SYMMETRY-AIDED DEEP CANONICAL OBJECT POSE ESTIMATION: A SIMPLE AND STRONG BASELINE

Anonymous authors
Paper under double-blind review

ABSTRACT

Canonical object pose estimation is a fundamental task in 3D computer vision, which targets establishing a unified coordinate system for objects of various categories. While object centroids are commonly utilized as coordinate origins, learning cross-category canonical orientations remains an open challenge. This paper presents a simple yet strong baseline, motivated by two key observations. First, the majority of real-world objects possess nearly symmetric geometric structures, which can provide axial cues for coordinate system hypothesis generation. Second, semantic priors are capable of resolving directional ambiguity and facilitating consistent axis alignment. Built upon the above observations, we design a three-stage pipeline for canonical pose estimation. Initially, we extract and aggregate geometric axes derived from object symmetry to generate candidate coordinate system hypotheses. Subsequently, we train a neural network to select semantically consistent "up" and "front" directions from annotated data, enabling the construction of a coarse canonical coordinate system. Finally, we conduct orientation refinement on the established coordinate system to mitigate the approximation errors introduced by symmetric initialization and discrete axis selection. We curate dedicated datasets for model training and evaluation. Experimental results validate the promising performance of our method on canonical object pose estimation tasks. We will release our code and data to facilitate future research.

1 INTRODUCTION

Object models acquired via CAD modeling, 3D reconstruction, and 3D generation often reside in arbitrary, unstandardized coordinate spaces. Canonical object pose estimation aims to establish unified coordinate systems for diverse object categories, making it a fundamental building block in 3D computer vision. In common practice, object centroids can serve as a unified coordinate origin (Chang et al., 2015; Wang et al., 2019); nevertheless, learning category-generalized canonical orientations remains an open challenge.

Recent advances in foundational models, empowered by large-scale training strategies, enable orientation estimation for arbitrary objects from images (Wang et al., 2024; 2026) and have demonstrated promising practical performance. However, 3D orientation estimates inferred from 2D images inevitably suffers from errors introduced primarily by perspective projection and image distortion. In this paper, we target canonical pose estimation directly from 3D assets.

In the literature, numerous 3D object datasets (Jin et al., 2025; Lu et al., 2025; Wu et al., 2023; Jin et al., 2026) have been constructed to facilitate the definition and estimation of canonical poses. Among these datasets, CanoVerse (Jin et al., 2026) represents the most recent effort, comprising 320K annotated 3D assets. Its practical effectiveness has been validated through consistent performance gains across various downstream tasks. Despite its intra-category alignment quality, CanoVerse suffers from cross-category inconsistencies, as its annotation standards rely on ambiguous human-defined rules. For example, vehicles such as cars and trains adopt inconsistent forward-direction definitions, while visually similar objects (e.g., towels and blankets) are annotated with conflicting canonical orientations. Such annotation ambiguities severely degrade generalization across categories, resulting in limited downstream performance and leaving the problem of unified canonical pose estimation unresolved.

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

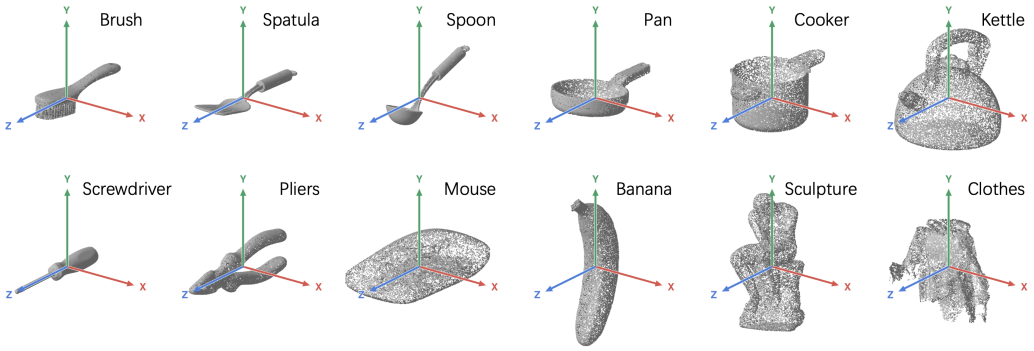


Figure 1: Our canonical pose estimation for real-world scanned objects across diverse categories. We also showcase challenging instances of non-symmetric clothing with incomplete scans. Despite its simplicity, our method achieves promising canonical pose estimation results and aligns different objects to the up (y-axis) and front (z-axis) directions.

In this work, we revisit the definition of a unified canonical coordinate system and refine the dataset by introducing additional ambiguity annotations. We standardize directional axes across semantically similar object categories according to object functionality and affordance. For example, the front direction of vehicle-related categories (e.g., airplanes, bicycles, cars, and trains) is consistently defined as the "front" direction. Similarly, the stabbing direction of daggers, pitchforks, and swords is aligned to the "front" axis. Objects with inherently ambiguous up/front directions receive dedicated labels: balloons and trees lack a well-defined front direction, while spheres and cubes possess no canonical up or front axes.

More importantly, we propose a simple yet strong baseline trained on this enhanced dataset. Our method leverages both geometric and semantic cues, motivated by two key observations: First, most real-world objects exhibit near-symmetric geometry, and symmetry naturally provides reliable cues for candidate coordinate axes. Second, semantic priors enable consistent axis direction alignment, and such priors can be learned from annotated data. Accordingly, we design a three-stage pipeline inspired by these observations: 1) generate plausible coordinate-system hypotheses using symmetry-derived geometric cues; 2) select the hypothesis with semantically consistent up and front directions; and 3) refine the estimated orientation. Despite its simplicity, this pipeline achieves strong canonical pose estimation performance across multiple datasets under diverse evaluation metrics.

We summarize our contributions as follows:

- We introduce a cross-category canonicalization principle and a three-stage canonical pose estimation pipeline. Our pipeline generates symmetry-derived coordinate hypotheses, performs semantics-aware direction selection, and conducts orientation refinement.
- We improve the CanoVerse dataset in a principled manner to unify its up and front conventions, annotate inherent directional ambiguities, and build a training set containing 250K objects. We further capture 24 diverse real-world objects to build a real-scan test set.
- Experiments on three datasets, including CanoVerse, OmniObject3D, and our newly collected real scans, show that our simple baseline achieves strong performance. Our method surpasses prior state-of-the-art approaches and provides a practical baseline for canonical object pose estimation.

2 Related Work

Object pose estimation. It is a long-standing problem in computer vision, with most prior work focusing on image-based 6-DoF pose estimation. Existing methods broadly fall into two paradigms. Instance-level methods estimate object poses with respect to pre-existing 3D models (Brachmann et al., 2014; Peng et al., 2019) or reference images (Sun et al., 2022; He et al., 2022; Wen et al., 2024). Category-level pose estimation (Wang et al., 2019; Chen et al., 2020a;b; Fan et al., 2022; Lin

et al., 2022; Wan et al., 2023; Lin et al., 2023) leverages canonicalized templates for objects within the same category. However, these approaches primarily estimate poses with respect to coordinate systems defined at either the instance or category level. They differ from our goal, which centers on defining and estimating a unified canonical pose across diverse object categories.

Driven by advances in large-scale model training, the Orient Anything series (Wang et al., 2024; 2026) achieves unified orientation estimation across diverse object categories. While their research goal is closely related to ours, their methods perform image-based estimation. Consequently, such approaches inevitably suffer from systematic information loss caused by perspective projection and geometric distortion. In contrast, our work targets canonical pose estimation on 3D assets (point clouds, the most accessible representation in 3D) and focuses on both defining and learning a unified canonical orientation for general objects.

Object and pose dataset. Existing pose estimation datasets cover both image-based (Hinterstoisser et al., 2012; Xiang et al., 2017; Wang et al., 2019) and shape-based settings (Wu et al., 2015; Chang et al., 2015; Jin et al., 2025; Lu et al., 2025; Zhang et al., 2024; Wu et al., 2023). Although these datasets provide valuable resources for pose learning, their scales remain relatively limited, with the total number of reliably canonicalized 3D objects summing to approximately 180K. Recently, CanoVerse (Jin et al., 2026) has significantly expanded this scope by providing 320K labeled 3D object models sourced from the Objaverse series (Deitke et al., 2022; 2023). While the dataset promotes intra- and cross-category pose alignment, its annotations rely on ambiguous human-defined conventions, leading to severe inter-category orientation conflicts and inherent directional ambiguities.

These inconsistencies arise from conflicting human annotation rules. Under a functionality criterion, the heads of axes, forks, and brooms are consistently defined as the front direction. In contrast, under a direction-of-action rule, fork tines remain forward-facing, while axe edges point sideways and broom bristles point downward. Such conflicts are even more pronounced for vehicle categories: cars are annotated with a z-axis-aligned front direction, whereas trains adopt an opposite front orientation. More critically, certain object categories lack definable full-axis orientations: balloons and trees have a clear up direction but no consistent front direction, while spherical objects exhibit no unambiguous up or front orientation at all. In this work, we propose a unified functional-based direction definition for canonical pose standardization. We further explicitly group each object instance to indicate whether its orientation is fully definite or inherently ambiguous along certain axes.

3 METHOD

Problem definition. Given a colored point cloud of a 3D object asset with an arbitrary pose, we first center the point cloud at the origin and normalize its scale to fit within a unit sphere. Canonical pose estimation targets reorienting the point cloud into a unified, consistent canonical frame. We adopt a right-handed Cartesian coordinate system, where only two principal axes require explicit definition, as the last axis is uniquely determined via the right-hand cross-product rule. Specifically, we define the y-axis as the "up" direction (support direction opposing gravity) and the z-axis as the "front" direction (the primary functional direction of the object). Our method outputs the point cloud consistently aligned to this canonical coordinate system.

Pipeline overview. Our method leverages complementary geometric and semantic cues for accurate canonical pose estimation. The core pipeline follows a simple yet effective paradigm: generating geometry-aware coordinate system hypotheses from object symmetry, selecting a semantically consistent optimal hypothesis across object categories, and refining the coarse orientation to eliminate residual errors. As illustrated in Figure 2, our framework consists of three progressive stages.

In the first stage, we extract symmetry-derived axes via RANSAC (Fischler & Bolles, 1981) and assemble a set of plausible coordinate system hypotheses. This ensures robustness against imperfect real-world asymmetry, and a greedy strategy is then applied to retain high-quality axis combinations for hypothesis generation. Details are described in Section 3.1. In the second stage, we deploy a neural network to select the optimal coordinate system from the generated candidates. Trained on our revised dataset, the model learns to prioritize semantically consistent up and front directions across diverse object categories. We elaborate on this hypothesis selection stage in Section 3.2.

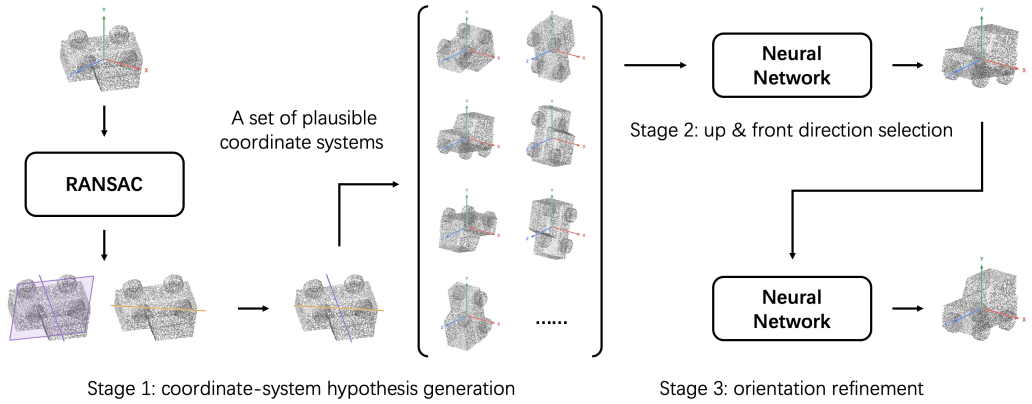


Figure 2: Illustration of our canonical pose estimation approach. Our pipeline comprises three stages. Given a point cloud with arbitrary orientation, the first stage generates coarse proposals for canonical coordinate systems derived from symmetry cues (Section 3.1). The second stage selects cross-object consistent up and front directions learned from object semantics (Section 3.2). Finally, the third stage refines the orientation of the selected coordinate system to produce fine-grained canonical poses.

In the final stage, a dedicated refinement network takes the coarsely aligned point cloud from the previous stage as input and predicts a residual rotation to further refine the orientation, enhancing our accuracy in canonical pose estimation. The details of pose refinement are presented in Section 3.3.

3.1 Coordinate-system hypothesis generation

The first stage of our method performs coordinate-system hypothesis generation. Its goal is to propose a set of initial coordinate systems suitable for subsequent tuning. For this task, we need to identify at least two orthogonal axes. Instead of relying on PCA (Jolliffe, 2002), as adopted in prior works (Jin et al., 2026), we find that symmetry cues yield better performance.

To sample candidate axes under the assumption of near-symmetric objects, we consider two common symmetry types: reflection symmetry and rotational symmetry. Both types can be effectively estimated using RANSAC (Fischler & Bolles, 1981).

Symmetry-derived axes from RANSAC. RANSAC (Random Sample Consensus) enables robust model estimation from noisy data, making it well-suited for symmetry estimation on near-symmetric objects. At its core, we sample candidate symmetry planes and axes, then evaluate their fitness by applying the corresponding symmetry transformation to all object points.

For both reflectional and rotational symmetry, we assume the symmetry plane or axis passes through the origin (object centroid). Under this setup, hypothesis generation reduces to sampling directions: for reflection symmetry, the symmetry plane is orthogonal to the sampled direction and passes through the origin; for rotational symmetry, the symmetry axis coincides with the sampled direction anchored at the origin. To generate symmetry hypotheses, we simply sample random directions uniformly from the unit sphere.

Next, we apply the candidate symmetry transformation to the point cloud and compute the Chamfer distance between the transformed points and the original ones. For reflection symmetry, the transformation corresponds to reflection across the hypothesized symmetry plane. For rotational symmetry, we test rotations by $360/N$ degrees with $N \in \{2, 3, 4, ...\}$. Hypotheses of both symmetry types are sorted in ascending order of Chamfer distance (i.e., fitness).

Coordinate-system hypothesis composition. We take the top-K symmetry-derived axes obtained from RANSAC to construct coordinate-system hypotheses. Specifically, we adopt a greedy strategy to select two orthogonal axes from the sorted list with the smallest Chamfer distances. Note that

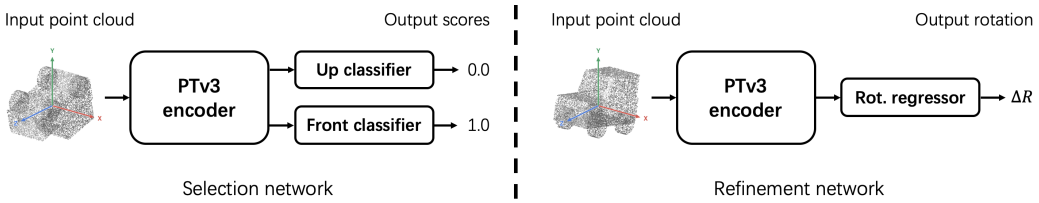


Figure 3: Illustration of our selection and refinement networks.

corner cases may arise where no valid pair of orthogonal axes can be found. In such scenarios, we fall back to PCA and use its first two principal components as the candidate axes.

Once two axes are determined, the third axis is computed via the cross product. However, these axes only encode directional lines without signed orientations, leading to 24 possible directional ambiguities. Accordingly, we generate 24 coordinate-system hypotheses for the subsequent selection stage. The output of this first stage is formulated as a set of reoriented point clouds, each aligned to one of the coordinate-system hypotheses.

3.2 Hypothesis Selection

In the second stage of our pipeline, we select the optimal coordinate-system hypothesis such that the y-axis aligns with the up direction and the z-axis aligns with the front direction. We adopt a neural network to achieve this selection. The network takes an object point cloud as input and outputs scores indicating whether the y-axis points upward and the z-axis points forward.

Selection network. As illustrated in Figure 3, we adopt a standard Point Transformer v3 (Wu et al., 2024) as our encoder to extract features from the input point cloud. The encoded features are fed into two lightweight three-layer MLPs serving as prediction heads for the up and front directions, respectively. We formulate the prediction for each direction as a separate binary classification task. Each head outputs a scalar score in $[0, 1]$, representing the predicted probability that the point cloud's y-axis points upward and the z-axis points forward.

Training supervision. The selection network is trained with cross-entropy loss. Since the input point clouds are only coarsely canonicalized in Stage 1, we cannot assume small orientation errors. Therefore, we assign a binary label of 1 to a point cloud hypothesis when its orientation deviates from the ground truth by less than 90 deg. In other words, among the 24 orientation hypotheses arising from directional ambiguity, 4 hypotheses are assigned label 1 (high target scores), while the remaining 20 are assigned label 0 (low target scores).

Inference. At test time, we sum the up and front scores for each candidate hypothesis and select the hypothesis with the highest aggregated score. The output of this second stage is a single point cloud selected from the hypothesis set generated in Stage 1. After these first two stages, the point cloud achieves initial alignment to a semantically consistent canonical pose.

3.3 Hypothesis Refinement

The final stage of our pipeline takes the coarsely reoriented point cloud as input and infers a relative rotation matrix to further refine the orientation toward the canonical pose.

Refinement network. For simplicity, our refinement network adopts the same architectural design (Figure 3) as the selection network, with the only difference being a rotation matrix prediction head (Dong et al., 2025). The network regresses rotations using a 9D representation (Levinson et al., 2020), which denotes the delta rotation required to refine the input pose toward the canonical orientation.

Training and inference. During training, the network is optimized to minimize the relative rotation error between the predicted delta rotation and the ground-truth one. At inference time, we

feed the coarsely aligned point cloud into the network, estimate the delta rotation, and apply the predicted transformation to the point cloud. Ultimately, our three-stage pipeline outputs a reoriented point cloud aligned to the canonical pose.

4 EXPERIMENTS

Implementation details. All input point clouds are downsampled to a fixed resolution of 8192 points. Color values are normalized to the range [0, 1]. For point clouds without color information, we assign a default color value of 0. In Stage 1, we set the number of random RANSAC samples to 20K for reflection and rotational symmetry estimation. Empirically, rotational angles of $180°$, $120°$, and $90°$ are sufficient to cover the dominant symmetric patterns of common objects. For both the selection and refinement networks, the Point Transformer v3 encoder is initialized with pre-trained weights from Utonia (Zhang et al., 2026) and outputs 576-dimensional features. The prediction heads adopt lightweight architectures. Specifically, each selection head is a three-layer MLP (576-512-256-1) that maps a feature vector to a scalar score. ReLU activations are applied after the first two layers, and scores over the hypothesis set are normalized via softmax. Following the standard pose regression architecture proposed in (Dong et al., 2025), the regression head maps the 576-dimensional feature to a rotation matrix.

Training data. We revised the coordinate system annotations and split the official CanoVerse dataset (Jin et al., 2026) into four subsets: SO(3), Up-only, Front-only, and CO. Through manual inspection, we confirm that objects in the SO(3) subset have consistent, well-defined "up" and "front" directions; representative examples include airplanes, alligators, buses, cameras, and dolls. Up-only and Front-only contain objects with an unambiguous up direction only (e.g., apple, balloon, cup) and an unambiguous front direction only (e.g., dagger, needle, pen), respectively. CO (cube-like orientation ambiguity) covers the remaining objects lacking both clear up and front definitions, such as balls and necklaces (ropes). We collect objects from the first three subsets within the official training split, yielding 250k objects as our training data.

Evaluation protocol. We evaluate our method on three datasets: CanoVerse (Jin et al., 2026), OmniObject3D (Wu et al., 2023), and our newly collected real-scanned object dataset. For CanoVerse, we adopt its official test split. Similar to our training data, we group it into four subsets: SO(3), Up-only, Front-only, and CO. For the SO(3) subset, we evaluate rotation accuracy under $10°$ and $30°$ error thresholds, and report the average SO(3) rotation error between the predicted and ground-truth orientations. For Up-only and Front-only subsets, we only compute the angular error along the single unambiguous axis of each object. For CO objects with full directional ambiguity, we adopt a relaxed evaluation protocol: we calculate rotation errors across all 24 valid canonical orientations and take the minimum error as the final measurement. This protocol evaluates alignment quality relative to the closest canonical pose, without relying on fixed predefined up or front conventions.

Although our model is trained solely on CanoVerse, we evaluate its out-of-distribution generalization ability on OmniObject3D and our real-world scanned data. Our real dataset is captured using a MetroX Pro scanner and consists of 24 daily-life objects, including 12 SO(3), 7 Up-only, 4 Front-only, and 1 CO object samples. We manually annotate the ground-truth canonical poses following the unified labeling framework and functional canonical definition proposed in this work. For fair and consistent evaluation across all three data sources, we uniformly convert all orientation formats to our right-handed coordinate system (y-up and z-front) before metric calculation.

4.1 COMPARISON

We select four representative state-of-the-art methods from existing literature for comparison with our proposed approach. The compared methods include two image-based orientation estimation approaches built on foundation models, namely Orient Anything (Wang et al., 2024) and Orient Anything V2 (Wang et al., 2026), as well as two point cloud-based pose estimation methods, GenPose++ (Zhang et al., 2024) and VI-Net (Lin et al., 2023). As summarized in the CanoVerse benchmark, these methods represent the major state-of-the-art solutions for canonical object orientation estimation. For the Orient Anything series, we uniformly sample six viewpoints around each object

and render corresponding images as inputs. Since GenPose++ and VI-Net share a problem setting closer to ours, we retrain them using the same training data as ours.

Table 1: Quantitative evaluation on the SO(3), Up-only, and Front-only test subsets of CanoVerse. We report accuracy (%) at 10 deg and 30 deg thresholds, as well as the absolute rotation error (deg). The best and second-best results are marked in bold and underlined, respectively.

Method	CanoVerse SO(3)		CanoVerse Up-only		CanoVerse Front-only	
	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓
Orient.	6.52 / 27.37	97.41	12.02 / 25.82	84.54	3.56 / 12.39	87.42
Orient. V2	7.22 / 30.55	73.42	14.58 / 34.54	60.21	12.35 / 31.77	57.24
GenPose++	57.19 / 68.18	<u>47.34</u>	**72.55 / 78.25**	<u>30.73</u>	40.46 / 43.45	90.39
VI-Net	<u>58.36</u> / <u>69.71</u>	47.51	58.74 / 64.48	48.14	<u>69.80</u> / <u>71.08</u>	<u>34.52</u>
Ours	**68.69 / 84.08**	**25.16**	<u>72.34</u> / <u>77.79</u>	**26.98**	**85.47 / 86.66**	**19.75**

Comparison on CanoVerse test sets. To validate the effectiveness of our method, we first conduct an in-distribution evaluation on CanoVerse with our revised test sets. Quantitative results are presented in Table 1 and Table 2, and visual comparisons are shown in Figure 4. Despite the simplicity of our approach, we observe strong performance across all four test subsets. Our method significantly outperforms all competing methods on the SO(3) and Front-only subsets. For the Up-only and CO subsets, our results are either superior or highly competitive compared with the best competitor. Interestingly, we find that Orient Anything V2 performs well on the CO subset, possibly owing to its symmetry-aware distribution learning for handling orientation ambiguity.

Table 2: Quantitative evaluation on the CO subset of CanoVerse.

Method	CanoVerse CO	
	Acc. (%) ↑	Err. (deg) ↓
Orient.	9.17 / 53.71	28.88
Orient. V2	19.69 / **80.41**	<u>19.91</u>
GenPose++	<u>37.27</u> / 65.15	21.31
VI-Net	34.70 / 69.41	20.39
Ours	**50.53** / <u>77.70</u>	**15.94**

Comparison on OmniObject3D and our real scans. To avoid potential bias, we do not partition OmniObject3D into subsets and directly use its official test split. As shown in Table 3, our method demonstrates superior or highly competitive performance to the best competing methods.

For real scans, our method outperforms all competing approaches, demonstrating its superiority for practical usage. We visualize qualitative outputs from different methods across the three datasets in Figure 4. We observe that Orient Anything V2 exhibits greater robustness than its predecessor, yet it still lacks sufficient accuracy and consistency due to its image-based formulation. GenPose++ and VI-Net struggle heavily when processing inputs with large initial rotational deviations. By contrast, our method reliably handles such challenging cases via hypothesis sampling and semantic selection; the refinement network only needs to perform local orientation tuning. Additional visual results on our real-scanned data are provided in Figure 1. Our approach yields consistent canonical pose estimations across diverse objects.

Table 3: Quantitative evaluation on OmniObject3D and our real scans.

Method	OmniObject3D		Real scans	
	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓
Orient.	5.27 / 13.77	102.17	4.35 / 17.39	97.49
Orient. V2	6.73 / 21.26	**76.48**	8.70 / 26.09	80.01
GenPose++	17.94 / 26.21	90.31	17.39 / 21.74	96.15
VI-Net	<u>21.35</u> / <u>26.58</u>	90.94	<u>26.08</u> / <u>34.78</u>	<u>79.99</u>
Ours	**28.05 / 33.13**	<u>82.15</u>	**42.03 / 63.77**	**47.79**

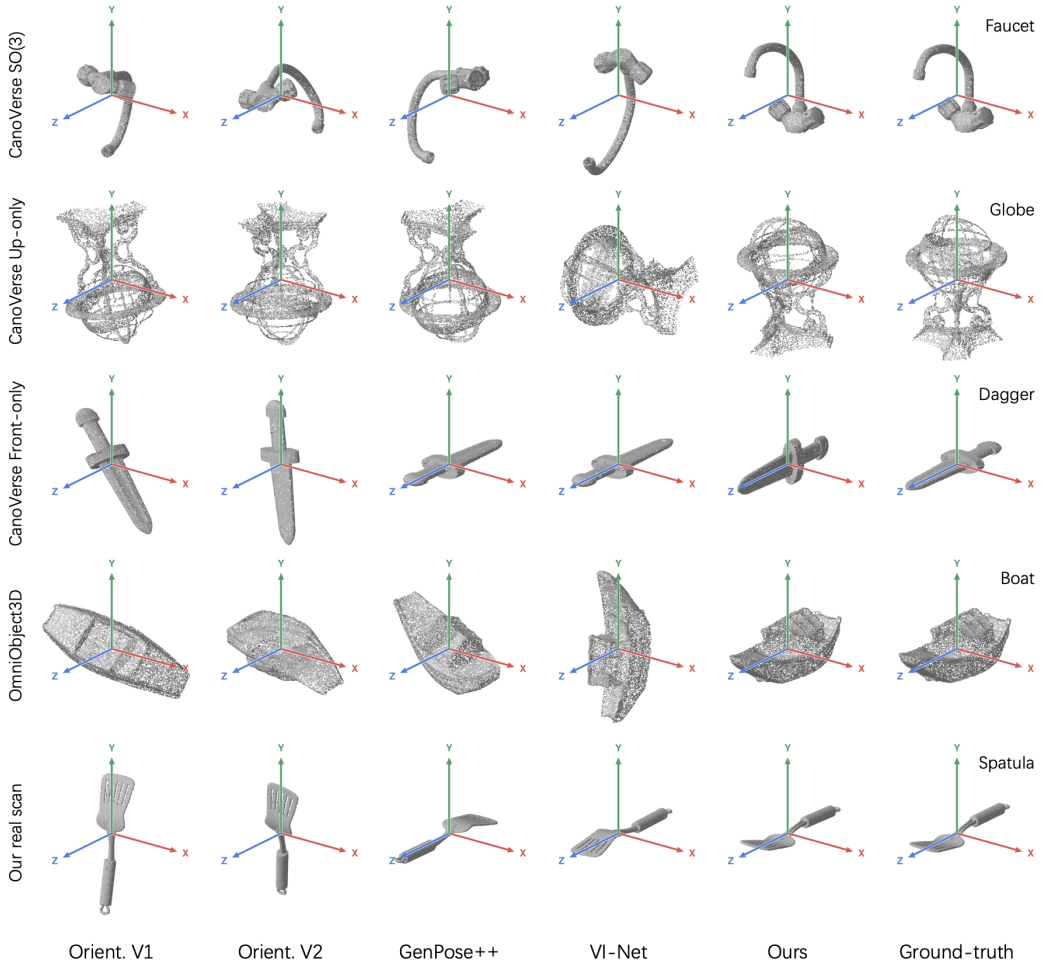


Figure 4: Visual results sampled from our evaluation on CanoVerse, OmniObject3D, and our real scans. Point clouds are rendered under predicted canonical orientations. While competing methods exhibit various failure modes, our method achieves consistent orientation across diverse objects.

4.2 Ablation Studies

In this section, we present a set of ablation studies to verify the necessity of each stage in our proposed method and analyze their individual performance.

The necessity of the three-stage pipeline. We first ablate different pipeline stages and perform evaluations on CanoVerse. As reported in Table 4, Stage 1 can generate reasonable coordinate hypotheses and achieves decent accuracy on the CO subset when used alone. However, lacking semantic directional cues, it fails on SO(3) cases that require consistent up and front estimations learned from data. Using Stage 3 alone (i.e., reorientation from arbitrarily oriented inputs) poses a challenge for our lightweight network design, verifying the importance of coarse coordinate hypothesis generation and selection. The variants 1+3 (no selection) and 1+2 (no refinement) also yield degraded performance on SO(3).

Symmetry- and PCA-based Coordinate Hypothesis Generation. We also replace our Stage 1 with a PCA-based coordinate axis proposal module. This variant is implemented using the classic principal component analysis algorithm, where we select the first two principal components as candidate canonical coordinate axes. We train Stage 2 and Stage 3 with this modified pipeline and observe degraded results (Table 4). This validates our design choice of symmetry-based coordinate hypothesis generation.

Table 4: Ablation study on our three-stage pipeline.

Pipeline stages	CanoVerse SO(3)		CanoVerse CO	
	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓
1	2.57 / 3.60	128.71	50.33 / 76.74	**15.41**
3	0.60 / 8.79	96.46	0.93 / 19.12	40.27
1+2	54.39 / 77.56	28.18	50.33 / 76.74	**15.41**
1+3	8.39 / 16.11	95.74	3.28 / 32.65	35.29
1+2+3 (Full)	**68.69 / 84.08**	**25.16**	**50.53 / 77.70**	15.94
PCA+2+3	63.34 / 83.06	25.96	46.74 / 76.30	16.82

Decoupled up and front prediction in selection network. We validate the effectiveness of our decoupled prediction design by constructing a unified single-head baseline. In this baseline, a single head is trained to identify the one positive hypothesis out of all 24 candidates, with the remaining 23 hypotheses treated as negative samples. As shown in Table 5, merging the two specialized heads degrades the accuracy across all three subsets. This demonstrates that the decoupled design is important for both individual directional axes and full canonical orientations.

Table 5: Analysis of the decoupled up and front prediction design in our selection network.

Output format	CanoVerse SO(3)		CanoVerse Up-only		CanoVerse Front-only	
	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓
Coupled	62.22 / 75.83	33.13	71.42 / 76.71	29.09	83.19 / 84.05	21.38
Decoupled	**68.69 / 84.08**	**25.16**	**72.34 / 77.79**	**26.98**	**85.47 / 86.66**	**19.75**

The effect of the training set for refinement network. The refinement network is trained to minimize the relative rotation error between predicted and ground-truth delta reorientations. Specifically, full orientation matrix errors are computed for SO(3) objects, while single-axis directional errors are adopted for Up-only and Front-only objects. We showcase the benefits of training with mixed data subsets in Table 6. It fully leverages the complementary characteristics of the three subsets and achieves superior performance compared with training on any single subset alone.

Table 6: Analysis of different training strategies for our refinement network.

Training set	CanoVerse SO(3)		CanoVerse Up-only		CanoVerse Front-only	
	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓	Acc. (%) ↑	Err. (deg) ↓
SO(3)	67.52 / 82.07	25.12	69.68 / 76.50	27.27	85.14 / 86.47	19.27
Up-only	51.58 / 75.98	29.54	71.46 / 76.96	26.67	85.42 / 86.42	19.47
Front-only	51.33 / 77.44	28.78	68.22 / 76.71	27.37	**85.47** / 86.47	**19.01**
Combined	**68.69 / 84.08**	**25.16**	**72.34 / 77.79**	**26.98**	**85.47 / 86.66**	19.75

5 CONCLUSION

In this work, we study canonical object pose estimation and present a baseline, dataset, and evaluation suite for this task. We leverage geometric symmetry and semantic functional priors to establish a unified canonical pose definition across diverse object categories. We develop an effective three-stage pipeline that first generates symmetry-based coordinate system hypotheses, selects semantically consistent directions to resolve inherent orientation ambiguities, and finally refines coarse poses for accurate alignment. Furthermore, we revise an existing large-scale dataset and construct curated training and real-world test sets with unified orientation conventions and explicit ambiguity annotations, establishing a more reliable benchmark for canonical pose learning. Extensive experiments on both synthetic and real-scan data demonstrate that our simple framework outperforms state-of-the-art methods, serving as a practical baseline for future research.

AI USE STATEMENT

In this work, we used generative AI tools to assist with code debugging and grammar checks during paper writing. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pp. 536–551. Springer, 2014.

Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11970–11979. IEEE, 2020a.

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

Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. *arXiv preprint arXiv:2212.08051*, 2022.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. *arXiv preprint arXiv:2307.05663*, 2023.

Siyan Dong, Shuzhe Wang, Shaohui Liu, Lulu Cai, Qingnan Fan, Juho Kannala, and Yanchao Yang. Reloc3r: Large-scale training of relative camera pose regression for generalizable, fast, and accurate visual localization. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16739–16752. IEEE, 2025.

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

Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24 (6):381–395, 1981.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022.

Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pp. 548–562. Springer, 2012.

Li Jin, Yujie Wang, Wenzheng Chen, Qiyu Dai, Qingzhe Gao, Xueying Qin, and Baoquan Chen. One-shot 3d object canonicalization based on geometric and semantic consistency. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16850–16859. IEEE, 2025.

Li Jin, Yuchen Yang, Weikai Chen, Yujie Wang, Dehao Hao, Tanghui Jia, Yingda Yin, Zeyu Hu, Runze Zhang, Keyang Luo, et al. Canoverse: 3d object scalable canonicalization and dataset for generation and pose. *arXiv preprint arXiv:2603.07144*, 2026.

Ian T Jolliffe. *Principal component analysis*. Springer, 2002.

Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of svd for deep rotation estimation. *Advances in Neural Information Processing Systems*, 33:22554–22565, 2020.

Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *European Conference on Computer Vision*, pp. 19–34. Springer, 2022.

Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13955–13965. IEEE, 2023.

Yichong Lu, Yuzhuo Tian, Zijin Jiang, Yikun Zhao, Yuanbo Yang, Hao Ouyang, Haoji Hu, Huimin Yu, Yujun Shen, and Yiyi Liao. Orientation matters: Making 3d generative models orientation-aligned, 2025. URL https://arxiv.org/abs/2506.08640.

Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4561–4570, 2019.

Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6815–6824. IEEE, 2022.

Boyan Wan, Yifei Shi, and Kai Xu. Socs: Semantically-aware object coordinate space for category-level 6d object pose estimation under large shape variations. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14019–14028. IEEE, 2023.

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2637–2646. IEEE, 2019.

Zehan Wang, Ziang Zhang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything: Learning robust object orientation estimation from rendering 3d models. *arXiv:2412.18605*, 2024.

Zehan Wang, Ziang Zhang, Jiayang Xu, Jialei Wang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything v2: Unifying orientation and rotation understanding. *Advances in Neural Information Processing Systems*, 38:52622–52645, 2026.

Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17868–17879. IEEE, 2024.

Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 803–814. IEEE, 2023.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4840–4851. IEEE, 2024.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1912–1920, 2015.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

Jiyao Zhang, Weiyao Huang, Bo Peng, Mingdong Wu, Fei Hu, Zijian Chen, Bo Zhao, and Hao Dong. Omni6dpose: A benchmark and model for universal 6d object pose estimation and tracking. In *European Conference on Computer Vision*, pp. 199–216. Springer, 2024.

Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds. *arXiv preprint arXiv:2603.03283*, 2026.

APPENDIX

A RANSAC PARAMETER TUNING

We randomly sampled 3K objects from CanoVerse (Jin et al., 2026) and conducted an analysis of how the number of random samples affects RANSAC performance. As shown in Figure 5, the minimum Chamfer distance at the 50th and 90th percentiles across all objects drops rapidly as the sampling count initially increases; thereafter, marginal improvements diminish. Accordingly, we adopt 20K samples as our default setting.

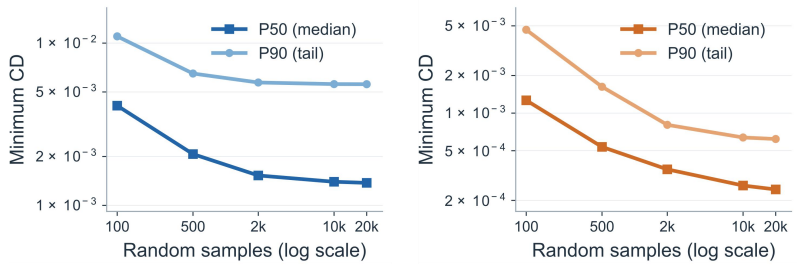


Figure 5: Left: rotational symmetry; right: reflection symmetry. As the number of random samples increases, the probability of finding optimal symmetry planes/axes rises, exhibiting diminishing returns. We select 20K samples as our default setting.

B MORE VISUAL RESULTS ON OUR REAL SCAN

Due to limited space, we show 12 real scans in Figure 1 in the main paper. The remaining 12 visual results are shown in Figure 6.

C LIMITATION

We observe two main failure cases, as illustrated in Figure 7. First, for objects with strong symmetry such as spheres, numerous candidate axes exist. This leads to ambiguity when selecting orthogonal axes to build the coordinate system; such objects also suffer from directional ambiguity in defining the up and front directions, e.g., balls and wheels. Second, for highly asymmetric objects, reasonable candidate axes are difficult to identify via symmetry error (Chamfer distance) sorting. For these scenarios, we resort to PCA instead of symmetry search to construct candidate axes.

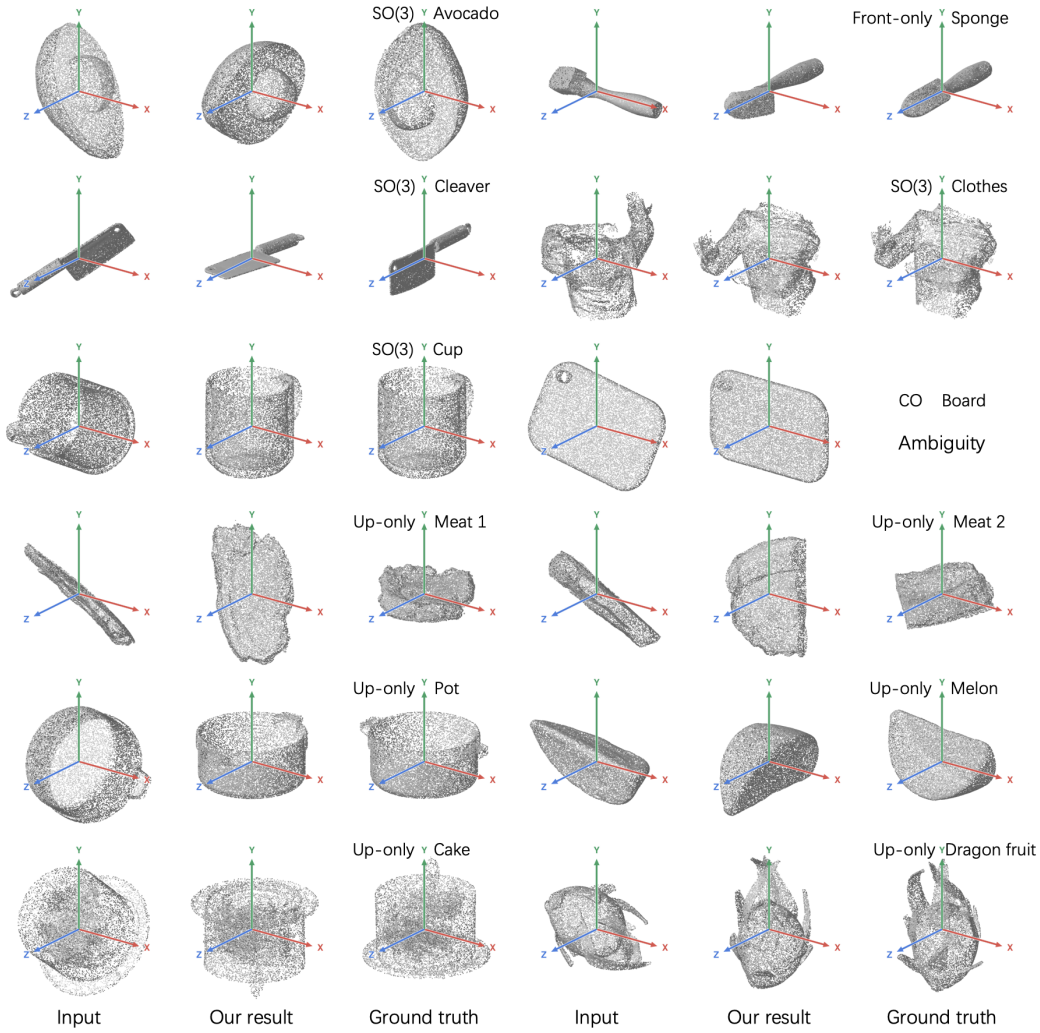


Figure 6: Visual results of our canonical pose estimation on real scanned objects.

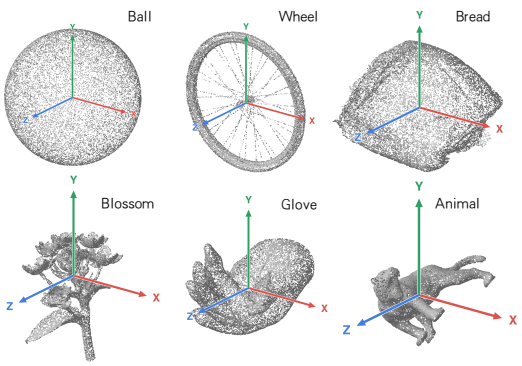


Figure 7: Up: strong symmetry; bottom: asymmetry.