EVERY9D-22M: LARGE-SCALE REAL-WORLD 9D CANONICALIZATION OF EVERYDAY OBJECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Estimating the 9D pose of everyday objects from a single real-world image remains challenging. This is largely due to the lack of large-scale supervision. Most existing datasets either rely heavily on synthetic renderings or provide limited coverage of real-world objects. We address this gap with Every9D-22M, a dataset of 9D pose annotations for 22.8M real-world images from 114K object-centric videos spanning 630 everyday object categories, making it nearly two orders of magnitude larger than any prior real-world 9D pose dataset in at least one of category, object, or image count. To achieve this scale, we leverage object-centric videos by reconstructing object-level point clouds via multi-view geometry and aligning similar instances into a shared canonical coordinate frame. Canonical poses are manually annotated for only a small set of reference objects (fewer than 0.01% of all images) and propagated to the remaining instances via cross-instance alignment. All propagated canonical poses are then verified from multiple viewpoints. We further introduce cross-category orientation rules that induce category-level symmetries, enabling symmetry-aware evaluation. Beyond establishing dedicated training and evaluation splits as a benchmark for 9D pose foundation models, we show that training on Every9D-22M generalizes better to other datasets than training on existing datasets, and outperforms prior foundation models in rotation accuracy on all benchmarks except the viewpoint-biased ImageNet3D, where joint training with ImageNet3D surpasses them as well. Data and code will be made publicly available.



Figure 1: **Every9D-22M.** Example canonicalized objects from our dataset. Each object is reconstructed as a 3D Gaussian Splat from an object-centric video in uCO3D. Cross-instance alignment establishes shared canonical coordinate frames, which are subsequently propagated to all video frames, yielding 9D pose annotations for 22.8M real-world images across 630 object categories.

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

1 INTRODUCTION

Single-image 9D object pose estimation, comprising 3D rotation, 3D translation, and 3D object extent, is fundamental for robotics and AR/VR applications, as it enables spatial reasoning and physical interaction from a single RGB image. It requires models to infer consistent object-centric coordinate frames across diverse categories and instances. While recent methods have made progress in category-level 3D pose estimation (Stark et al., 2010; Zhou et al., 2018; Xiao et al., 2021; Jesslen et al., 2024) and full 6D pose estimation (Wang et al., 2019b; Chen & Dou, 2021; Lin et al., 2022; Liu et al., 2023a; Zheng et al., 2023; Zhang et al., 2024a), reliable 9D pose estimation for general everyday objects remains an open challenge under unconstrained real-world conditions. Consistent object canonicalization also matters beyond robotics: text-to-3D and image-to-3D systems benefit from orientation-aligned training data, as canonical object frames improve controllability and geometric consistency (Liu et al., 2024; Zhang et al., 2024b; Lu et al., 2025), and correspondence learning and 3D morphable modeling commonly assume canonicalized objects to ensure stable alignment across instances (Mariotti et al., 2024; Dünkel et al., 2025; Choy et al., 2016; Groueix et al., 2018; Park et al., 2019; Deng et al., 2021; Zheng et al., 2021; Sun et al., 2022; Sommer et al., 2025). Scalable real-world canonicalized supervision beyond synthetic data is therefore needed.

However, most existing large-scale 3D datasets rely on rendered assets (Deitke et al., 2023b;a; Zhang et al., 2025b; Liu et al., 2023b; Collins et al., 2022; Khanna et al., 2024; Fu et al., 2021), which enable scale but are created under synthetic and controlled conditions. Models trained predominantly on such renderings, including the recent OrientAnything models for pose estimation (Wang et al., 2024b; 2026), struggle to generalize to real images affected by sensor noise, occlusions, motion blur, and appearance variability. To mitigate this sim-to-real gap, OrientAnything resorts to ImageNet3D (Ma et al., 2024), yet image-level 3D pose annotations are prohibitively expensive and fundamentally ambiguous from single views. Existing real-world pose datasets are thus forced to trade off category diversity against instance scale: those covering many categories provide only tens of thousands of annotated images, while those reaching millions of images cover only a handful of categories.

To overcome this limitation, we leverage object-centric videos as a scalable source of real-world supervision. For each video, we reconstruct an object-level point cloud via multi-view reconstruction, yielding a geometric representation robust to single-view ambiguities. We then cluster visually similar instances and align their point clouds through cross-instance geometric and appearance matching, establishing shared canonical coordinate frames that are propagated from a small set of manually annotated reference objects. Applying this framework to 114K object-centric videos of everyday objects from uCO3D (Liu et al., 2025) yields Every9D-22M, with 9D pose annotations for 22.8M real-world images across 630 categories. Manual annotation covers fewer than 0.01% of images, while every propagated pose is verified from multiple viewpoints, amounting to 199 hours of human effort in total: 67 hours of reference annotation and 132 hours of verification.

Our main contributions are the following:

1. **Scalable real-world 9D canonicalization.** We introduce an iterative reference-based canonicalization framework that leverages object-centric videos for multi-view 3D reconstruction and cross-instance alignment. By propagating canonical frames from a small set of annotated reference objects, our approach enables 9D pose supervision at scale, with every propagated annotation subsequently verified from multiple viewpoints.

2. **Every9D-22M: A large-scale real-world 9D pose dataset.** The resulting dataset covers 22.8M images across 630 categories, nearly two orders of magnitude more than any prior real-world 9D pose dataset in at least one of category, object, or image count. Fewer than 0.01% of images require manual annotation.

3. **Diverse large-scale benchmark for 9D pose foundation models.** We establish standardized symmetry-aware evaluation protocols and baseline models. Training on Every9D-22M generalizes better to other datasets than training on existing datasets and outperforms prior foundation models in rotation accuracy on all benchmarks except the viewpoint-biased ImageNet3D, where joint training with ImageNet3D surpasses them as well.

2 Related Work

We review prior work on 9D object pose annotation from real-world imagery, organized by the level of annotation effort, and compare real-world 9D pose datasets quantitatively in Table 1.

2.1 Image-Level Pose Annotations

Obtaining pose annotations at the image level is the most labor-intensive paradigm. Prior works (Xiang et al., 2014; 2016; Ma et al., 2024) select an exemplary CAD mesh for a given category and manually align it to each image by adjusting its pose parameters. This is inherently sensitive to discrepancies between the template mesh and the true object: the projected alignment can appear accurate in the 2D image plane while remaining geometrically inconsistent in 3D. Given scale ambiguity and the absence of real-world camera intrinsics, these approaches evaluate only 3D object orientations. ImageNet3D (Ma et al., 2024) is the largest such benchmark with approximately 86K images, whereas we scale to 22.8M annotated images with only around 1K manual pose annotations.

WildDet3D (Huang et al., 2026) instead annotates 3D bounding boxes at the image level, with orientation defined primarily by the gravity direction, whereas we leverage semantically grounded cross-category intrinsic and interactive orientation rules, detailed in the supplementary material.

2.2 Video-Level Pose Annotations

Video-level annotation reduces ambiguity by exploiting multi-view consistency across frames. Where each video contains a single object instance, human effort scales with the number of videos, as in Objectron (Ahmadyan et al., 2021), which annotates poses manually per video, and in Wild6D (Fu & Wang, 2022) and HANDAL (Guo et al., 2023), which propagate keyframe and tracker initializations and refine them manually. With multiple objects per scene and instances reappearing across videos, the effort scales with both quantities: NOCS-REAL275 (Wang et al., 2019b) relies entirely on manual labeling, while PhoCaL (Wang et al., 2022) and HouseCat6D (Jung et al., 2024) match previously scanned objects to scene observations using robot-arm keypoint measurements.

Labeling each instance within a sequence nevertheless remains labor-intensive, limiting scale. Objectron (Ahmadyan et al., 2021) is the largest dataset by number of annotated objects, comprising approximately 17K instances, but it covers only 9 categories, nearly two orders of magnitude fewer than our 630. Omni6DPose-ROPE (Zhang et al., 2024a), the second largest by object instances, contains only 581 objects, again nearly two orders of magnitude fewer than our 114K.

uCO3D (Liu et al., 2025) normalizes object-centric reconstructions at the scene level by enforcing a gravity-aligned vertical axis, normalizing scale and translation, and resolving horizontal orientation using PCA. This removes the arbitrary coordinate frames of structure-from-motion pipelines, but does not guarantee semantic object orientation: the vertical scene axis need not match the semantic top–

Table 1: Comparison of real-world model-free pose datasets. ⋆Note that depth estimates in our dataset are obtained using a monocular depth prediction method (Yang et al., 2024) and are subsequently aligned with SfM. For Every9D-22M, we count only categories given five test instances.

Dataset	In-/Outdoor	Modality	Eval.	#Categories	#Objects	#Videos	#Images	#Manual
PASCAL3D+ (Xiang et al., 2014)	✓	RGB	3D	12	12K	/	12K	12K
ObjectNet3D (Xiang et al., 2016)	✓	RGB	3D	100	57K	/	57K	57K
ImageNet3D (Ma et al., 2024)	✓	RGB	3D	200	86K	/	86K	86K
Objectron (Ahmadyan et al., 2021)	✓	RGB	9D	9	17K	15K	4M	17K
Wild6D (Fu & Wang, 2022)	✓	RGBD	9D	5	162	486	10K	162
HANDAL (Guo et al., 2023)	✓	RGBD	9D	17	212	2K	308K	2K
NOCS-REAL275 (Wang et al., 2019b)	×	RGBD	9D	6	42	18	8K	90
PhoCaL (Wang et al., 2022)	×	RGBD	9D	8	60	24	3.9K	60
HouseCat6D (Jung et al., 2024)	×	RGBD	9D	10	194	41	23.5K	4.8K
Omni6DPose-ROPE (Zhang et al., 2024a)	×	RGBD	9D	149	581	363	332K	1.8K
Every9D-22M (Ours)	✓	RGBD⋆	9D	**630**	**114K**	**114K**	**22.8M**	1K

bottom direction of the object, and PCA-based yaw alignment is not semantic either. In contrast, we canonicalize objects semantically, according to meaningful orientation rules shared across instances.

2.3 Reference-Level Pose Annotations

Methods exploiting automatic cross-instance alignment remain limited (Sommer et al., 2024; Chi et al., 2026; Jin et al., 2025). The first method aligning instances to a single reference (Sommer et al., 2024) is limited to 1K videos across 23 categories, does not filter incorrect poses after alignment, and requires category information to select a reference. C3PO (Chi et al., 2026) canonicalizes 200 videos across 20 categories using category supervision, extending single-reference canonicalization to partially observed objects through pairwise all-to-all alignment, whose quadratic complexity limits scalability. More recently, Jin et al. (2025) canonicalize 32K filtered Objaverse assets using one annotated reference per category, again assuming category information; their support-plane strategy is questionable for the noisy, partial point clouds obtained from structure-from-motion with imperfect masks, and has not been shown to generalize to real-world videos. Our method requires no category information, allowing cross-instance alignment across categories and reducing human effort.

2.4 Self-Supervised Pose Annotations

MagicPony (Wu et al., 2023) recovers 6D pose and shape self-supervised from real-world data, but targets articulated animals, reports no quantitative pose accuracy, and has not been shown to generalize to rigid everyday objects. Purely geometric approaches (Li et al., 2021; Sajnani et al., 2022) have so far demonstrated strong performance primarily on a limited set of categories, such as airplanes, cars, and bottles, and typically assume complete point cloud observations, with accuracy degrading substantially for partial ones; Jin et al. (2025) further shows that Condor (Sajnani et al., 2022) performs particularly poorly in low-instance regimes. Given the long-tailed distribution of object instances in real-world data and the noise introduced by structure-from-motion reconstruction and imperfect object masks, these limitations become critical in our setting.

3 Method

Our objective is large-scale, high-quality 9D pose annotation for real-world object-centric videos. To scale annotation while keeping orientation consistent across instances, we adopt a reference-based canonicalization strategy: a small set of manually annotated reference objects defines canonical coordinate frames that are propagated automatically to many instances and their video frames through cross-instance and multi-view alignment.

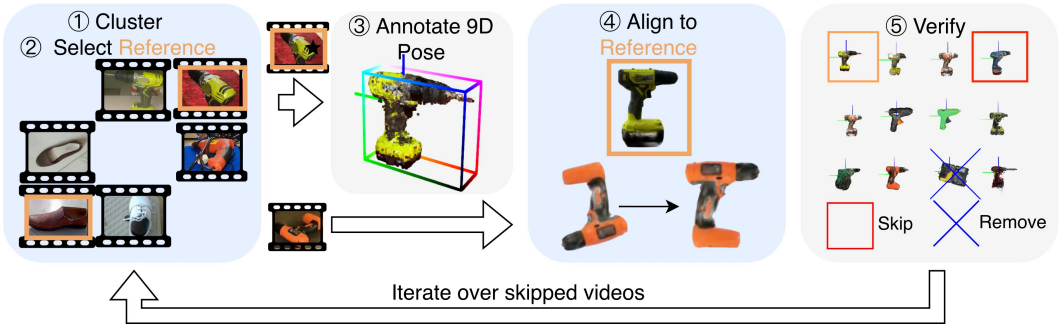


Figure 2: **Canonicalization Framework.** Our framework to canonicalize object-centric videos consists of 5 steps. 1) We cluster all object-centric videos. 2) We select a reference per cluster. 3) We annotate a 9D pose for the reference. 4) We align each object to its reference based upon geometry and appearance. 5) We verify all objects, removing objects with bad reconstruction and skipping objects with wrong pose annotation. Next, we repeat all steps over all skipped videos. Also note that manual labour is only required for step 3) 9D Pose annotation of the reference and 5) verification.

Our framework comprises five stages, illustrated in Figure 2: clustering videos (Section 3.1), selecting a reference per cluster (Section 3.2), annotating its 9D pose following our orientation categories (Section 3.3), aligning the cluster's remaining videos to that reference (Section 3.4), and verifying the propagated annotations (Section 3.5); we then iterate over the videos left over (Section 3.6). We additionally train a 9D pose estimation network as a baseline (Section 3.7).

3.1 Clustering Videos

The inputs to this stage are object-centric videos recorded by non-experts, typically capturing the object from diverse viewpoints, often close to a full 360° around it. They vary substantially in distance, elevation, illumination, background clutter, and motion blur.

To assign videos to clusters sharing a common reference instance, we compute a single embedding per video. Specifically, we extract DINOv3 (Siméoni et al., 2025) features from 32 uniformly sampled frames and average them after normalization, which stays within 1.5 points of using 64 frames on the CO3D benchmark at half the cost. We then apply cosine k-means ($k = 500$) to the normalized embeddings.

Table 2: **Canonicalization design validation** on a small annotated subset of CO3D. We validate (a) the choice of a medoid reference over a random reference and (b) the benefit of multiple alignment iterations. Row 1 is UOP3D (Sommer et al., 2024).

#	Method	Clustering	Ref.	Iter.	GT / iter.	Acc@30° ↑
1	UOP3D	Category	Random	—	10%	82.2%
2	Ours	DINOv3	Random	—	10%	76.1%
3	Ours	DINOv3	Medoid	—	10%	83.0%
4	Ours	DINOv3	Medoid	2	5%	86.5%

We validate clustering on a small CO3D subset (Reizenstein et al., 2021) with annotated 3D poses (Guo et al., 2023), see Table 2. Category-agnostic clustering nearly matches the accuracy of ground-truth category grouping, indicating near category-level cluster quality. Unlike category-based grouping, it lets us tune the number of clusters to control how many references need manual annotation, is robust to incorrect category labels, and exploits cross-category visual similarity. Accordingly, two alignment iterations on half the ground-truth annotations improve performance, even on this subset of 20 categories with 10 similar instances each.

3.2 Reference Selection

Instead of a random reference (Sommer et al., 2024), we choose the cluster medoid, i.e., the video whose image-feature embedding has the smallest average distance to all others in the cluster. This yields a more representative reference and significantly improves accuracy, see Table 2.

3.3 9D Pose Annotation

We first obtain camera poses and a sparse 3D point cloud with VGGSfM (Wang et al., 2024a), and extract it by multi-view masking (Sommer et al., 2024) using text-conditioned LangSAM (Kirillov et al., 2023; Medeiros, 2023) masks stabilized by XMem (Cheng & Schwing, 2022).

Fitting a centered oriented 3D bounding box to the masked point cloud turns each orientation annotation into a 9D pose annotation. For precision, annotators use an interactive tool, placing axes coarsely and refining them in 1° increments; all labels were cross-verified by a second annotator.

To define orientation consistently across instances, we establish cross-category rules (Appendix S4), grouping object categories into orientation categories (Appendix S4.4), based on three principles: (a) *intrinsic orientation*, such as a mounting direction or the primary direction of motion; (b) *human–object interaction*, whether an object faces toward or away from the user and how it is grasped; and (c) *object–object interaction*, reflecting functional usage such as painting or squeezing direction. For example, the canonical front of a chair is opposite the backrest, while for scissors the forward axis aligns with the cutting direction. Orientation categories that leave axes underdetermined additionally induce symmetries, which we derive automatically: a bottle defines only *contain upward*, giving a continuous symmetry about that axis, whereas a tray defines *along grab two hands left right*, an axis rather than a direction, giving a discrete symmetry.

3.4 ALIGNMENT TO REFERENCE

We estimate the rigid transformation aligning a reconstructed instance with its cluster's canonical reference, jointly leveraging geometric structure and visual appearance: each video is represented as a 3D surface endowed with DINO features (Siméoni et al., 2025) projected onto a reconstructed mesh, following (Sommer et al., 2024).

From the masked sparse reconstruction (Section 3.3), we estimate a coarse object surface via alpha shapes (Edelsbrunner & Mücke, 1994) and project multi-view DINO features onto it, giving a surface feature representation $\mathcal{S} = \{\mathcal{V}, \mathcal{F}\}$, with vertices $\mathcal{V} = \{\mathbf{v}_i \in \mathbb{R}^3\}_{i=1}^{|\mathcal{V}|}$ and multi-view features $\mathcal{F} = \left\{ \left\{ \mathbf{f}_i^k \in \mathbb{R}^D \right\}_{k=1}^{|\mathcal{F}_i|} \right\}_{i=1}^{|\mathcal{V}|}$. Here, $\mathbf{f}_i^k$ is the k-th feature observation of vertex i.

We estimate the relative alignment T between two surfaces $\mathcal{S}_i$ and $\mathcal{S}_j$, composed of rotation, translation and scale, using RANSAC initialization followed by gradient-based refinement, as in (Sommer et al., 2024). The alignment minimizes a weighted combination of geometric and appearance distances: $\mathrm{dist}(\mathcal{S}_i, \mathcal{S}_j) = (1-\alpha)\,\mathrm{dist}_{\mathrm{geo}}(\mathcal{S}_i, \mathcal{S}_j) + \alpha\,\mathrm{dist}_{\mathrm{app}}(\mathcal{S}_i, \mathcal{S}_j)$, where $\alpha \in [0, 1]$ controls the trade-off between geometry and appearance. In practice, we set $\alpha = 0.2$.

The geometric term $\mathrm{dist}_{\mathrm{geo}}$ is a weighted Chamfer distance between vertices; the appearance term $\mathrm{dist}_{\mathrm{app}}$ is evaluated in 3D but with nearest neighbors in feature rather than spatial proximity.

Specifically, for each vertex i, we identify the target vertex whose associated feature set yields the smallest aggregated nearest-neighbor feature distance: $\mathbf{v}_{\mathrm{nn}}^{\mathrm{feat}}(i) = \arg\min_{j \in \{1,\ldots,|\mathcal{V}_j|\}} \sum_l \min_k \left\| \mathbf{f}_j^k - \mathbf{f}_i^l \right\|_2$, for $i \in \{1, \ldots, |\mathcal{V}_i|\}$. The bi-directional appearance distance is computed between the corresponding 3D vertex locations, each correspondence weighted by its feature-space cycle consistency (Sommer et al., 2024). Finally, we fit an oriented 3D bounding box to the aligned masked point cloud and propagate the resulting 9D pose to all frames of the video.

3.5 9D POSE VERIFICATION

For each sequence, we visualize the canonicalized reconstruction from three canonical views (front/top/right) and check whether the pose matches the object's semantic axes and the reconstruction is of sufficient quality.

We categorize failures into two groups. **Filtered** ones are permanently removed because the geometry is unreliable, e.g., poor reconstruction (moving objects, noise, missing parts) or masks containing multiple instances. **Skipped** sequences are kept for later processing, as the failure stems from the reference rather than the geometry: overly diverse clusters, fine-grained state changes (e.g., open vs. closed scissors), or semantically different objects grouped together (e.g., a spoon aligned to a pan). As summarized in Table S3, 74.9% of sequences are accepted across both iterations, while 18.1% are filtered and 7.0% remain skipped.

3.6 ITERATIVE PROCESS

We apply the pipeline for two iterations: after each pass, we remove filtered sequences and reprocess the skipped subset by re-clustering, selecting new references, annotating poses, and repeating alignment and verification. Overall, we annotate 1000 reference objects across both iterations.

3.7 MONOCULAR 9D POSE ESTIMATION

To demonstrate the usefulness of Every9D-22M for training monocular pose estimators, we train a simple baseline model for 9D pose prediction. Following prior work (Wang et al., 2019a; Zhang et al., 2024a; 2025a), we fuse geometry and appearance by re-projecting each pixel into 3D using its depth and attaching its patch feature. As point cloud backbone we employ LitePT (Yue et al., 2025) rather than PointNet++ (Qi et al., 2017) or DGCNN (Phan et al., 2018; Yu et al., 2021; Zhang et al., 2025a), combining sparse 3D convolutions (Graham et al., 2018) with point transformer layers (Wu et al., 2024), and encode images with a pretrained DINO backbone (Oquab et al., 2023). The features are concatenated and passed through a two-layer MLP predicting translation, size, and 6D rotation (Zhou et al., 2019), trained with a geodesic rotation loss and ℓ_1 losses on translation and extent.

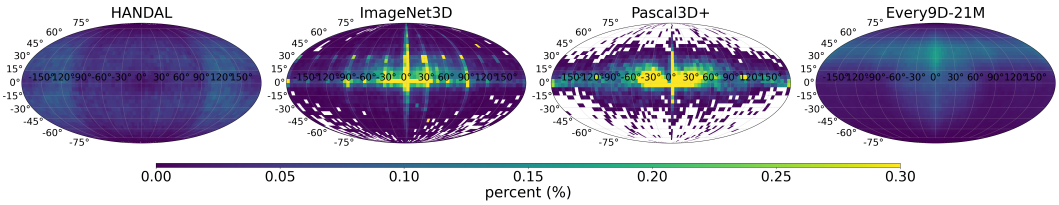


Figure 3: **Mollweide projection of camera-direction histograms in the canonical object frame** Every9D-22M is the only dataset providing broad, near-uniform coverage of the viewing sphere; ImageNet3D and Pascal3D+ concentrate sharply along the equator (canonical photograph angles), while HANDAL covers the hemisphere uniformly but contains very few categories.

4 DATASET

We present a comprehensive overview of our dataset, which comprises 22.8M images, 114K annotated objects, and 630 categories with five test instances each across many super-categories. The images are obtained by sampling 200 frames per video. In practice, given our limited compute budget, we sample only 16 frames per video in our experiments.

Category Distribution. The distribution over super-categories is visualized in Figure S1. As shown, the dataset exhibits broad semantic coverage, spanning diverse domains from indoor to outdoor object categories, thereby capturing substantial real-world diversity. Importantly, the category distribution remains largely unchanged, with only minor deviations from the original dataset. These changes primarily affect categories such as animals, where dynamic videos break the assumption of a single consistent coordinate frame, and meals, which often contain multiple small objects.

Viewpoint Distribution. Figure 3 illustrates the viewpoint distribution of the canonicalized objects. The dataset provides coverage from all viewing directions, indicating that objects are well represented across the full pose sphere. A slight bias toward frontal views and viewpoints mildly elevated above the horizontal plane is observable. This tendency can be attributed to common recording conditions in real-world videos: many objects naturally rest or are positioned upright (e.g., bicycles), or are placed such that their frontal surface faces upward (e.g., books). Consequently, front-facing and slightly top-down perspectives occur more frequently, while overall maintaining broad viewpoint diversity.

Train–Test Split. Every9D-22M contains 960 categories in total. For each category, the test set contains at most five object instances, for which we select only the highest-quality annotations; all test labels were additionally double-checked by human annotators. The test set covers 898 categories, 630 of which contain five object instances; evaluation is conducted on all test categories. The number of categories and object instances per split is reported in Table S1. All remaining videos, covering 895 of the 960 categories, form the training set, which is not limited in the number of object instances per category and typically contains many more instances per category than the test set.

Label Quality. Since the test set consists only of the highest-quality, double-checked annotations, the following quality assessment applies exclusively to the training set. To assess the accuracy of its generated 9D pose labels, we uniformly sample 125 approved training objects and independently annotate their poses by two human annotators; whenever the two annotations differ by more than 30°, the object is re-examined to establish the ground-truth orientation. Category-level symmetries are accounted for by accepting all equivalent orientations as valid. We find that 92.8% of the generated labels are accurate within 30° and 85.6% within 15°. Of the 9 labels with errors exceeding 30°, 5 correspond to unavoidable 180° ambiguities that cannot be resolved from appearance alone (e.g., a switched-off smartphone), 2 involve objects with an inherently ill-defined canonical orientation (e.g., a round hairbrush), and only 2 are genuine annotation mistakes, i.e., a false acceptance rate of merely 1.6%. Errors between 15° and 30° predominantly involve objects lacking reliable geometric cues, such as curled cats, squeezed teddy bears, or nearly spherical fruits; notably, even the two independent annotators disagreed by more than 15° in 4 of these 10 cases. Overall, the small number of genuine annotation errors and the predominance of intrinsically ambiguous cases demonstrate the reliability of the generated labels.

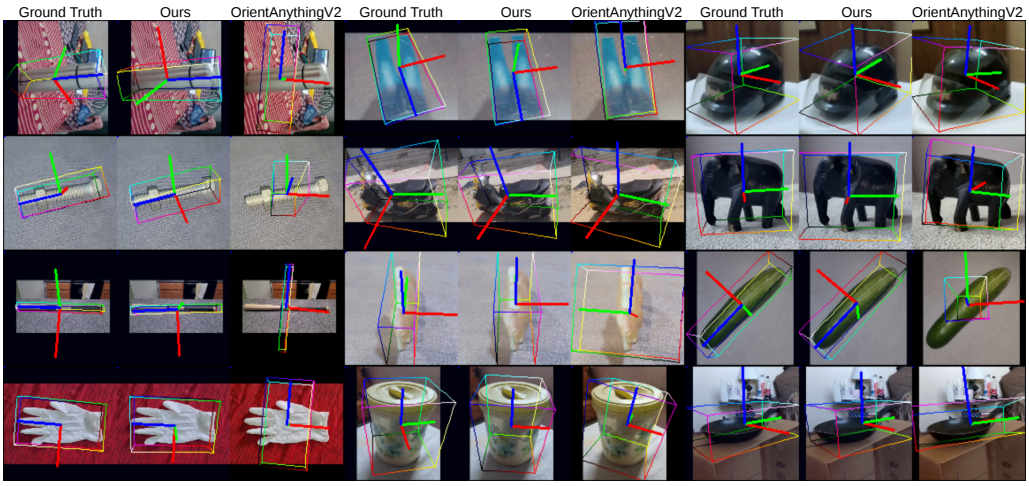


Figure 4: **Qualitative Results on Every9D-22M (test split).** Since OrientAnythingV2 does not predict 3D bounding boxes, we visualize its rotation estimates using ground-truth 3D boxes. Predicted axes are color-coded as red (left), green (back), and blue (top).

5 Experiments

We validate the key design choices underlying our canonicalization framework in Table 2, and evaluate on the disjoint training and test splits of Every9D-22M (Section 4). As shown in Table 3, training on Every9D-22M: (a) generalizes better to other datasets than any other training dataset; (b) outperforms prior foundation models in rotation accuracy on all benchmarks except ImageNet3D; and (c) reveals a strong viewpoint bias of existing datasets, which our jointly trained model overcomes.

5.1 Experimental Details

Datasets. We evaluate on all real-world pose datasets in Table 1 with more than 15 categories, namely ImageNet3D (Ma et al., 2024), HANDAL (Guo et al., 2023), Omni6DPose-ROPE (Zhang et al., 2024a), and Every9D-22M, except for ObjectNet3D (Xiang et al., 2016). ImageNet3D extends ObjectNet3D, aiming in particular to improve its annotation quality.

Symmetry-Aware Evaluation. The cross-category orientation rules (Appendix S4) identify category-level symmetries through underconstrained orientations, as described in Section 3.3.

5.2 Comparison with Prior Work

Every9D-22M generalizes best across datasets. In the lower part of Table 3, we train our model on individual datasets and evaluate on all others. On every benchmark other than its own, the model trained on Every9D-22M-train achieves the highest rotation accuracy among all training datasets. Compared to the next best training dataset, it improves coarse (30°) accuracy on ImageNet3D by **15.2 / 13.9** points (symmetry-unaware / symmetry-aware, vs. Omni6DPose-SOPE), on HANDAL by **3.9 / 3.3** points (vs. Omni6DPose-SOPE), and on Omni6DPose by **30.2 / 29.0** points (vs. ImageNet3D). On Omni6DPose, it also improves 3D IoU by **25.9** points and, despite the domain gap, recovers most of the accuracy of training on Omni6DPose-SOPE (54.0 vs. 62.8).

Every9D-22M outperforms prior foundation models. Every9D-22M poses a substantial challenge for state-of-the-art foundation models: the best of them reaches only 23.8 / 47.4 coarse (30°) accuracy on Every9D-22M-test. Our model trained on Every9D-22M-train outperforms the best foundation model in coarse rotation accuracy by **14.4 / 14.0** points on Every9D-22M-test, by **8.6 / 21.0** points on HANDAL, and by **18.7 / 18.8** points on Omni6DPose (symmetry-unaware / symmetry-aware). ImageNet3D is the only exception, discussed below. The qualitative comparisons in Figure 4 highlight three key advantages of our approach over OrientAnythingV2: improved robustness to unusual object orientations, more accurate estimation of subtle angular differences, and stronger resilience under

Dataset	ImageNet3D		HANDAL			Omni6DPose-ROPE			Every9D-test		
	Acc@30°		Acc@30°		3D IoU	Acc@30°		3D IoU	Acc@30°		3D IoU
Symmetry-Aware	×	✓	×	✓		✓*	✓		×	✓	
OrientAnythingV1 (Wang et al., 2024b)	41.6	48.4	2.9	7.4	–	27.7	29.8	–	18.6	37.5	–
OrientAnythingV2‡ (Wang et al., 2026)	55.7	62.7	2.1	3.8	–	31.2	33.4	–	23.8	46.1	–
WildDet3D (RGB-D) (Huang et al., 2026)	27.4	37.5	1.3	2.7	7.0	25.1	26.9	4.3	12.4	33.5	22.6
SAM3D (RGB-D) (Chen et al., 2025)	46.3	55.0	6.7	8.8	**29.3**	35.3	35.3	39.5	23.6	47.4	40.8
Ours (ImageNet3D + Every9D)	**59.1**	**65.4**	**12.2**	**23.2**	23.0	**51.8**	**51.4**	**39.6**	**33.7**	**57.7**	**48.0**
Ours (OrientAnythingV2†)	15.2	24.6	0.7	2.6	13.9	11.9	13.9	10.2	3.2	13.7	9.0
Ours (ImageNet3D)	**58.6**	**64.2**	4.2	10.0	15.4	23.8	25.1	10.7	10.2	28.0	9.6
Ours (HANDAL)	3.4	7.1	**64.7**	**77.7**	**46.4**	15.7	17.1	15.9	4.6	17.1	22.3
Ours (Omni6DPose-SOPE†)	15.7	25.3	11.4	26.5	27.5	**62.8**	**61.1**	**45.2**	22.8	46.9	40.7
Ours (Every9D)	30.9	39.2	15.3	29.8	26.9	54.0	54.1	41.8	**38.2**	**61.4**	**50.6**

Table 3: **Quantitative Results Across Datasets.** We report rotational accuracy at the coarse (30°) threshold, without and with symmetry-aware evaluation, and the average 3D IoU. The upper block compares our approach, trained jointly on ImageNet3D and Every9D, against foundation models; the lower block reports our approach trained on individual datasets. Best and second-best results within each block are **bold** and underlined. Dashes mark rotation-only models. *For Omni6DPose, this column is evaluated with the dataset's instance-level symmetries instead of our category-level ones. ‡OrientAnythingV2 is also trained on ImageNet3D. †Synthetic training datasets.

challenging lighting conditions. Furthermore, WildDet3D orientations appear strongly biased by gravity, resulting in poor rotation accuracy. In terms of 3D IoU, our model outperforms WildDet3D by **28.0** points and SAM3D by **9.8** points on Every9D-22M-test.

Existing datasets exhibit a strong viewpoint bias. Models trained on a single dataset perform well in-domain but degrade on other datasets, most pronouncedly for ImageNet3D, whose viewpoints concentrate sharply along the equator (Figure 3). Our model trained on ImageNet3D achieves 58.6 / 64.2 coarse accuracy in-domain, but only 4.2 / 10.0 on HANDAL and 10.2 / 28.0 on Every9D-22M-test; OrientAnythingV2, whose training data includes ImageNet3D, shows the same pattern (2.1 / 3.8 on HANDAL). Conversely, models trained on broader viewpoint coverage perform considerably worse on ImageNet3D, which thus rewards a prior towards canonical photograph viewpoints. To leverage both, we jointly train our model with mixed batches, where half of each batch is sampled from Every9D-22M-train and half from ImageNet3D (Appendix S3). The jointly trained model outperforms all prior foundation models on ImageNet3D, improving over OrientAnythingV2 by **3.4 / 2.7** points, and slightly improves over training on ImageNet3D alone (59.1 / 65.4 vs. 58.6 / 64.2), while substantially improving over it on all other benchmarks (Table 3).

6 Limitations

Symmetry handling at training time. Our evaluation accounts for category-level symmetries, but the baseline model itself predicts a single deterministic pose per image. Modeling symmetry-induced ambiguity through multi-hypothesis or distributional predictions is a promising direction for improving performance on rotationally symmetric categories.

Depth and dynamics. Depth annotations are inherited from monocular DepthAnythingV2 (Yang et al., 2024) estimates rather than ground-truth sensor depth, with scale aligned to the structure-from-motion reconstruction. The dataset also does not cover dynamic object interactions: each video is assumed to capture a static instance from multiple viewpoints.

7 Conclusion

We introduced **Every9D-22M**, 9D pose annotations for 22.8M real-world images across 630 everyday object categories, canonicalized from 114K object-centric videos with full human verification of every pose. Training on it generalizes better than on existing datasets, and joint training with ImageNet3D surpasses all prior foundation models in rotation accuracy on every benchmark.

AI USE STATEMENT

We used generative AI tools as general-purpose writing and coding assistants. Specifically, they were used to polish the phrasing of text that we wrote ourselves, and to assist with routine implementation tasks such as boilerplate code, plotting scripts, and refactoring. We did not use generative AI tools to generate research ideas, to derive or verify any of the claims made in this paper, to produce or label any part of Every9D-22M, or to retrieve and summarize related work. All AI-assisted text and code was reviewed by the authors; AI-assisted code was verified by inspection and by the experiments reported in Section 5 and the supplementary material. We take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

Every9D-22M is built entirely on top of uCO3D (Liu et al., 2025), which we use in accordance with its official license terms (CC BY 4.0); we credit the original authors accordingly. The underlying videos are object-centric recordings of everyday objects rather than human-centered data, so the dataset contains no personally identifying subjects by construction, and we introduce no new data collection involving human subjects. The manual annotation and verification described in Appendix S5 was performed by the authors, not by crowdworkers. We discuss the broader risks of large-scale 9D pose estimation, including surveillance misuse, category and geographic imbalance in the source videos, and the propagation of monocular depth errors into downstream systems, in Appendix S7.

REPRODUCIBILITY STATEMENT

The canonicalization framework is described in Section 3, with the design choices behind it ablated in Appendix S1. The dataset construction, filtering criteria, and train/test splits are given in Section 4 and Appendix S2, and the symmetry-aware evaluation protocol together with the orientation rules for every category is specified in Appendix S4. Full training details for the baseline models, including losses and weights, point cloud normalization, optimizer, schedule, batch size, and parameter counts, are reported in Appendix S3. The annotation and verification interfaces and the human effort they required are documented in Appendix S5. Data and code will be public upon acceptance.

REFERENCES

Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7822–7831, 2021.

Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 2773–2782, 2021.

Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, et al. Sam 3d: 3dfy anything in images. *arXiv preprint arXiv:2511.16624*, 2025.

Ho Kei Cheng and Alexander G Schwing. Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *European conference on computer vision*, pp. 640–658. Springer, 2022.

Yu Chi, Leonhard Sommer, Olaf Dünkel, Dominik Muhle, Daniel Cremers, Christian Theobalt, and Adam Kortylewski. C3PO: Canonicalization of 3d pose from partial views with generalizable correspondence features. In *Thirteenth International Conference on 3D Vision*, 2026.

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*, pp. 628–644. Springer, 2016.

Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21126–21136, 2022.

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*, 36:35799–35813, 2023a.

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*, pp. 13142–13153, 2023b.

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*, pp. 10286–10296, 2021.

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*, pp. 5834–5844, 2025.

Herbert Edelsbrunner and Ernst P Mücke. Three-dimensional alpha shapes. *ACM Transactions On Graphics (TOG)*, 13(1):43–72, 1994.

Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129(12): 3313–3337, 2021.

Yang Fu 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.

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

Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9224–9232, 2018.

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*, pp. 216–224, 2018.

Andrew Guo, Bowen Wen, Jianhe Yuan, Jonathan Tremblay, Stephen Tyree, Jeffrey Smith, and Stan Birchfield. Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11428–11435. IEEE, 2023.

Weikai Huang, Jieyu Zhang, Sijun Li, Taoyang Jia, Jiafei Duan, Yunqian Cheng, Jaemin Cho, Mattew Wallingford, Rustin Soraki, Chris Dongjoo Kim, et al. Wilddet3d: Scaling promptable 3d detection in the wild. *arXiv preprint arXiv:2604.08626*, 2026.

Artur Jesslen, Guofeng Zhang, Angtian Wang, Wufei Ma, Alan Yuille, and Adam Kortylewski. Novum: Neural object volumes for robust object classification. In *European Conference on Computer Vision*, pp. 264–281. Springer, 2024.

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

HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22498–22508, 2024.

Mukul Khanna, Yongsen Mao, Hanxiao Jiang, Sanjay Haresh, Brennan Shacklett, Dhruv Batra, Alexander Clegg, Eric Undersander, Angel X Chang, and Manolis Savva. Habitat synthetic scenes dataset (hssd-200): An analysis of 3d scene scale and realism tradeoffs for objectgoal navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16384–16393, 2024.

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*, pp. 4015–4026, 2023.

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.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

Yunzhi Lin, Jonathan Tremblay, Stephen Tyree, Patricio A Vela, and Stan Birchfield. Single-stage keypoint-based category-level object pose estimation from an rgb image. In *2022 International Conference on Robotics and Automation (ICRA)*, pp. 1547–1553. IEEE, 2022.

Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13978–13988, 2023a.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *Advances in neural information processing systems*, 36:44860–44879, 2023b.

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*, pp. 6881–6891, 2024.

Xingchen Liu, Piyush Tayal, Jianyuan Wang, Jesus Zarzar, Tom Monnier, Konstantinos Tertikas, Jiali Duan, Antoine Toisoul, Jason Y Zhang, Natalia Neverova, et al. Uncommon objects in 3d. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14102–14113, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

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

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*, 37:96127–96149, 2024.

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*, pp. 19521–19530, 2024.

Luca Medeiros. lang-segment-anything. `https://github.com/luca-medeiros/lang-segment-anything`, 2023. GitHub repository.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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*, pp. 165–174, 2019.

Anh Viet Phan, Minh Le Nguyen, Yen Lam Hoang Nguyen, and Lam Thu Bui. Dgcnn: A convolutional neural network over large-scale labeled graphs. *Neural Networks*, 108:533–543, 2018.

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.

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*, pp. 10901–10911, 2021.

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*, pp. 16969–16979, 2022.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

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*, pp. 22787–22796, 2024.

Leonhard Sommer, Olaf Dünkel, Christian Theobalt, and Adam Kortylewski. Common3d: Self-supervised learning of 3d morphable models for common objects in neural feature space. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 6468–6479, 2025.

Michael Stark, Michael Goesele, and Bernt Schiele. Back to the future: Learning shape models from 3d cad data. In *Bmvc*, volume 2, pp. 5, 2010.

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*, pp. 20845–20855, 2022.

Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3343–3352, 2019a.

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*, pp. 2642–2651, 2019b.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21686–21697, 2024a.

Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 21222–21231, 2022.

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 preprint arXiv:2412.18605*, 2024b.

Zehan Wang, Ziang Zhang, Jiayang Xu, Jialei Wang, Tianyu Pang, Chao Du, Hengshuang Zhao, and Zhou Zhao. Orient anything v2: Unifying orientation and rotation understanding. *arXiv preprint arXiv:2601.05573*, 2026.

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*, pp. 8792–8802, 2023.

Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4840–4851, 2024.

Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pp. 75–82. IEEE, 2014.

Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Choy, Hao Su, Roozbeh Mottaghi, Leonidas Guibas, and Silvio Savarese. Objectnet3d: A large scale database for 3d object recognition. In *European conference on computer vision*, pp. 160–176. Springer, 2016.

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)*, pp. 74–84. IEEE, 2021.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024.

Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 12498–12507, 2021.

Yuanwen Yue, Damien Robert, Jianyuan Wang, Sunghwan Hong, Jan Dirk Wegner, Christian Rupprecht, and Konrad Schindler. Litept: Lighter yet stronger point transformer. *arXiv preprint arXiv:2512.13689*, 2025.

Jinyu Zhang, Haitao Lin, Jiashu Hou, Xiangyang Xue, and Yanwei Fu. Beyond'templates': Category-agnostic object pose, size, and shape estimation from a single view. *arXiv preprint arXiv:2510.11687*, 2025a.

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

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, 2024b.

Yibo Zhang, Li Zhang, Rui Ma, and Nan Cao. Texverse: A universe of 3d objects with high-resolution textures. *arXiv preprint arXiv:2508.10868*, 2025b.

Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 17163–17173, 2023.

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*, pp. 1429–1439, 2021.

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)*, pp. 318–334, 2018.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019.