, M.Z., Lee, R., Liu, K., Ambrus, R., Nishiwaki, K.: DiffusionNOCs: Managing Symmetry and Uncertainty in Sim2Real Multi-Modal Category-level Pose Estimation. In: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7406–7413 (2024). <https://doi.org/10.1109/IROS58592.2024.10802487>
17. Irshad, M.Z., Zakharov, S., Ambrus, R., Kollar, T., Kira, Z., Gaidon, A.: ShAPO: Implicit Representations for Multi-object Shape, Appearance, and Pose Optimization. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. vol. 13662, pp. 275–292. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-20086-1_16
18. Jain, A., Giguere, S., Lioutikov, R., Niekum, S.: Distributional Depth-Based Estimation of Object Articulation Models. In: Proceedings of the 5th Conference on Robot Learning (2022)
19. Kingma, D.P., Dhariwal, P.: Glow: Generative Flow with Invertible 1x1 Convolutions. In: Advances in Neural Information Processing Systems. vol. 31. Curran Associates, Inc. (2018)
20. Kolotouros, N., Pavlakos, G., Jayaraman, D., Daniilidis, K.: Probabilistic Modeling for Human Mesh Recovery. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 11605–11614 (2021)
21. Lei, J., Deng, C., Shen, W.B., Guibas, L.J., Daniilidis, K.: NAP: Neural 3D Articulated Object Prior. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems. vol. 36, pp. 31878–31894. Curran Associates, Inc. (2023)
22. Li, X., Wang, H., Yi, L., Guibas, L.J., Abbott, A.L., Song, S.: Category-Level Articulated Object Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3706–3715 (2020)
23. Li, Y., Leng, W.H., Fang, Y., Eisner, B., Held, D.: FlowBotHD: History-Aware Diffuser Handling Ambiguities in Articulated Objects Manipulation. In: Proceedings of The 8th Conference on Robot Learning (2025)
24. Liu, J., Sun, W., Liu, C., Yang, H., Zhang, X., Mian, A.: MH6D: Multi-Hypothesis Consistency Learning for Category-Level 6-D Object Pose Estimation. IEEE Transactions on Neural Networks and Learning Systems pp. 1–14 (2024). <https://doi.org/10.1109/TNNLS.2024.3360712>
25. Liu, L., Huang, A., Wu, Q., Guo, D., Yang, X., Wang, M.: KPA-Tracker: Towards Robust and Real-Time Category-Level Articulated Object 6D Pose Tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 3684–3692 (2024). <https://doi.org/10.1609/aaai.v38i4.28158>
26. Liu, L., Xue, H., Xu, W., Fu, H., Lu, C.: Toward Real-World Category-Level Articulation Pose Estimation. IEEE Transactions on Image Processing **31**, 1072–1083 (2022). <https://doi.org/10.1109/TIP.2021.3138644>
27. Liu, X., Zhang, J., Hu, R., Huang, H., Wang, H., Yi, L.: Self-Supervised Category-Level Articulated Object Pose Estimation with Part-Level SE(3) Equivariance. In: The Eleventh International Conference on Learning Representations (2023)
28. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21013–21022 (2022)

29. Meng, X., Huo, Y., Zhang, L., Liu, L., Jiang, H., Zhong, Y., Zhang, P., Lu, C., Liu, J.: Exploring category-level articulated object pose tracking on se (3) manifolds. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 40, pp. 8016–8024 (2026)
30. Michel, F., Krull, A., Brachmann, E., Yang, M.Y., Gumhold, S., Rother, C.: Pose Estimation of Kinematic Chain Instances via Object Coordinate Regression. In: *British Machine Vision Conference (BMVC)* (2015)
31. Mu, J., Qiu, W., Kortylewski, A., Yuille, A., Vasconcelos, N., Wang, X.: A-SDF: Learning Disentangled Signed Distance Functions for Articulated Shape Representation. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 13001–13011 (2021)
32. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive Body Capture: 3D Hands, Face, and Body From a Single Image. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10975–10985 (2019)
33. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space. In: *Advances in Neural Information Processing Systems*. vol. 30. Curran Associates, Inc. (2017)
34. Sun, T., Zhu, L., Huang, S., Song, S., Armeni, I.: Rectified point flow: Generic point cloud pose estimation. In: *Advances in Neural Information Processing Systems (NeurIPS)* (2025)
35. Tian, Y., Zhang, J., Huang, G., Wang, B., Wang, P., Pang, J., Dong, H.: RoboKey-Gen: Robot Pose and Joint Angles Estimation via Diffusion-based 3D Keypoint Generation. In: *2024 IEEE International Conference on Robotics and Automation (ICRA)*. pp. 5375–5381 (2024). <https://doi.org/10.1109/ICRA57147.2024.10611423>
36. Wang, J., Cui, Y., Guo, D., Li, J., Liu, Q., Shen, C.: PointAttN: You Only Need Attention for Point Cloud Completion. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 38, pp. 5472–5480 (2024). <https://doi.org/10.1609/aaai.v38i6.28356>
37. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. In: *Advances in Neural Information Processing Systems*. vol. 34, pp. 27171–27183. Curran Associates, Inc. (2021)
38. Wang, R., Mao, W., Li, H.: DeepSimHO: Stable pose estimation for hand-object interaction via physics simulation. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 79685–79697. Curran Associates, Inc. (2023)
39. Wang, Y., Zhang, X., Wu, R., Li, Y., Shen, Y., Wu, M., He, Z., Wang, Y., Dong, H.: Adamanip: Adaptive articulated object manipulation environments and policy learning. In: *The Thirteenth International Conference on Learning Representations* (2025), <https://openreview.net/forum?id=Luss2sa0vc>
40. 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/CVF International Conference on Computer Vision (ICCV)*. pp. 13209–13218 (2021)
41. Wu, R., Cheng, K., Zhao, Y., Ning, C., Zhan, G., Dong, H.: Learning environment-aware affordance for 3D articulated object manipulation under occlusions. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) *Advances in Neural Information Processing Systems*. vol. 36, pp. 60966–60983. Curran Associates, Inc. (2023)

42. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., Yi, L., Chang, A.X., Guibas, L.J., Su, H.: SAPIEN: A SimulATED Part-Based Interactive ENvironment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11097–11107 (2020)
43. Xu, C., Cheng, L.: Efficient Hand Pose Estimation from a Single Depth Image. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2013)
44. Xu, C., Govindarajan, L.N., Zhang, Y., Cheng, L.: Lie-X: Depth Image Based Articulated Object Pose Estimation, Tracking, and Action Recognition on Lie Groups. *International Journal of Computer Vision* **123**(3), 454–478 (2017). <https://doi.org/10.1007/s11263-017-0998-6>
45. Xu, W., Liu, L., Zhang, L., Zhang, R., Wu, H., Guo, D., Wang, M.: KineDiff3D: Kinematic-Aware Diffusion for Category-Level Articulated Object Shape Reconstruction and Generation (2025). <https://doi.org/10.48550/arXiv.2510.17137>
46. Yen-Chen, L., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.Y.: iNeRF: Inverting Neural Radiance Fields for Pose Estimation. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1323–1330 (2021). <https://doi.org/10.1109/IROS51168.2021.9636708>
47. Yu, Q., Hao, C., Yuan, X., Zhang, L., Liu, L., Huo, Y., Agarwal, R., Lu, C.: Generalizable Articulated Object Perception with Superpoints. In: ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5 (2025). <https://doi.org/10.1109/ICASSP49660.2025.10890874>
48. Yu, X., Jiang, H., Zhang, L., Wu, L.Y., Ou, L., Liu, L.: EfficientCAPER: An End-to-End Framework for Fast and Robust Category-Level Articulated Object Pose Estimation. In: Advances in Neural Information Processing Systems. vol. 37, pp. 31968–31989 (2024). <https://doi.org/10.52202/079017-1005>
49. Yuan, C., Wen, C., Zhang, T., Gao, Y.: General Flow as Foundation Affordance for Scalable Robot Learning. In: Proceedings of The 8th Conference on Robot Learning. pp. 1541–1566. PMLR (2025)
50. Zhang, J., Wu, M., Dong, H.: Generative Category-level Object Pose Estimation via Diffusion Models. In: Advances in Neural Information Processing Systems. vol. 36, pp. 54627–54644 (2023)
51. Zhang, L., Jiang, H., Huo, Y., Zhong, Y., Wang, J., Wang, X., Wang, R., Liu, L.: R²-Art: Category-Level Articulation Pose Estimation from Single RGB Image via Cascade Render Strategy. In: Proceedings of the AAAI Conference on Artificial Intelligence (AAAI). vol. 39, pp. 9985–9993 (2025). <https://doi.org/10.1609/aaai.v39i9.33083>
52. Zhang, X., Wang, Y., Wu, R., Xu, K., Li, Y., Xiang, L., Dong, H., He, Z.: Adaptive articulated object manipulation on the fly with foundation model reasoning and part grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13032–13042 (October 2025)
53. Zhou, J., Xu, C., Tang, K., Ge, Y., Guo, T., Cheng, L.: Vpho: Joint visual-physical cue learning and aggregation for hand-object pose estimation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 13683–13691 (2026)
54. Zhu, Z., Wang, J., Qin, Y., Sun, D., Jampani, V., Wang, X.: ContactArt: Learning 3D Interaction Priors for Category-level Articulated Object and Hand Poses Estimation. In: 2024 International Conference on 3D Vision (3DV). pp. 201–212 (2024). <https://doi.org/10.1109/3DV62453.2024.00028>