C3PO: Canonicalization of 3D Pose from Partial Views with Generalizable Correspondence Features

Yu Chi[1,*] Leonhard Sommer[2] Olaf Dünkel[3] Dominik Muhle[1]
Daniel Cremers[1] Christian Theobalt[3] Adam Kortylewski[4]

[1]Technical University of Munich [2]University of Freiburg
[3]Max Planck Institute for Informatics, Saarland Informatics Campus
[4]CISPA Helmholtz Center for Information Security

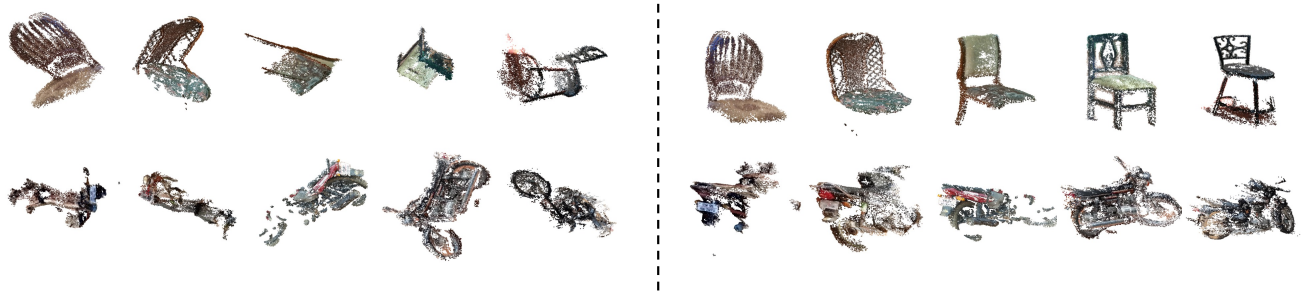

Figure 1. **Canonicalization of partial 3D reconstructions.** Given non-canonicalized partial 3D shapes reconstructed from object-centric videos (left), our method aligns them into a shared canonical space (right), enabling consistent comparison and downstream processing.

Abstract

Progress in 3D object understanding relies on the category-level canonicalization of 3D objects, i.e., bringing 3D instances into a consistent position and orientation. Most related works assume complete 3D representations, while real-world applications often require solving the more challenging task of canonicalizing from partial views, i.e., short videos that cover only a part of the object. We introduce C3PO, a method capable of canonicalizing partial views from arbitrary object categories by enforcing geometric and feature-level appearance consistency of overlapping views. We represent partial views as 3D point clouds obtained via structure-from-motion, where each point carries a feature vector that is extracted from 2D images using a novel feature extractor capable of estimating generalizable correspondence features. Notably, our correspondence features are learned on a large dataset and generalize to object categories not seen during training. On top of this, we introduce an efficient pairwise-registration framework that aligns partial object representations into a globally consistent canonical frame. Experiments on synthetic and real-world benchmarks demonstrate that C3PO significantly outperforms existing methods.

1. Introduction

Tasks that require understanding of 3D objects, such as object recognition [41], object retrieval [10, 11], 3D pose estimation [50, 54, 62], and learning category-level morphable models [6, 9, 20, 31, 45, 61], have been extensively studied in the computer vision community. A common prerequisite across these tasks is category-level canonicalization: mapping all objects of a given category into a shared, standardized pose, orientation, and scale, the canonical frame. This allows comparing, analyzing, and processing them consistently. Canonicalization has also been shown to benefit the generation of high-quality 3D assets [25, 59].

In most existing benchmarks, category-level canonicalization is achieved through manual alignment during dataset

*Work done during an internship at MPI for Informatics.

preparation, ensuring that all instances are consistently oriented before training and evaluation. While this has enabled progress in many downstream tasks, the manual labeling severely limits scalability, both in terms of the number of categories and captured instances per category.

Automating canonicalization becomes even more challenging when only partial observations of an object are available, e.g., 3D point clouds reconstructed from a short video showing part of the object. Yet, this scenario is common in real-world applications such as robotics, AR/VR, and 3D content creation, where full object geometry is rarely accessible. Achieving accurate canonicalization from incomplete data would not only remove the manual bottleneck but also extend the benefits of consistent alignment to long-tail object categories and in-the-wild captures. Figure 1 illustrates this problem setting.

Prior works on automatic canonicalization have typically considered complete 3D coverage of each object. Some study the canonicalization of point clouds from synthetic data [38, 44, 46] or real data [22], where each object instance is fully observed from one or more scans and alignment is based solely on geometry. However, purely geometric approaches struggle for categories with multiple symmetry axes, such as keyboards and laptops, where appearance cues are essential to resolve ambiguity. Other works address canonicalization from videos with 100% orbit coverage[40, 43], but most in-the-wild videos capture only partial views, which can be observed for egocentric footage (e.g., EGO4D[19]) and for many object-centric videos (e.g., CO3D [35], MVImgNet [55]).

In this work, we study *category-level canonicalization* from partial views of object-centric videos, considering 10%, 25%, and 50% coverage. To this end, we introduce two benchmarks: one using synthetically rendered ShapeNet videos [4], and one using real-world object-centric videos from CO3D [35].

To address this challenging task, we propose **C3PO**, **C**anonicalization of **3**D **PO**se from partial views with generalizable correspondence features. Unlike prior work [43], which canonicalizes each source object to a single target object, we jointly canonicalize multiple partial source objects and enforce global consistency among them through overlapping regions. This requires simultaneously estimating canonical transformations and identifying overlapping regions, a particularly challenging optimization problem. We tackle this with an efficient global-consistency framework based on pairwise registration, combined with a correspondence feature extractor trained on the large-scale ImageNet3D dataset [26]. Our learned features can reliably determine, for example, whether an RGB image of an unseen instance was captured from the left or right side of the object in canonical space. Thanks to large-scale training, they generalize beyond the categories of the training distribution. Finally, within the global-consistency optimization, we incorporate semantic distances between partial views into the loss function, allowing the model to effectively account for viewpoint variations. Across both synthetic and real-world benchmarks, C3PO significantly outperforms the prior state-of-the-art method for challenging partial-view settings (10%, 25%, 50%).

2. Related Work

Category-level canonicalization of partial views for common object categories remains largely unexplored. Pioneering works are bounded to animal categories [23, 52, 53] or require CAD models and RGB-D input [58]. However, we discuss methods for canonicalization of 3D shapes in Sec. 2.1, and methods designed for 100% orbit coverage in Sec. 2.2. Further, we provide an overview for correspondence feature extractors in Sec. 2.3.

2.1. Canonicalization of Shapes

Self-supervised category-level canonicalization has advanced significantly with $SE(3)$-invariant and equivariant networks. Early research prioritized rotation and translation invariance [8, 32–34], while subsequent methods developed equivariant architectures [5, 7, 16, 47, 48, 60, 63]. By combining both, recent works enable canonicalization of arbitrary point clouds [22, 38, 44, 46]. However, these approaches require extensive per-category training data, limiting their scalability to ”long-tail” object categories.

Partial-to-partial shape matching, despite its significant practical importance, remains relatively underexplored. For humans and animals, however, several works exist. [24] introduce a theoretical framework for partial-to-partial matching but lack practical evaluation. [2] study overlapping regions between partial shapes without addressing shape matching. [1] and [36] advance the problem using deep learning and combinatorial optimization, respectively; however, Sm-comb is limited to water-tight shapes. [14] propose a method that neither requires 3D ground-truth labels nor hole filling. All these works focus on humans and animals, where topology changes can be ignored. In contrast, our work targets arbitrary object categories, including those with topology changes, such as chairs.

Overall, all these methods are not leveraging appearance. Therefore, posing the task of canonicalization close to infeasible for object categories with two more or symmetry axes such as mouse, keyboard, or laptop.

2.2. Canonicalization of Complete Views

With the increased availability of large scale datasets of object-centric videos such as CO3D [35], category-level canonicalization has received more attention. The completely unsupervised method, proposed by [17], builds upon DINO [3]. First, they find two images that are taken from

a similar viewpoint using the image features. Second, they filter pixel-wise feature correspondences to get relative pose estimation. Further, the authors extend their work in [18], where they find a consensus over many images from both videos. Sommer et al. [43] introduce an efficient representation to take into account more images, adding a geometric constraint besides the appearance correspondences, and formulating the cyclic distances for down-weighting in 3D euclidean instead of 2D pixel space. In contrast to these works, we are addressing the ambiguities of DINO correspondences given few viewpoints, and are maximizing the appearance and geometric consistency over many videos.

2.3. Correspondence Feature Learning

Matching semantic keypoints across images from arbitrary object categories is a challenging but important computer vision task [28]. Self-supervised models have shown remarkable performance in zero-shot semantic correspondences from self-superivsed models like DINOv2 [3] and Stable Diffusion [37]. However, recent studies [12, 27, 57] have demonstrated that DINOv2 features struggle to accurately distinguish between object symmetries and individual parts. Mariotti et al. [27] introduced a novel method for estimating semantic correspondences that enhances DINO features with 3D awareness through a coarse geometric spherical prior. Further, Zhang et al. [57] leverages viewpoint augmentations to account for viewpoint dependency. Shtedritski et al. [42] showed that by leveraging 3D shape representations, the consistency of 2D-3D correspondences may improve overall performance significantly. However, the necessity of a detailed categorical mesh, restricts this applicability to a small selected set of animals. In contrast to others, we train a feature extractor that generalize to novel categories leveraging images of over a hundred categories.

3. Method

In this section, we describe our approach for canonicalizing partial views. First, we transform the partial views into the object-centric surface features representation, see Sec. 3.1. These are then aligned to ensure global geometric and appearance consistency (Sec. 3.2). Furthermore, we introduce a method for obtaining generalizable correspondence features from foundational models that facilitates the global consistency in comparison with DINOv2 and, hence, improves the canonicalization Sec. 3.3.

3.1. Partial Views as Surface Features

We represent the partial views as object-centric surface features, to partially capture the geometry as well as the appearance of each object. This is similar to other works that represent categories or complete objects as 3D object geometries with attached surface features [29, 43, 49, 56]. We use the combination of DINOv2 and generalizable correspondence features in this step. The surface features $\mathrm{S} = \{\mathrm{V}, \mathrm{F}\}$ comprise a set of vertices V and a set of features F. By using structure-from-motion [39] we obtain depth maps, and by using masked backprojecting we obtain a point cloud. We approximate this geometry with a triangular mesh using alpha shapes [13], resulting in the vertices $\mathrm{V} = \{\mathrm{v}_i \in \mathbb{R}^3\}_{i=1}^{|\mathrm{V}|}$. Further, we capture the appearance per vertex with semantic features from multiple viewpoints as follows $\mathrm{F} = \{\{\mathrm{f}_i^k \in \mathbb{R}^D\}_{k=1}^{|\mathrm{F}_i|}\}_{i=1}^{|V|}$. Features are accumulated by projecting visible vertices into the 2D feature image from an image encoder.

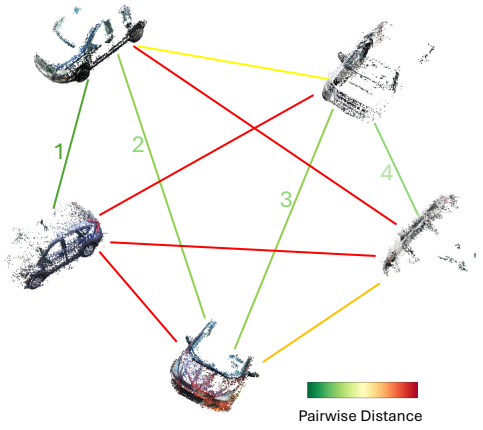

Figure 2. **Pairwise registration for global consistency.** Illustration of pairwise registration for achieving global consistency across multiple partial scans. This method involves sequentially aligning scans to build a coherent global 3D structure. The numbers on the edges indicate the registration sequence. Initially, scans predominantly capturing the left side are aligned (edge 1). The third scan, including views from both the left and right sides, is then integrated (edge 2), serving as a bridge for subsequent right-side dominant scans (edges 3 and 4). This ensures global consistency even in the absence of direct overlap between some scan pairs. Edge colors reflect the distance between scan nodes: red indicates larger distances, while green indicates shorter distances.

3.2. Global Consistency via Pairwise Registration

Achieving global consistency with respect to geometry and appearance by transforming all object-centric surface features in one canonical coordinate frame is highly challenging. Especially, because some partial views may have no or only minimal pairwise overlap. Imagine the illustrative scenario where one partial view covers an object from the left side and another one covers another object from the right side. As there is no common region of the two objects the problem itself is ill-posed. Hence global consistency for many partial views is required. We achieve this by canonicalizing all partial views in a way that the geometry and the

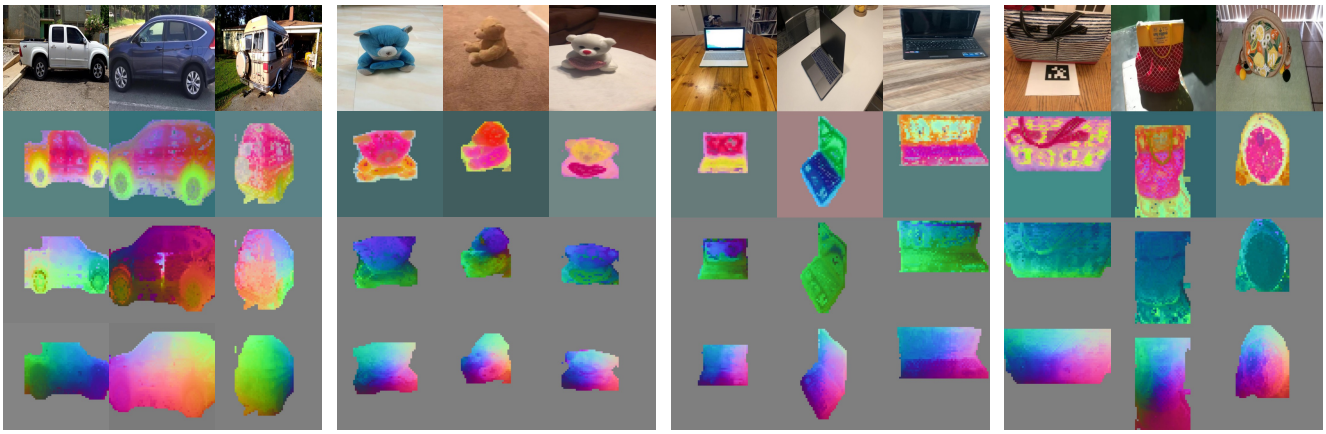

Figure 3. **Feature comparison across seen and unseen categories.** Comparison of DINOv2, Spherical Mapper, and Generalizable Correspondence features on *car*, *teddybear*, *laptop*, and *handbag*. The *car* category is seen by the Spherical Mapper, while the others are unseen; all categories are unseen for the Generalizable Correspondence Feature. DINOv2 shows symmetric ambiguities across viewpoints. The Spherical Mapper distinguishes left/right for seen categories but fails to generalize to unseen ones, producing noisier predictions. In contrast, the Generalizable Correspondence Feature yields robust, generalizable 3D-aware predictions across all categories.

appearance of overlapping partial views match. In total we have K partial views $\mathcal{S} = \{\mathrm{S}_i\}_{i=1}^{K}$ which we can transform into the canonical space with the corresponding transformations $\mathcal{T} = \{\mathrm{T}_i = (R_i \in \mathbb{R}^{3\times3}, t_i \in \mathbb{R}^3) : \mathbb{R}^3 \mapsto \mathbb{R}^3\}_{i=1}^{K}$, as follows:

$$\mathrm{T}(\mathrm{S}) = \{\{\mathrm{T}(\mathrm{v}_i)\}_{i=1}^{|\mathrm{V}|}, \mathrm{F}\}. \tag{1}$$

For global consistency, we introduce the unknown graph $\mathrm{G} = (\mathrm{N}, \mathrm{E})$ of overlapping partial views, resulting in the following optimization problem:

$$\underset{\mathcal{T},\mathrm{E}}{\operatorname{argmin}} \sum_{(i,j)\in\mathrm{E}} \mathcal{D}(\mathrm{T}_i(\mathrm{S}_i), \mathrm{T}_j(\mathrm{S}_j)). \tag{2}$$

Here, $\mathcal{D}(\mathrm{T}_i(\mathrm{S}_i), \mathrm{T}_j(\mathrm{S}_j))$ denotes the geometric and appearance distance between two canonicalized partial views. Note, to avoid the trivial solution of an empty graph, we enforce the graph to be connected. As more edges may only increase the cost, this means for the number of edges $|\mathrm{E}| = |\mathrm{N}|-1$. Hence, the graph we are searching is actually a tree, and the optimal relative transformation between two neighboring nodes is independent of the others, resulting in

$$\underset{\mathrm{E}}{\operatorname{argmin}} \sum_{(i,j)\in\mathrm{E}} \underset{{}^{i}\mathrm{T}_j}{\operatorname{argmin}}\, \mathcal{D}(\mathrm{S}_i, {}^{i}\mathrm{T}_j(\mathrm{S}_j)), \tag{3}$$

with

$${}^{i}\mathrm{T}_j = \mathrm{T}_i^{-1}\mathrm{T}_j. \tag{4}$$

This independence enables us to first efficiently estimate ${}^{i}\mathrm{T}_j$ and $\mathcal{D}(\mathrm{S}_i, {}^{i}\mathrm{T}_j(\mathrm{S}_j))$ by building on the RANSAC-based[15] method, as proposed by Sommer et al. [43], and then optimize for the optimal tree graph. Otherwise, estimating all $\mathcal{T}$, would become too exhaustive, as the number of proposals would grow exponentially with the number of partial views. We find the optimal tree graph G with edges E using a greedy search algorithm. Therefore, we start with the two partial views whose pairwise overlap yields the minimum distance. Then, we iteratively add the partial view with the minimum distance to one of the previously selected partial views. The method is illustrated in Fig. 2.
Regarding the pairwise geometric and appearance distance, we built upon [43], and define the distance for two overlapping partial views S_i and S_j as follows

$$\mathcal{D}(\mathrm{S}_i, \mathrm{S}_j) = (1-\alpha)\mathcal{D}_{\mathrm{geo}}(\mathrm{S}_i, \mathrm{S}_j) + \alpha\mathcal{D}_{\mathrm{app}}(\mathrm{S}_i, \mathrm{S}_j), \tag{5}$$

where the geometric distance $\mathcal{D}_{\mathrm{geo}}$ equals a weighted chamfer distance for the vertices. And similarly, the appearance distance $\mathcal{D}_{\mathrm{app}}$ is measuring the distance between vertices in euclidean space, but the nearest neighbors ψ are found in feature space and not in euclidean space, as follows

$$\psi(i) = \underset{j\in1...|\mathrm{V}_j|}{\operatorname{argmin}} \sum_{l} \min_{k} ||\mathrm{f}_j^k - \mathrm{f}_i^l||\,, \quad \text{for } i \in 1...|\mathrm{V}_i|. \tag{6}$$

Each vertex-based distance is weighted with respect to its feature correspondence cyclic consistency. Therefore, semantic parts which are only captured in one partial view, are wrongly matched, resulting in a low cycle consistency and are thus down-weighted.

3.3. Generalizable Correspondence Features

The large-scale training of foundational models, such as DINOv2 [30], lead to the emergence of features that capture semantic concepts, which also allows finding semantic correspondences between two images. However, a limitation of the direct application of such foundational features

Figure 4. **Visualization of the Generalizable Correspondence Feature across categories.** Visualization of the Generalizable Correspondence Feature for the categories *bicycle*, *toy truck*, *motorcycle*, *toy bus*, *toy plane*, and *chair*. These features show that instances captured from similar viewpoints are closely grouped in the feature space, demonstrating effective canonicalization across different categories.

arises from the feature ambiguities of object parts with similar appearance. This sometimes results in the inability to distinguish between object symmetries, i.e. left-right ambiguity [57], or to differentiate between identical semantic parts located at different positions. This missing capability is, however, critical for our task, since we aim at canonicalizing 3D objects from partial views. Mariotti et al. [27] address the limitations for solving semantic correspondences by proposing a spherical mapper, i.e., a neural network that maps DINOv2 features $\mathcal{F}$ to points $s \in \mathcal{S}^2$ on a canonicalized spherical coordinate system: $s = f_s(\mathcal{F})$. This approach conceptually addresses our need for resolving features ambiguities of DINOv2 features. For example, the feature of a patch from the left side of a car is mapped to the opposite hemisphere than the features of a patch from the right side of a car. However, we observe that the original work, trained on only the 18 categories of the SPair-71K dataset [28], does not generalize to novel categories. Therefore, we introduce a generalizable approach that avoids the requirement of knowing beforehand which object categories shall be matched at test time. Specifically, we train a generalizable correspondence feature extractor on large scale 3D pose annotations from ImageNet3D [26]. To achieve this, we quantize the given azimuth angles to 8 bins and acquire the object masks using SAM [21]. To ensure a generalizable setting, we carefully remove all categories for training that overlap with our evaluation, eventually applying the scaled spherical mapper in an out-of-distribution scenario. Qualitatively, we show in Fig. 3, that our feature extractor generalizes to novel categories. Further, quantitative experiments confirm that our approach leads to a better generalizability in semantic correspondence, hence largely facilitating the canonicalization of 3D objects from partial views.

After training the correspondence feature extractor, we follow prior work [27], and concatenate the trained features with pretrained DINOv2 features. This combination preserves the rich semantic information from DINOv2 while incorporating crucial viewpoint information from the generalizable correspondence feature.

4. Experiments

In this section, we introduce the experimental setup in 4.1, how we train the generalizable correspondence features in Section 4.2, present how we test our methods in both synthetic and real-world datasets in Section 4.3. In the end, we show the ablation of key components in our method in 4.4.

4.1. Experimental setup

Dataset for training generalizable correspondence features. We train the spherical mapper on a large-scale dataset of ImageNet3D[26] that consists of more than 86 000 images of 200 rigid categories that are annotated with 3D pose label. In order to show the generalizability of our new features, we remove all the CO3Dv2 and ShapeNet dataset categories in the training. We test our feature on unseen categories from CO3Dv2 and ShapeNetv2 to show the better generalization compared to the spherical mapper.

Dataset for Evaluation of Canonicalization Results from Partial Views. We test our method and baseline on the real-world CO3Dv2 dataset[35] and the synthetic ShapeNet dataset[4]. All tested categories from CO3Dv2 and ShapeNet dataset were unseen during the training of the spherical mapper.

The CO3Dv2 dataset[35] includes object-centric full videos across 50 diverse categories. We focused our testing on the same 20 benchmark categories that were previously examined by the baseline UOP3D [43] within the CO3Dv2 dataset. Each category features maximum five annotated

canonicalized sequences used as ground truth labels. We explored various partial video settings with frame ratios of 10% , 25%, 50%, starting always from the first frame of each sequence. Using COLMAP[39], we generated partial scans from these partial video sequences.

We use ShapeNet [4], which provides ground-truth canonicalization labels, to evaluate 10 categories. This extends the CO3Dv2 categories by four classes: table, bench, display, and telephone. For each category, we render 30 objects at partial fractions of 10%, 25%, and 50%. The camera always points to the object origin with $\theta = 0$, a random elevation in $[0°, 30°]$, and an azimuth that starts from a random angle and increases to cover the desired partial fraction.

Baseline. Our baselines are UOP3D [43], which introduces a canonicalization method for object-centric full-view video sequences, and OrientAnything [51], a foundation model for object orientation estimation in single-view images, trained on 2M synthetically rendered images. UOP3D utilizes the DINOv2 feature, back-projecting it onto the entire mesh. We evaluate our method against UOP3D across various settings, including partial views of 10%, 25% and 50%. We use the base version of DINOv2 for both the UOP3D and our method. The feature dimension of DINOv2 feature is 768. For both UOP3D and our method, we select $\alpha = 0.2$ in distance formulation 5. Since OrientAnything [51] predicts both orientation and confidence for each frame, we compute the canonical orientation for the whole partial video sequence by averaging orientations weighted by their confidence scores.

Evaluation Metrics. Our evaluation follows the standard benchmark settings used by UOP3D[43]. We compute the rotational error between each pair in the test dataset. The evaluation metric is determined by the proportion of pairs where the rotational error is less than $30°$.

Training Details for generalizable correspondence features. Current SOTA viewpoint-dependent features trained on the Spair-71K dataset cannot generalize well on the unseen categories from the training set. To address this shortcoming, we train the model on ImageNet3D[26] following the architecture and training scheme of [27], keeping the base variant of DINOv2 frozen as backbone.

4.2. Generalizable Correspondence Features

Quantitative Analysis. The original spherical mapper is trained on SPair-71K dataset, which has overlap categories with CO3Dv2 dataset. For evaluation, we pick unseen categories of SPair-71K training categories from CO3D categories. Those categories will be both unseen for the original spherical mapper and our adapted version, the Generalizable Correspondence Features. As shown in Table 1, the model trained on a larger data scale performs better than the original Spherical Mapper on unseen categories.

Qualitative Analysis. The effectiveness of the Generalizable Correspondence Features is demonstrated through the visualization results of unseen categories from training dataset presented in Figure 4. This figure highlights several key insights. Firstly, the features vary significantly with changes in viewpoint, indicating that distant viewpoints yield distinctly different feature representations. In addition, instances from the same category exhibit similar features when observed from similar viewpoints, confirming that the feature successfully canonicalizes these instances. Lastly, the proposed features generalize across different categories, as similar viewpoints across various object types lead to similar feature spaces.

Category	Acc@30° (SPH)	Acc@30° (GCF)
Teddybear	33.3	**100.0**
Toaster	**30.0**	**30.0**
Handbag	60.0	**90.0**
Laptop	10.0	**100.0**
Keyboard	**60.0**	**60.0**
Hairdryer	0.0	**60.0**
Toilet	**100.0**	**100.0**
Toytruck	30.0	**40.0**
Mouse	**10.0**	**10.0**
Mean	37.03	**65.56**

Table 1. $30°$ **accuracy on unseen object categories.** Comparison of $30°$ accuracy on unseen categories between the original Spherical Mapper Features (SPF) and Generalizable Correspondence Feature (GCF).

4.3. Canonicalization from Partial Views

Quantitative Analysis. We report experimental results on the CO3Dv2 dataset (Table 2) and the ShapeNet dataset (Table 3). Across both real-world and synthetic settings, our method consistently outperforms baselines at all tested ratios. As OrientAnything is a foundation model developed for frame-wise canonical orientation estimation, its predictions across video frames can be subject to temporal inconsistencies. By contrast, our approach naturally integrates multi-view information and enforces globally consistent alignment across 3D shapes reconstructed from partial videos, which results in superior performance with substantially lower training cost. The difference in dataset characteristics is also notable: CO3Dv2 provides at most five annotated sequences per category, whereas ShapeNet offers thirty. The greater number of sequences in ShapeNet reduces variance in performance, likely due to a more uniform distribution of partial views in the canonical space.

Qualitative Analysis. We compare the canonicalization results of UOP3D baseline and our method in Figure 5. In UOP3D, the first scan on the left serves as the reference; all other source scans attempt to align to this reference. For the

												Avg (20)
10%	UOP3D	**100.0**	**33.3**	50.0	**20.0**	**60.0**	**100.0**	50.0	50.0	**33.3**	**40.0**	42.9
	OrientAnything	40.0	0.0	60.0	10.0	50.0	83.3	10.0	**100.0**	**33.3**	20.0	42.2
	Ours	**100.0**	16.7	**100.0**	**20.0**	**60.0**	**100.0**	**60.0**	83.3	**33.3**	**40.0**	**47.0**
25%	UOP3D	75.0	20.0	50.0	**25.0**	**60.0**	**100.0**	**60.0**	83.3	**100.0**	**20.0**	48.3
	OrientAnything	40.0	0.0	90.0	10.0	40.0	**100.0**	20.0	**100.0**	**100.0**	**20.0**	55.5
	Ours	**90.0**	**40.0**	**100.0**	20.0	**60.0**	**100.0**	50.0	**100.0**	**100.0**	**20.0**	**57.3**
50%	UOP3D	**50.0**	45.0	85.0	30.0	55.0	**100.0**	**100.0**	83.0	**100.0**	65.0	68.8
	OrientAnything	40.0	0.0	**100.0**	10.0	40.0	**100.0**	40.0	**100.0**	**100.0**	20.0	60.0
	Ours	40.0	**60.0**	**100.0**	**40.0**	**60.0**	**100.0**	**100.0**	**100.0**	**100.0**	**100.0**	**71.7**

Table 2. **Comparison under varying supervision ratios on CO3Dv2 dataset.** Comparison of baselines (UOP3D, OrientAnything) and our method (Ours) at 10%, 25%, and 50% supervision ratios on selected CO3Dv2 categories. The evaluation metric is 30° accuracy. Results are averaged over all 20 categories, while only 10 categories are shown.

												Avg (10)
10%	UOP3D	39.1	25.7	33.7	42.3	41.0	32.1	44.1	33.0	72.5	33.0	39.7
	Ours	**51.5**	**32.0**	**43.9**	**55.9**	**46.4**	**40.9**	**45.5**	**52.2**	**92.4**	**40.0**	**50.1**
25%	UOP3D	46.7	31.1	**33.3**	45.1	42.1	35.9	49.2	38.6	84.1	42.9	44.9
	Ours	**87.1**	**35.4**	25.7	**48.3**	**52.2**	**42.1**	**59.5**	**48.0**	**100.0**	**57.9**	**55.6**
50%	UOP3D	85.5	30.3	44.5	54.9	**51.1**	39.8	**64.9**	41.0	88.4	50.0	55.0
	Ours	**93.3**	**41.1**	**58.4**	**66.9**	46.4	**40.0**	54.0	**52.0**	**93.3**	**57.9**	**60.3**

Table 3. **Comparison under varying supervision ratios on ShapeNet dataset.** Comparison of the baseline (UOP3D) and our method (Ours) at 10%, 25%, and 50% supervision ratios on the synthetic ShapeNet dataset. The evaluation metric is 30° accuracy.

	ShapeNet			CO3Dv2		
	10%	25%	50%	10%	25%	50%
Ours	**50.1**	**55.6**	**60.3**	**47.0**	**57.3**	71.7
w/o GCF	43.5	47.4	57.9	46.0	53.9	**75.7**
w/o GO	45.0	50.9	59.3	44.4	50.3	69.4
w/o GCF & GO	39.7	44.9	55.0	42.9	48.3	68.8

Table 4. **Ablation study on GCF and global optimization.** Ablation study of the Generalizable Correspondence Feature (GCF) and Global Optimization (GO) on the ShapeNet and CO3Dv2 datasets. The evaluation metric is 30° accuracy.

laptop category, the pairwise alignment is problematic due to the suboptimal choice of a noisy and partial reference mesh, leading to poor canonicalization outcomes for subsequent scans. Conversely, our method initiates canonicalization with two clean and comprehensive scans from the selection of lowest distance between the partial scans, allowing subsequent partial scans to align more effectively into the canonical space. Furthermore, Figure 6 showcases the canonicalization results for a collection of partial scans, including both annotated and unannotated sequences to create a more comprehensive dataset. Within each category, the sequence order from left to right also delineates the canonicalization process.

4.4. Ablation Studies

Effect of Generalizable Correspondence Features. The Generalizable Correspondence Features significantly enhance the performance, as demonstrated in Table 4. By incorporating viewpoint-aware elements, these features improve the accuracy of finding matches. This improvement helps to effectively mitigate the challenges of invalid correspondences, often resulting from the feature ambiguities of repeated object parts that are observed with DINOv2. Furthermore, the GCFs enable more precise weighting during the evaluation of pairwise registration quality, as defined in Eq. (5), which is crucial for global optimization.

Effect of Global Optimization. Table 4 highlights the benefits of global optimization, which ensures precise pairwise pose registration across a collection of partial scans, demonstrating an improved global pose accuracy.

5. Conclusion

In this work, we introduced C3PO, a novel framework for category-level canonicalization of 3D objects from partial

Figure 5. **Qualitative comparison of canonicalization results.** Visualization comparing the UOP3D baseline and our method for the *laptop* (10%) and *hydrant* (50%) categories. For UOP3D, the leftmost scan serves as the reference to which the remaining scans are aligned. For our method, scans are processed sequentially from left to right, reflecting the canonicalization order.

Figure 6. **Canonicalization of partial scans across categories.** Visualization of canonicalized partial scans for the categories *car* (25%), *bicycle* (10%), and *backpack* (10%). Both annotated and unannotated video sequences are integrated into a unified collection of partial sequences. Scans are processed sequentially from left to right, indicating the canonicalization order.

views. Unlike prior approaches that rely on full object coverage or fixed references, our method jointly canonicalizes multiple partial reconstructions by enforcing both geometric and appearance consistency across overlapping regions. Central to this is are Generalizable Correspondence Features (GCF) that capture viewpoint-aware correspondences from large-scale data and generalize effectively to unseen categories. Together with our global-consistency optimization via pairwise registration, C3PO achieves robust canonicalization even under severe partial-view settings.

We established two new benchmarks on synthetic (ShapeNet) and real-world (CO3Dv2) datasets to rigorously evaluate canonicalization from partial observations. Across both, C3PO consistently and significantly outperformed state-of-the-art baselines, with particularly strong gains in challenging low-coverage scenarios. By enabling scalable and accurate canonicalization without manual alignment or full geometry, C3PO opens the door to applying canonicalization at scale, including to long-tail object categories and in-the-wild video data.

Acknowledgements

Adam Kortylewski gratefully acknowledges support for his Emmy Noether Research Group, funded by the German Research Foundation (DFG) under Grant No. 468670075.

References

[1] Souhaib Attaiki, Gautam Pai, and Maks Ovsjanikov. Dpfm: Deep partial functional maps. In *2021 International Conference on 3D Vision (3DV)*, pages 175–185. IEEE, 2021. 2

[2] David Bensaïd, Noam Rotstein, Nelson Goldenstein, and Ron Kimmel. Partial matching of nonrigid shapes by learning piecewise smooth functions. In *Computer Graphics Forum*, page e14913, 2023. 2

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 2, 3

[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. 2, 5, 6

[5] Haiwei Chen, Shichen Liu, Weikai Chen, Hao Li, and Randall Hill. Equivariant point network for 3d point cloud analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14514–14523, 2021. 2

[6] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *European conference on computer vision*, pages 628–644. Springer, 2016. 1

[7] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021. 2

[8] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppf-foldnet: Unsupervised learning of rotation invariant 3d local descriptors. In *Proceedings of the European conference on computer vision (ECCV)*, pages 602–618, 2018. 2

[9] Yu Deng, Jiaolong Yang, and Xin Tong. Deformed implicit field: Modeling 3d shapes with learned dense correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10286–10296, 2021. 1

[10] 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. 1

[11] 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. 1

[12] Olaf Dünkel, Thomas Wimmer, Christian Theobalt, Christian Rupprecht, and Adam Kortylewski. Do it yourself: Learning semantic correspondence from pseudo-labels. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5834–5844, 2025. 3

[13] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 29(4):551–559, 1983. 3

[14] Viktoria Ehm, Maolin Gao, Paul Roetzer, Marvin Eisenberger, Daniel Cremers, and Florian Bernard. Partial-to-partial shape matching with geometric consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27488–27497, 2024. 2

[15] Martin A. Fischler and Robert C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981. 4

[16] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se (3)-transformers: 3d roto-translation equivariant attention networks. *Advances in neural information processing systems*, 33:1970–1981, 2020. 2

[17] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Zero-shot category-level object pose estimation. In *European Conference on Computer Vision*, pages 516–532. Springer, 2022. 2

[18] Walter Goodwin, Ioannis Havoutis, and Ingmar Posner. You only look at one: Category-level object representations for pose estimation from a single example, 2023. 3

[19] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18995–19012, 2022. 2

[20] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 216–224, 2018. 1

[21] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 5

[22] 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. 2

[23] Zizhang Li, Dor Litvak, Ruining Li, Yunzhi Zhang, Tomas Jakab, Christian Rupprecht, Shangzhe Wu, Andrea Vedaldi, and Jiajun Wu. Learning the 3d fauna of the web. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9752–9762, 2024. 2

[24] Or Litany, Emanuele Rodolà, Alexander M Bronstein, and Michael M Bronstein. Fully spectral partial shape matching. In *Computer Graphics Forum*, pages 247–258. Wiley Online Library, 2017. 2

[25] Qihao Liu, Yi Zhang, Song Bai, Adam Kortylewski, and Alan Yuille. Direct-3d: Learning direct text-to-3d generation

on massive noisy 3d data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6881–6891, 2024. 1

[26] Wufei Ma, Guofeng Zhang, Qihao Liu, Guanning Zeng, Adam Kortylewski, Yaoyao Liu, and Alan Yuille. Imagenet3d: Towards general-purpose object-level 3d understanding. *Advances in Neural Information Processing Systems*, 38, 2024. 2, 5, 6

[27] Octave Mariotti, Oisin Mac Aodha, and Hakan Bilen. Improving semantic correspondence with viewpoint-guided spherical maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19521–19530, 2024. 3, 5, 6

[28] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence, 2019. 3, 5

[29] Natalia Neverova, David Novotny, Marc Szafraniec, Vasil Khalidov, Patrick Labatut, and Andrea Vedaldi. Continuous surface embeddings. *Advances in Neural Information Processing Systems*, 33:17258–17270, 2020. 3

[30] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1

[32] Adrien Poulenard, Marie-Julie Rakotosaona, Yann Ponty, and Maks Ovsjanikov. Effective rotation-invariant point cnn with spherical harmonics kernels. In *2019 International Conference on 3D Vision (3DV)*, pages 47–56. IEEE, 2019. 2

[33] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017.

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2

[35] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2, 5

[36] Paul Roetzer, Paul Swoboda, Daniel Cremers, and Florian Bernard. A scalable combinatorial solver for elastic geometrically consistent 3d shape matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 428–438, 2022. 2

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[38] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[39] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 501–518. Springer, 2016. 3, 6

[40] Nima Sedaghat and Thomas Brox. Unsupervised generation of a viewpoint annotated car dataset from videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1314–1322, 2015. 2

[41] Nima Sedaghat, Mohammadreza Zolfaghari, Ehsan Amiri, and Thomas Brox. Orientation-boosted voxel nets for 3d object recognition. *arXiv preprint arXiv:1604.03351*, 2016. 1

[42] Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. Shic: Shape-image correspondences with no keypoint supervision. In *ECCV*, 2024. 3

[43] 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 (CVPR)*, pages 22787–22796, 2024. 2, 3, 4, 5, 6

[44] Riccardo Spezialetti, Federico Stella, Marlon Marcon, Luciano Silva, Samuele Salti, and Luigi Di Stefano. Learning to orient surfaces by self-supervised spherical cnns. *Advances in Neural information processing systems*, 33:5381–5392, 2020. 2

[45] Shanlin Sun, Kun Han, Deying Kong, Hao Tang, Xiangyi Yan, and Xiaohui Xie. Topology-preserving shape reconstruction and registration via neural diffeomorphic flow. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20845–20855, 2022. 1

[46] 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. 2

[47] Hugues Thomas, Charles R Qi, Jean-Emmanuel Deschaud, Beatriz Marcotegui, François Goulette, and Leonidas J Guibas. Kpconv: Flexible and deformable convolution for point clouds. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6411–6420, 2019. 2

[48] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. *arXiv preprint arXiv:1802.08219*, 2018. 2

[49] Angtian Wang, Adam Kortylewski, and Alan Yuille. Nemo:

Neural mesh models of contrastive features for robust 3d pose estimation. *arXiv preprint arXiv:2101.12378*, 2021. 3

[50] Angtian Wang, Peng Wang, Jian Sun, Adam Kortylewski, and Alan Yuille. Voge: a differentiable volume renderer using gaussian ellipsoids for analysis-by-synthesis. In *The Eleventh International Conference on Learning Representations*, 2022. 1

[51] 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. 6

[52] Shangzhe Wu, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Dove: Learning deformable 3d objects by watching videos. *International Journal of Computer Vision*, pages 1–12, 2023. 2

[53] Shangzhe Wu, Ruining Li, Tomas Jakab, Christian Rupprecht, and Andrea Vedaldi. Magicpony: Learning articulated 3d animals in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8792–8802, 2023. 2

[54] Yang Xiao, Yuming Du, and Renaud Marlet. Posecontrast: Class-agnostic object viewpoint estimation in the wild with pose-aware contrastive learning. In *2021 International Conference on 3D Vision (3DV)*, pages 74–84. IEEE, 2021. 1

[55] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9150–9161, 2023. 2

[56] Yanjie Ze and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *Advances in Neural Information Processing Systems*, 35:27469–27483, 2022. 3

[57] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3076–3085, 2024. 3, 5

[58] Kaifeng Zhang, Yang Fu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang. Self-supervised geometric correspondence for category-level 6d object pose estimation in the wild. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[59] 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. 1

[60] Yongheng Zhao, Tolga Birdal, Jan Eric Lenssen, Emanuele Menegatti, Leonidas Guibas, and Federico Tombari. Quaternion equivariant capsule networks for 3d point clouds. In *European conference on computer vision*, pages 1–19. Springer, 2020. 2

[61] Zerong Zheng, Tao Yu, Qionghai Dai, and Yebin Liu. Deep implicit templates for 3d shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1429–1439, 2021. 1

[62] Xingyi Zhou, Arjun Karpur, Linjie Luo, and Qixing Huang. Starmap for category-agnostic keypoint and viewpoint estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 318–334, 2018. 1

[63] Minghan Zhu, Maani Ghaffari, William A Clark, and Huei Peng. E2pn: Efficient se (3)-equivariant point network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1223–1232, 2023. 2