One-shot 3D Object Canonicalization based on Geometric and Semantic Consistency

Li Jin[1,7] Yujie Wang[3,4] Wenzheng Chen[5,6] Qiyu Dai[3]
Qingzhe Gao[3] Xueying Qin[1,7] Baoquan Chen[2†]

[1]School of Software, Shandong University
[2]State Key Laboratory of General Artificial Intelligence, Peking University
[3]School of Intelligence Science and Technology, Peking University
[4]UNC Chapel Hill [5]Wangxuan Institue of Computer Technology, Peking University
[6]State Key Laboratory of Multimedia Information Processing, Peking University, Beijing, P.R. China
[7]Engineering Research Center of Digital Media Technology, Ministry of Education, P.R. China

Abstract

3D object canonicalization is a fundamental task, essential for various downstream tasks. Existing methods rely on either cumbersome manual processes or priors learned from extensive, per-category training samples. Real-world datasets, however, often exhibit long-tail distributions, challenging existing learning-based methods, especially in categories with limited samples. We address this by introducing the first one-shot category-level object canonicalization framework that operates under arbitrary poses, requiring only a single canonical model as a reference (the "prior model") for each category. To canonicalize any object, our framework first extracts semantic cues with large language models (LLMs) and vision-language models (VLMs) to establish correspondences with the prior model. We introduce a novel joint energy function to enforce geometric and semantic consistency, aligning object orientations precisely despite significant shape variations. Moreover, we adopt a support-plane strategy to reduce search space for initial poses and utilize a semantic relationship map to select the canonical pose from multiple hypotheses. Extensive experiments on multiple datasets demonstrate that our framework achieves state-of-the-art performance and validates key design choices. Using our framework, we create the Canonical Objaverse Dataset (COD), canonicalizing 32K samples in the Objaverse-LVIS dataset, underscoring the effectiveness of our framework on handling large-scale datasets. Project page at https://github.com/JinLi998/CanonObjaverseDataset

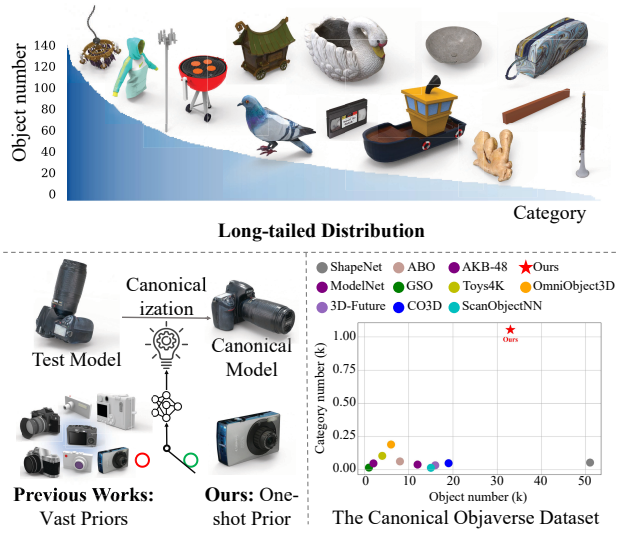

Figure 1. **A one-shot approach for 3D object canonicalization.** The number of objects in different categories follows a long-tail distribution in the real world, as shown by Objaverse-LVIS [6] (top). Existing methods rely on many prior conditions, which are often hard to meet. Our one-shot approach effectively addresses this issue (bottom left). We also introduce the Canonical Objaverse Dataset (bottom right), which features the largest number of categories among current canonical 3D datasets.

1. Introduction

3D object canonicalization, which transforms 3D shapes into a canonical space, ensuring consistent rotation, translation, and scaling across different instances, is crucial for creating large-scale 3D object datasets. Canonical 3D object datasets serve as the foundation for many downstream

†: Baoquan Chen is the corresponding author.
The work is done while the first author was visiting Peking University.

tasks, including 3D shape recognition and retrieval [8, 9], 3D object understanding [10, 12], and pose estimation [14, 30]. Furthermore, in the field of Artificial Intelligence Generated Content for 3D (AIGC3D), recent studies [16, 22, 29] indicate that training with canonical datasets significantly enhances generation quality.

Traditional methods, like ShapeNet [4] and NOCS [25], rely heavily on manual canonicalization, which is labor-intensive and prone to inconsistencies, particularly when handling large, diverse 3D model collections. The limitations in efficiency and consistency of manual canonicalization have driven recent advances in learning-based approaches [2, 21]. These methods facilitate the canonicalization of diverse shapes within the same category by leveraging self-supervised learning and learning implicit category-level canonical fields. However, learning-based methods depend on geometric constraints and require substantial amounts of objects for training, making data collection resource-intensive and labor-demanding.

Noteworthy, real-world datasets, such as Objaverse [6], exhibit a long-tail distribution [15, 27], challenging self-supervised approaches due to limited data in most categories. Additionally, category-level canonical fields are closely tied to object semantics; for example, camera lenses should maintain a consistent orientation in canonical fields. As a result, learning a canonical field based solely on geometric constraints—without incorporating semantic relationships—becomes particularly challenging for categories with significant intra-category shape variations.

To address the challenges, we propose a one-shot category-level canonicalization approach that considers both geometric and semantic consistency. Unlike previous methods requiring numerous objects as priors, our approach only needs a single canonical object (the prior model). For handling objects from arbitrary categories, we design a zero-shot 3D perception module that utilizes large language models (LLMs) and vision-language models (VLMs) to perceive 3D shapes in arbitrary poses. To improve efficiency, we employ a support-plane-based object initialization strategy (the "support-plane strategy") to generate multiple initial pose hypotheses. For aligning test models with the prior model, we develop a joint energy function that combines semantic and geometric consistency for coarse and fine alignment respectively. Finally, we filter canonicalization hypotheses through a semantic relationship map to determine the canonical pose.

We evaluate our approach on several benchmark datasets [4, 5, 25, 27], achieving improvements of 63% to 85% on the GEC metric over existing methods. Additionally, applying our method to the Objaverse-LVIS dataset [6] enables us to create the Canonical Objaverse Dataset (COD) that features 1,054 categories and 32k shapes, showcasing its effectiveness in handling large datasets. Our contributions are:

- We introduce the first one-shot category-level object canonicalization framework operating under arbitrary poses, requiring only a single model per category and eliminating the need for extensive training samples.
- We propose a carefully designed joint strategy that uniquely integrates semantic and geometric cues for more accurate alignment. Also, by devising a multi-hypothesis initialization and selection scheme, we achieve superior pose alignment performance.
- We present the Canonical Objaverse Dataset (COD), featuring the largest number of categories among current canonical 3D datasets, providing a valuable resource for research in downstream tasks.

2. Related Work

3D Object Canonicalization: Traditional canonical datasets, such as ShapeNet [4] and NOCS [25], are created through computer-assisted manual alignment of 3D shapes. However, manipulating 3D objects with manual processes at a large scale is cumbersome.

To tackle these challenges, learning-based methods [21, 24] seek to derive category-level canonical fields from diverse shapes using a self-supervised approach. ConDor [21] employs Tensor Field Networks to create a canonical field, producing equivariant poses. The method [2] has been further extended to canonicalize objects represented by Neural Radiance Fields. Additionally, ShapeMatcher [9] improves segmentation and retrieval functions through object canonicalization. Some studies [23] focus on canonicalizing similar objects by aligning them in a reference video, particularly under common poses.

These methods demonstrate strong generalization capabilities. However, real-world 3D shape data, following a long-tail distribution [6, 15, 27], with limited sufficient samples, which restricts the practical application of these methods. Moreover, most approaches rely on geometric energy functions for constraints. Since individual objects can vary significantly in shape, minimizing geometric similarity does not ensure semantic consistency.

3D Object Dataset: Despite advancements in AI with models like ChatGPT [1] and CLIP [20], the development of large-scale 3D models is hindered by a shortage of 3D datasets, limiting progress in tasks like object generation.

While existing datasets have significantly contributed to the field, they still face limitations. 3D canonical datasets such as ShapeNet [4], ModelNet3D [28], NOCS [25], and OmniObject3D [27] are vital for 3D perception, reconstruction, and generation. ShapeNet [4] contains 51,300 CAD models across 55 categories, while OmniObject3D [27] includes 6,000 scanned objects with textures and multi-view photos across 190 categories. However, these datasets depend on manual alignment and lack sufficient size for training large models.

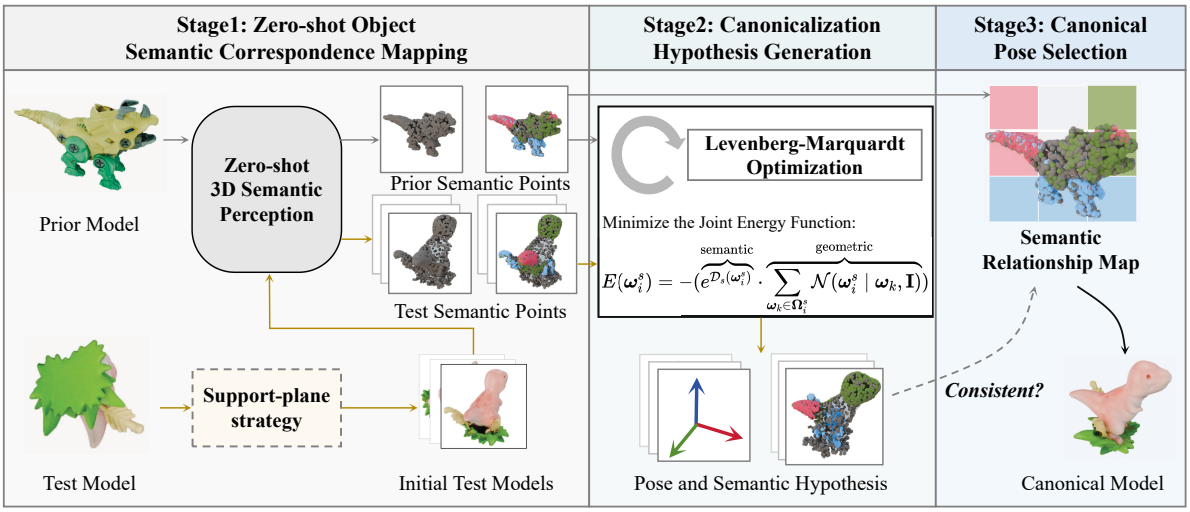

Figure 2. **Method overview.** Our approach enables category-level object canonicalization using a single prior model for each category. We begin by utilizing large language models (LLM) and vision-language models (VLM) to capture the 3D semantics of both the prior model and the test model, establishing semantic correspondences (left). Next, we generate canonical pose hypotheses and introduce a joint energy function that integrates semantic and geometric cues, facilitating accurate alignment with the prior model (middle). Finally, we identify the optimal canonical pose using a semantic relationship map (right) by evaluating the consistency of semantic positions.

In contrast, the internet provides a vast but challenging resource of 3D data, much of which is unaligned, low-quality, and inaccurately labeled. Objaverse-LVIS [6] offers a curated set of 47,000 objects across 1,156 LVIS categories, with CLIP-selected candidates. Objaverse 1.0 [6] provides 800K annotated 3D objects from Sketchfab, while Objaverse-XL [7] adds over 10 million web-sourced objects. Efficient cleaning and canonicalization of this data are essential for advancing 3D vision tasks.

3. Method

3.1. Overview

Our method aims at achieving category-level object canonicalization in a one-shot manner. Specifically, given a single canonical prior model as an alignment basis, our method automatically canonicalizes test models in arbitrary poses from the same category. The canonicalization process involves canonicalizing a test model's translation, scale and rotation. Following previous approaches [2, 21], we handle translation and scale through centering and scaling the object within its bounding box, while focusing primarily on the most challenging aspect—rotation canonicalization. As illustrated in Figure 2, our method comprises three main stages: Zero-shot Object Semantic Correspondence Mapping, Canonicalization Hypothesis Generation, and Canonical Pose Selection.

In Stage 1, as Figure 2 shows, we establish semantic correspondences between the test model and the prior model. As we aim at handling arbitrary categories without training

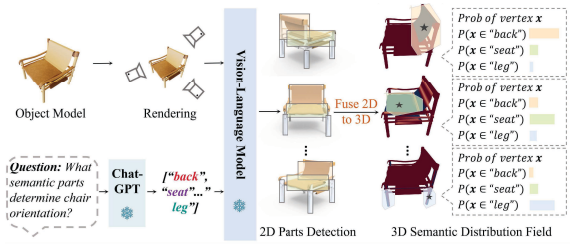

Figure 3. **Zero-shot 3D semantic perception.** We achieve 3D semantic perception by detecting parts from 2D images. In this example, we first render the object model and obtain semantic labels using Chat-GPT (left). Next, we utilize the VLM Glip to perform part detection for each semantic label (middle). Finally, our aggregation method projects the 2D detected parts onto 3D (right), assigning probabilities to the labels at each vertex of the model.

tailored models with extensive data, we develop a Zero-Shot 3D Semantic Perception Module (Section 3.2) leveraging large language models (LLMs) and visual-language models (VLMs). To mitigate potential detection errors incurred by merely considering a single initial pose, we introduce a support-plane strategy that initializes the test model with multiple rotational poses. This stage generates the prior semantic points and the test semantic points, which correspond to each other through semantic labels.

In Stage 2, canonicalization Hypothesis Generation (Section 3.3), we generate a set of canonical pose hypotheses for each sampled initial pose of the test model. We propose a joint energy function that considers both semantic and geometric cues, which enables more accurate alignment of the test model with the prior model.

In Stage 3, the Canonical Pose Selection (Section 3.4), we determine the final canonical pose, based on the generated pose hypotheses, by evaluating the consistency of relative semantic positions between the prior model and the test model.

3.2. Zero-Shot 3D Semantic Perception

Semantic cues are helpful to identify corresponding regions among different objects even with substantial shape variations. However, current supervised 3D segmentation methods [19, 26] fall short on arbitrary unseen categories. We opt to leverage the zero-shot semantic understanding abilities of LLMs and VLMs, originally developed for language and 2D images, to achieve zero-shot 3D semantic perception.

As Figure 3 shows, we begin by rendering 10 multi-view images of objects and use an LLM (ChatGPT [3]) to generate semantic part labels in a zero-shot manner. More details are provided in the Supplementary Material. Then we apply a VLM (GLIP [13]) to perform zero-shot parts detection for each view. Finally, these 2D parts are projected into 3D space to produce the final 3D perception.

The GLIP-based semantic perception module is not trained or finetuned on specialized datasets, often resulting in detection errors. To mitigate this, we constrain the 2D bounding box of a part's semantic instance within the object's 2D bounding box, which guides the adjustment of 2D detection results. Moreover, the ambiguities detection in GLIP can lead to inconsistencies across views. Inspired by [17], we introduce semantic probabilities to address this.

For an object model with v vertices, one of vertices is denoted as a vector $\mathbf{x}_l \in \mathbb{R}^3$, $l = 1, \ldots, v$. All vectors of vertices on this object are constructed as a matrix $\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \ldots & \mathbf{x}_v \end{bmatrix} \in \mathbb{R}^{3\times v}$. The probability P_l^s is defined as that $\mathbf{x}_l$ belongs to a specific semantic label $s \in \{1, 2, \ldots, n\}$, where n is the number of semantic labels. Then, we define the semantic confidence vector $\mathbf{c}_l \in \mathbb{R}^n$ of each vertex $\mathbf{x}_l$ as

$$\mathbf{c}_l = (P_l^1, P_l^2, \ldots, P_l^n)^\top, l \in \{1, \ldots, v\}. \tag{1}$$

All these $\mathbf{c}_l$ consist of an matrix $\mathbf{C} = \begin{bmatrix} \mathbf{c}_1 & \mathbf{c}_2 & \ldots & \mathbf{c}_v \end{bmatrix} \in \mathbb{R}^{n\times v}$. The semantic points are then defined as $(\mathbf{X}, \mathbf{C})$. For the prior model, we calculate its semantic points $(\mathbf{X}_r, \mathbf{C}_r)$, where the subscript r distinguishes variables for the prior model from those for the test model.

When dealing with test models, we notice that the long-tail issues in the training data for GLIP model [17] hinder reliable detection across diverse poses. Specifically, as shown in Figure 4, VLM can produce accurate detection only for common poses, such as upright ones, making zero-shot 3D perception insufficient for handling a wide range of poses. This creates a chicken-and-egg dilemma between

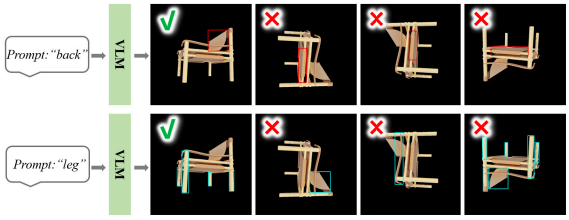

Figure 4. **Characteristics of 2D vision language model.** Taking Glip [13] as an example, the parts detection results for the same object vary under different poses.

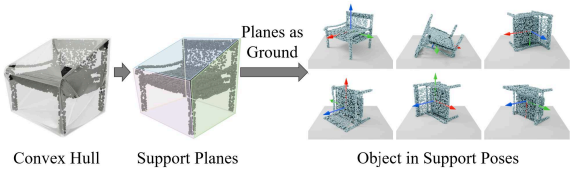

Figure 5. **Support-plane strategy.** We initialize the object's pose based on stable support on the ground. First, we compute the 3D convex hull of the object (left). Then, by calculating the relationship between the center and the projections of the convex hull facets, we identify potential support planes (middle). Finally, we align the support plane with the ground by rotating the object, establishing the initial pose (right).

achieving a canonical pose and accurate semantic understanding. Exhaustively exploring the rotation space to find an optimal initialization pose is also impractical. To address this, we propose a support-plane strategy that initializes the object pose with several candidate poses, corresponding to stable resting positions. This strategy efficiently identifies a minimal set of poses that optimize GLIP's detection accuracy. This initialization strategy is elaborated below.

Support-Plane-based Object Initialization. As Figure 5 shows, we identify static equilibrium for a rigid body when, as stated in [11], its center of mass lies above the support polygon. Using this criterion, we determine support planes for a given mesh and establish candidate poses. In the following, we use the subscript i and t to indicate initial and test models, respectively.

Specifically, for the i-th supporting surface, we calculate a rotation matrix $\mathbf{R}_i \in \mathbb{R}^{3\times3}, i \in \{1, \ldots, m\}$ to transform the test model $\mathbf{X}_t \in \mathbb{R}^{3\times v_t}$ so that the surface is parallel to the ground. Details on computing the rotation matrix $\mathbf{R}_i$ are provided in the Supplementary Material. Finally, we obtain an initialization rotation matrix set $\mathcal{R}_{\text{init}} = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_m\}$, where m represents the number of supporting surfaces. By transforming the test model $\mathbf{X}_t$ using the transformations within $\mathcal{R}_{\text{init}}$, we obtain a set of initial point cloud $\mathcal{X}_{\text{init}}$ for the test model $\mathbf{X}_t$, given by

$$\mathcal{X}_{\text{init}} = \{\mathbf{R}_1\mathbf{X}_t, \mathbf{R}_2\mathbf{X}_t, \ldots, \mathbf{R}_m\mathbf{X}_t\}. \tag{2}$$

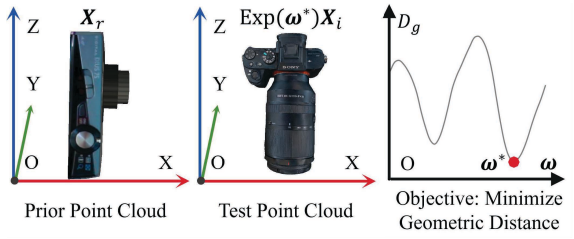

Figure 6. **Geometry canonicalization leads to orientation inconsistency.** The prior (left) semantic orientation is different from the test model (middle), optimized by the geometry constraint(right).

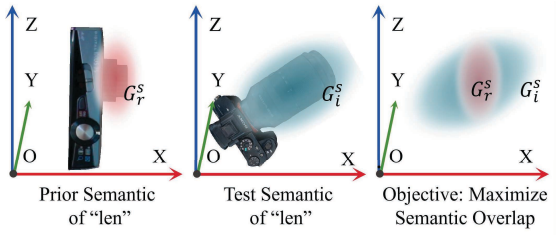

Figure 7. **Semantic canonicalization leads to inaccuracies in geometry.** The prior semantic (left) does not align with the geometric details of the test model (middle), when maximizing the overlap of semantic distributions (right).

Feeding the test model under various initial poses into the Zero-Shot 3D Semantic Perception Module, yields a set of initial semantic points $\mathcal{I}$ for the test model:

$$\mathcal{I} = \{(\mathbf{X}_i, \mathbf{C}_i) \mid \mathbf{X}_i \in \mathcal{X}_{\text{init}}, i = 1, \ldots, m\}. \tag{3}$$

Since the supporting surface defines the object's support axis, we only need to determine its horizontal orientation during the Canonicalization Hypothesis Generation phase.

3.3. Canonicalization Hypothesis Generation

Using the prior semantic points of the model $(\mathbf{X}_r, \mathbf{C}_r)$ and those for the test model, $\{(\mathbf{X}_i, \mathbf{C}_i)\}_{i=1}^m$, we generate a set of canonical hypotheses for the final canonical pose. In the process, we calculate an individual rotation $\mathbf{R}_i^s \in \mathbb{R}^{3\times3}$ for each initial test model and each semantic label. Specifically, we optimize $\mathbf{R}_i^s$ using its Lie algebra representation $\boldsymbol{\omega}_i^s \in \mathfrak{so}(3)$ by aligning the initial test semantic points $(\mathbf{X}_i, \mathbf{C}_i)$ with the prior semantic points $(\mathbf{X}_r, \mathbf{C}_r)$ based on the distribution of the s-th semantic label. This process helps improve the alignment, we introduce a joint energy function that incorporates both geometric and semantic cues.

Geometry-Based constraint. Inspired by [2, 21], geometry-based constraints enhance alignment by measuring the similarity of shapes, leading to greater overlap in similar regions. Therefore, we measure the geometric similarity using Chamfer distance $\mathcal{D}_g(\boldsymbol{\omega}_i^s)$:

$$\begin{aligned}\mathcal{D}_g(\boldsymbol{\omega}_i^s) = &\sum_{\mathbf{x}\in\mathbf{X_r}} \min_{\mathbf{y}\in\text{Exp}(\boldsymbol{\omega}_i^s)\mathbf{X}_i} \|\mathbf{x}-\mathbf{y}\|^2 \\ &+ \sum_{\mathbf{y}\in\text{Exp}(\boldsymbol{\omega}_i^s)\mathbf{X}_i} \min_{\mathbf{x}\in\mathbf{X}_r} \|\mathbf{y}-\mathbf{x}\|^2,\end{aligned} \tag{4}$$

where Exp() denotes the function that converts a Lie algebra vector into its corresponding rotation matrix by applying the matrix exponential function [18]. While the geometric constraint excels in local alignment, it often gets stuck in local minima due to a lack of semantic guidance. As Figure 6 illustrates, even minimizing the geometric constraint to maximize the overlap between the test model and the prior model can result in an undesired camera orientation.

Semantic-Based constraint. When incorporating semantic constraints, we observe that the 3D semantics from the zero-shot perception module are low-quality and noisy. To address this, we represent the semantics with a 3D Gaussian distribution, which effectively filters out noise. Specifically, we model the prior model distribution of s-th semantical label $\mathcal{G}_r^s$ as a normal distribution $\mathcal{N}(\boldsymbol{\mu}_r^s, \boldsymbol{\Sigma}_r^s)$. Similarly, the semantic distribution of the test model under i-th initialization, $\mathcal{G}_i^s$, is modeled as $\mathcal{N}(\boldsymbol{\mu}_i^s(\boldsymbol{\omega}_i^s), \boldsymbol{\Sigma}_i^s(\boldsymbol{\omega}_i^s))$. Details on calculating the mean and variance for these distributions are provided in the Supplementary Material. We then define a metric $\mathcal{D}_s(\boldsymbol{\omega}_i^s)$ measuring the degree of overlap of the semantic distributions between the prior model and the initial model:

$$\mathcal{D}_s(\boldsymbol{\omega}_i^s) = \left(1 + \frac{\langle \boldsymbol{\mu}_r^s, \boldsymbol{\mu}_i^s(\boldsymbol{\omega}_i^s)\rangle}{\|\boldsymbol{\mu}_r^s\|\|\boldsymbol{\mu}_i^s(\boldsymbol{\omega}_i^s)\|}\right)^2. \tag{5}$$

This metric serves as semantic guidance for optimizing the rotation $\boldsymbol{\omega}_i^s$, by increasing the alignment of the semantic distributions as the metric increases.

Despite semantic constraints providing information about the object's orientation, achieving precise alignment using semantics alone is challenging. As Figure 7 shows, semantic guidance only offers a rough direction, but lacks precise alignment of detailed components.

Joint energy function. To incorporate semantic and geometric guidance for coarse and fine alignment respectively, we design a joint energy function. Instead of a weighted combination of semantic and geometric constraints, our energy function builds a Gaussian mixture model to better capture local geometry while considering semantic confidence. The energy function $E(\boldsymbol{\omega}_i^s)$ is defined as

$$E(\boldsymbol{\omega}_i^s) = -(\overbrace{e^{\mathcal{D}_s(\boldsymbol{\omega}_i^s)}}^{\text{semantic}} \cdot \overbrace{\sum_{\boldsymbol{\omega}_k\in\boldsymbol{\Omega}_i^s} \mathcal{N}(\boldsymbol{\omega}_i^s \mid \boldsymbol{\omega}_k, \mathbf{I})}^{\text{geometric}}), \tag{6}$$

where $\boldsymbol{\Omega}_i^s = \{\boldsymbol{\omega}_i^s \mid \frac{\partial}{\partial\boldsymbol{\omega}}\mathcal{D}_g(\boldsymbol{\omega}_i^s) = 0\}$ and $\mathbf{I}$ is the identity matrix. By considering all extrema derived from the geometric constraints and weighting them according to the overlap with the prior model's semantic distribution, this energy

Table 1. **Few-shot 3D object canonicalization on the ShapeNet dataset.** Lower scores indicate better performance.

Method	Prior num.	*Car*		*Table*		*Chair*		*Plane*		*Couch*		*Lamp*		*Water.*	
		IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC
CaCa [24]	10	1.520	1.476	2.251	2.668	1.930	2.110	1.054	1.119	1.925	1.981	3.335	4.316	1.731	1.964
ConDor [21]	10	0.870	1.146	1.721	2.918	0.976	1.898	0.742	0.978	0.932	1.759	2.810	4.798	1.137	1.525
ShapeMat. [9]	10	1.242	2.301	1.214	2.181	1.071	1.996	0.878	1.855	1.171	2.956	**1.509**	4.461	0.734	3.099
Ours	1	**0.077**	**0.087**	**0.702**	**0.783**	**0.558**	**0.656**	**0.224**	**0.238**	**0.479**	**0.544**	2.651	**2.874**	**0.141**	**0.157**
Method	Prior num.	*Bench*		*Speaker*		*Cabinet*		*Firearm*		*Monitor*		*Cell.*		**Avg.**	
		IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC	IC	GEC
CaCa [24]	10	2.754	2.776	1.762	**2.021**	2.158	2.208	2.171	2.188	1.281	1.446	1.017	1.324	1.915	2.123
ConDor [21]	10	3.065	3.953	**1.161**	2.105	**0.962**	1.612	1.255	1.481	1.062	1.547	1.514	2.387	1.401	2.162
ShapeMat. [9]	10	2.149	4.226	1.303	2.203	1.517	2.238	2.625	4.140	2.076	3.082	3.499	4.675	1.614	3.032
Ours	1	**0.355**	**0.411**	1.945	2.134	0.987	**1.160**	**0.261**	**0.275**	**0.324**	**0.364**	**0.430**	**0.504**	**0.703**	**0.784**

function allows more precise alignment. The effectiveness of the energy function is investigated in the ablation studies (see Table 4).

As the energy function (Eq. (6)) is non-differentiable, we use finite differences to compute the gradient and optimize it using the Levenberg-Marquardt optimization method. Finally, the optimum for $\boldsymbol{\omega}_i^s$ is calculated via

$$\hat{\boldsymbol{\omega}}_i^s = \arg\min_{\boldsymbol{\omega}_i^s} E(\boldsymbol{\omega}_i^s). \tag{7}$$

Then a set of hypotheses for the final canonical pose, denoted as $\mathcal{R}_{\text{hypo}}$, is obtained by

$$\mathcal{R}_{\text{hypo}} = \left\{\hat{\mathbf{R}}_i^s = \text{Exp}(\hat{\boldsymbol{\omega}}_i^s)\mathbf{R}_i \mid i \in \{1, 2, \ldots, m\}\right\}_{s=1}^{n}. \tag{8}$$

3.4. Canonical Pose Selection

From multiple canonical pose hypotheses within $\mathcal{R}_{\text{hypo}}$, we identify the most accurate one as the final canonical pose by analyzing the relationships among semantic parts. As shown in the right part of Figure 2, an object in the correct canonical pose aligns with the prior model's semantic distribution, even with significant shape differences. Measuring this consistency allows us to filter out incorrect hypotheses that may arise from segmentation errors.

In this process, we first construct semantic relation maps in the canonical space for both the prior model and the test model. Specifically, we divide the canonical space evenly into B^3 blocks and calculate semantic weights w_j^s for each label in each cubic block $\mathbf{b}_j$ ($j \in \{1, 2, \ldots, B^3\}$) as follows:

$$w_j^s = \int_{\mathbf{b}_j} \mathcal{G}_r^s(\mathbf{x})\, d\mathbf{x}, \quad s = 1, \ldots, n. \tag{9}$$

Next, we assign the label with the highest weight to each block to form the semantic map: $S_j = \arg\max_s w_j^s$. This process yields the semantic relation map $\mathcal{M}_r = \{S_j \mid j \in \{1, 2, \ldots, B^3\}\}$ in canonical space for the prior model. Similarly, we generate a set of semantic relation maps $\{\mathcal{M}_i\}_{i=1}^m$ by using $\mathcal{G}_i^s$ for the test model under each canonical pose hypothesis.

Finally, by computing the cosine similarity between the vectorized M_r and each M_i, we obtain a score for each canonical pose hypothesis. From the top five highest-scoring hypotheses, we select the one with the highest **geometric** similarity as the final pose.

Table 2. **Few-shot 3D object canonicalization on the NOCS dataset.**

Method	**Prior num.**	*Laptop*		*Mug*		*Bowl*		*bot-*
		IC	**GEC**	**IC**	**GEC**	**IC**	**GEC**	**IC**
CaCa [24]	10	1.925	2.300	1.641	1.812	1.365	1.458	1.673
ConDor [21]	10	0.841	1.260	0.114	0.146	0.548	0.664	0.734
ShapeMat. [9]	10	2.549	4.124	0.550	0.682	1.612	1.831	1.599
Ours	1	**0.187**	**0.222**	**0.091**	**0.082**	**0.033**	**0.036**	**0.144**
Method	**Prior num.**	*-tle*	*Camera*		*Can*		**Avg.**	
		GEC	**IC**	**GEC**	**IC**	**GEC**	**IC**	**GEC**
CaCa [24]	10	1.790	2.089	2.236	0.727	0.733	1.570	1.722
ConDor [21]	10	0.768	0.686	1.236	0.871	1.393	0.679	0.876
ShapeMat. [9]	10	2.628	**0.624**	1.279	1.473	1.958	1.401	2.084
Ours	1	**0.149**	0.874	**1.067**	**0.099**	**0.099**	**0.145**	**0.129**

4. Experiments

4.1. Experiment Setup

Datasets. Following the prior work [2, 21], we assess our framework on 13 categories from the simulated dataset **ShapeNet** [4]. We also conduct evaluations on two real datasets: **NOCS** [25] containing 6 categories with 36 real-world textureless objects affected by artifacts including holes and noise; **DREDS** comprising 7 categories with 42 objects that are scanned to produce high-quality meshes with detailed textures. Furthermore, we apply our method to uncanonicalized objects from **OmniObject3D** [27] and **Objaverse-LVIS** [6].

Evaluation metrics. We adopt two metrics for evaluation: Instance-Level Consistency (IC) to measure the normalization consistency of a single instance model across different poses, and Ground Truth Equivariance Consistency (GEC) to evaluate the consistency of normalized models among different intra-category instances. Following the previous methods [2, 21], we compute numerical errors by sampling 1,024 points from the surface point clouds of each test model and the associated ground-truth model. Notably, we observe that the Category-Level Consistency (CC) metric degrades when the canonical transformation is the identity, as discussed in Supplementary Material.

Baselines. We compare our method with existing object

Table 3. **Few-shot 3D object canonicalization on the DREDS dataset.**

Method	Prior num.	*Aeroplane*		*Car*		*Bowl*		*Bottle*	
		IC	**GEC**	**IC**	**GEC**	**IC**	**GEC**	**IC**	**GEC**
CaCa [24]	10	0.785	0.869	1.985	1.881	1.569	1.895	2.632	3.080
ConDor [21]	10	0.823	1.391	0.605	0.782	0.898	1.202	1.108	1.261
ShapeMat. [9]	10	0.659	2.337	2.409	3.596	0.368	0.452	1.279	3.791
Ours	1	**0.051**	**0.058**	**0.103**	**0.118**	**0.011**	**0.012**	**0.031**	**0.034**
Method	**Prior num.**	*Camera*		*Can*		*Mug*		**Avg.**	
		IC	**GEC**	**IC**	**GEC**	**IC**	**GEC**	**IC**	**GEC**
CaCa [24]	10	1.294	1.451	1.797	2.095	0.661	0.655	1.532	1.704
ConDor [21]	10	1.035	1.529	1.048	1.753	0.126	0.189	0.806	1.158
ShapeMat. [9]	10	**0.377**	1.491	1.369	2.211	0.399	0.629	0.980	2.072
Ours	1	1.116	**1.177**	**0.030**	**0.037**	**0.018**	**0.019**	**0.194**	**0.208**

canonicalization approaches [9, 21, 24]. For a fair comparison, we retrained these methods in few-shot settings. Comparative results using the full training dataset are provided in the Supplementary Material.

4.2. Comparative Results

Evaluation on simulation Dataset. Table 1 reports the performance of different methods on the ShapeNet dataset [4]. Using only one prior model, our method distinctly outperforms existing methods. Specifically, our method demonstrates at least 49.8% and 63.1% improvements on average IC and GEC separately, demonstrating its effectiveness across diverse categories and shapes. Visual comparisons (first row) in Fig. 8 further demonstrate our method's effectiveness. Using only one prior model, which is in stark contrast to prior methods, our method achieves precise canonical perception. For instance, for the *Chair* category, previous methods result in compromised results by focusing solely on geometric alignment, while ours maintains consistent semantic orientation across shapes.

Evaluation on real Dataset. Real-world scanned data poses challenges due to limited data amount and quality variations, such as noise, holes, and missing textures, creating a domain gap with simulated data. Tables 2 and 3 report the numerical results on low-quality NOCS [25] and high-quality DREDS [5], where our method consistently outperforms prior approaches despite domain gaps. Fig. 8 (second and third rows) further highlights our method's robustness, even dealing with significant shape differences in *Camera* category or issues like missing textures and noise on *Laptop* category. Additional results can be found in Supplementary Material.

Evaluation in the wild data. We conduct qualitative tests on wild data from the OmniObject3D [27] and Objaverse-LVIS [6] datasets. As Figure 9 shows, our method performs well on objects with significant variations in both shape and color, demonstrating its robustness in practical applications. Additional results are provided in Supplementary Material.

4.3. Ablation and Analysis

Role of key designs from our framework. We analyze our key design choices: 1) the adoption of both geometric constraints and semantic cues from our zero-shot 3D semantic perception module, 2) the proposed joint energy function, and 3) the multi-hypothesis pose initialization and selection scheme. The results of ablation studies are presented in Table 4.

Table 4. **Results for ablation studies.** The experiments are conducted on the *DREDS* dataset. "Multi-Hypotheses" represents the multi-hypothesis pose initialization and selection strategy.

Geometric Constraint	Semantic Constraint	Weighted Combination	Full Energy Func.	Multi-Hypotheses	IC	GEC
✓					0.696	0.724
	✓				2.213	2.315
✓	✓	✓			0.802	0.870
✓	✓		✓		0.621	0.690
✓	✓		✓	✓	**0.194**	**0.208**

Table 4 shows that only using geometric constraints leads to a considerable increase in error, often due to local minima that lack semantic alignment (discussed in Section 3.3). Meanwhile, using only 3D semantic cues from the zero-shot perception causes a significant error increase, as semantics inherently contain noise and can only provide a rough alignment direction. A naive weighted combination of geometric and semantic constraints remains heavily affected by semantic noise, resulting in higher errors than using geometric constraints alone. In contrast, our proposed joint energy function effectively leverages semantic guidance with geometric constraints to achieve more precise alignment, as evidenced by significantly lowered errors. By recognizing the impact of different initial poses on semantic errors, our support-plane-based multi-hypothesis initialization strategy, combined with a selection process based on semantic relationship maps, further reduces errors considerably.

Demanding Sample Analysis for Prior Methods. We examine how the number of priors (training samples) affects the performance of Condor [21], the best-performing existing method. Specifically, we train Condor with varying numbers of priors and evaluate the models on the *Aeroplane* and *Laptop* categories, with results shown in Fig. 10. As illustrated, Condor generally converges after 200 priors, achieving a similar or slightly lower performance than our method. Noteworthy, real-world datasets often follow a long-tail distribution, which challenges existing techniques on categories with limited data. For instance, in the Objaverse-LVIS dataset, only 2 of 1,156 categories have over 200 objects, and only 74 have over 100. This means over 93% of categories lack the priors required by previous methods.

4.4. Application on Large-Scale Dataset

We apply our method to canonicalize unaligned objects in the large-scale Objaverse-LVIS dataset [6]. After the canonicalization process, we render the objects from fixed view-

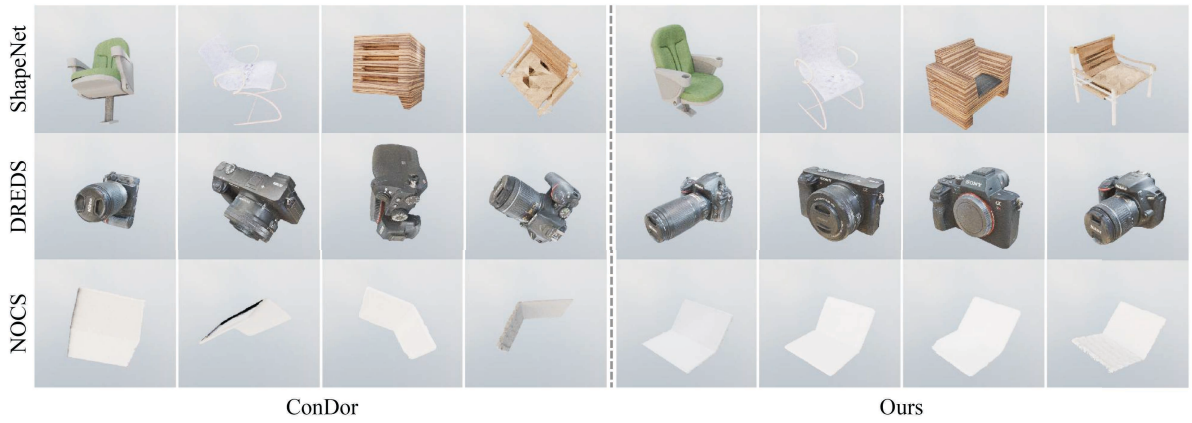

Figure 8. **Visual results of 3D object canonicalization on the ShapeNet dataset [4] (top), DREDS dataset [5] (middle), and NOCS dataset [26] (bottom).** The left and right columns display results from the ConDor method [21] and results from our method respectively.

Figure 9. **Visual results of 3D object canonicalization in the wild.** The top two rows are from OmniObject3D [27], while the bottom two rows are from Objaverse-LVIS [6].

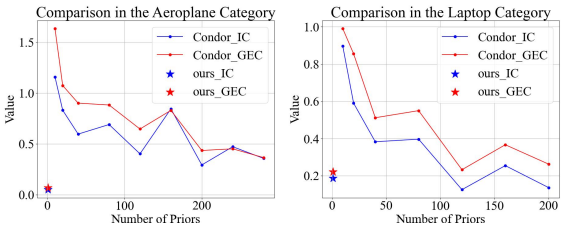

Figure 10. **Ablation studies on the effect of varying the number of priors.** We conduct experiments on the areoplane (left) and laptop (right) categories. Lower values indicate better performance.

points to filter out those with poor mesh quality, improper alignment, or incorrect labels. The resulting dataset, Canonical Objaverse Dataset (COD), includes 1,054 categories and 32K shapes, making it the largest 3D canonical dataset regarding category count. Figure 1 compares COD to existing canonical datasets, demonstrating its superiority.

5. Conclusion

We present a novel one-shot 3D object canonicalization framework that uses a single prior. Leveraging LLMs and VLMs for zero-shot 3D semantic perception, we design an energy function for coarse semantic and precise geometric alignment. Evaluations on simulated and real-world datasets show it outperforms prior methods, with the Canonical Objaverse Dataset (COD) demonstrating scalability for large datasets.

Acknowledgments

This work is partially supported by the National Key R&D Program of China under grants (No.2022ZD0160801, 2022YFB3303203), and the NSF of China (No.62172260). We sincerely thank Pengshuai Wang, Xifeng Gao and Jia Li for fruitful diseussions.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Rohith Agaram, Shaurya Dewan, Rahul Sajnani, Adrien Poulenard, Madhava Krishna, and Srinath Sridhar. Canonical fields: Self-supervised learning of pose-canonicalized neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4500–4510, 2023.

[3] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[4] 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.

[5] Qiyu Dai, Jiyao Zhang, Qiwei Li, Tianhao Wu, Hao Dong, Ziyuan Liu, Ping Tan, and He Wang. Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects. In *European Conference on Computer Vision*, pages 374–391. Springer, 2022.

[6] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Yan Di, Chenyangguang Zhang, Ruida Zhang, Fabian Manhardt, Yongzhi Su, Jason Rambach, Didier Stricker, Xiangyang Ji, and Federico Tombari. U-red: Unsupervised 3d shape retrieval and deformation for partial point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8884–8895, 2023.

[9] Yan Di, Chenyangguang Zhang, Chaowei Wang, Ruida Zhang, Guangyao Zhai, Yanyan Li, Bowen Fu, Xiangyang Ji, and Shan Gao. Shapematcher: Self-supervised joint shape canonicalization segmentation retrieval and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21017–21028, 2024.

[10] Niladri Shekhar Dutt, Sanjeev Muralikrishnan, and Niloy J Mitra. Diffusion 3d features (diff3f): Decorating untextured shapes with distilled semantic features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4494–4504, 2024.

[11] Hongbo Fu, Daniel Cohen-Or, Gideon Dror, and Alla Sheffer. Upright orientation of man-made objects. In *ACM SIGGRAPH 2008 papers*, pages 1–7. 2008.

[12] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T. Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[13] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022.

[14] Xiaolong Li, Yijia Weng, Li Yi, Leonidas J Guibas, A Abbott, Shuran Song, and He Wang. Leveraging se (3) equivariance for self-supervised category-level object pose estimation from point clouds. *Advances in neural information processing systems*, 34:15370–15381, 2021.

[15] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[16] Fangfu Liu, Diankun Wu, Yi Wei, Yongming Rao, and Yueqi Duan. Sherpa3d: Boosting high-fidelity text-to-3d generation via coarse 3d prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20763–20774, 2024.

[17] Minghua Liu, Yinhao Zhu, Hong Cai, Shizhong Han, Zhan Ling, Fatih Porikli, and Hao Su. Partslip: Low-shot part segmentation for 3d point clouds via pretrained image-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21736–21746, 2023.

[18] Yi Ma, Stefano Soatto, Jana Košecká, and Shankar Sastry. *An invitation to 3-d vision: from images to geometric models*. Springer, 2004.

[19] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *Advances in neural information processing systems*, 35:23192–23204, 2022.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] Rahul Sajnani, Adrien Poulenard, Jivitesh Jain, Radhika Dua, Leonidas J Guibas, and Srinath Sridhar. Condor: Self-supervised canonicalization of 3d pose for partial shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16969–16979, 2022.

[22] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[23] Leonhard Sommer, Artur Jesslen, Eddy Ilg, and Adam Kortylewski. Unsupervised learning of category-level 3d pose from object-centric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22787–22796, 2024.

[24] Weiwei Sun, Andrea Tagliasacchi, Boyang Deng, Sara Sabour, Soroosh Yazdani, Geoffrey E Hinton, and Kwang Moo Yi. Canonical capsules: Self-supervised capsules in canonical pose. *Advances in Neural information processing systems*, 34:24993–25005, 2021.

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019.

[26] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics (tog)*, 38(5):1–12, 2019.

[27] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023.

[28] 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*, pages 1912–1920, 2015.

[29] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024.

[30] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In *European Conference on Computer Vision*, pages 655–672. Springer, 2022.