Omni6D: Large-Vocabulary 3D Object Dataset for Category-Level 6D Object Pose Estimation

Mengchen Zhang[1,2], Tong Wu[3], Tai Wang[2], Tengfei Wang[2], Ziwei Liu[4], and Dahua Lin[2,3]

[1] Zhejiang University, Zhejiang, China
[2] Shanghai Artificial Intelligence Laboratory, Shanghai, China
[3] The Chinese University of Hong Kong, Hong Kong SAR
[4] Nanyang Technological University, Singapore
{zhangmengchen,wangtai,wangtengfei}@pjlab.org.cn,
{wt020,dhlin}@ie.cuhk.edu.hk, ziwei.liu@ntu.edu.sg

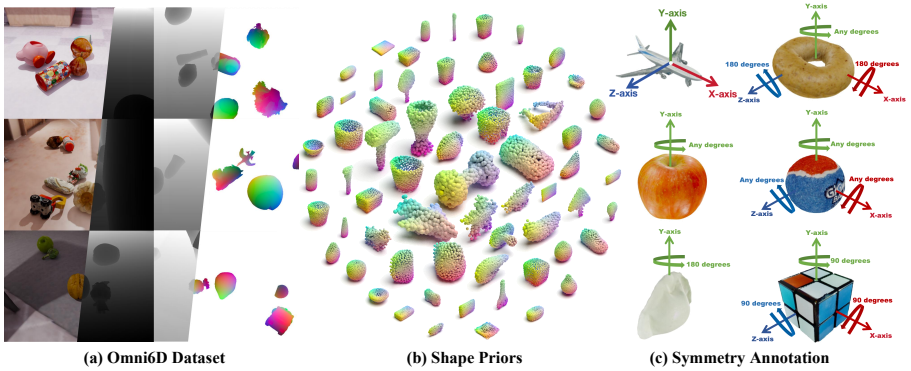


Fig. 1: Omni6D is a dataset for 6D object pose and size estimation with large vocabulary categories and rich annotations. **(a)** showcases ground truth of RGB image, depth map and NOCS map. **(b)** presents shape priors derived from a variational autoencoder [5] with adjusted canonical poses. **(c)** provides examples of the rotational symmetry of objects we have annotated, indicating the multiples of angles by which the shape remains unchanged when rotated around the xyz axes.

Abstract. 6D object pose estimation aims at determining an object's translation, rotation, and scale, typically from a single RGBD image. Recent advancements have expanded this estimation from instance-level to category-level, allowing models to generalize across unseen instances within the same category. However, this generalization is limited by the narrow range of categories covered by existing datasets, such as NOCS, which also tend to overlook common real-world challenges like occlusion. To tackle these challenges, we introduce **Omni6D**, a comprehensive RGBD dataset featuring a wide range of categories and varied backgrounds, elevating the task to a more realistic context. **1)** The dataset comprises an extensive spectrum of **166** categories, **4688** instances adjusted to the canonical pose, and over **0.8 million** captures, significantly broadening the scope for evaluation. **2)** We introduce a symmetry-aware

metric and conduct systematic benchmarks of existing algorithms on Omni6D, offering a thorough exploration of new challenges and insights. **3)** Additionally, we propose an effective fine-tuning approach that adapts models from previous datasets to our extensive vocabulary setting. We believe this initiative will pave the way for new insights and substantial progress in both the industrial and academic fields, pushing forward the boundaries of general 6D pose estimation.

Keywords: 6DoF Pose Estimation · Large Vocabulary Dataset · Metrics and Benchmarks

1 Introduction

6D pose estimation aims at predicting the position, orientation, and size of objects in a 3D space using RGB (D) images, enabling various applications such as augmented/virtual reality [26,33], robot manipulation [11,35], and scene understanding [15,28].

Early instance-level pose estimation approaches [32, 38, 39, 42, 43] typically involve providing instance CAD models and predicting poses of instances that were seen during training, restricting the generalization to unseen objects. In contrast, recent research has shifted towards category-level 6D object pose estimation [6–8,10,16,17,20,24,25,29,34,37,40,44–47], which learns category prior from a large number of instances within a category, allowing for pose estimation of new instances within the same category without the need for CAD models. By learning on a diverse range of categories, category-level approaches could be a more versatile solution for 6D pose estimation in real-world scenarios.

However, most existing datasets [22, 40, 44] are limited to a small number of object categories, typically less than 10, as shown in Tab. 1, hindering their practical applicability to complex scenes.

To overcome the limitations in previous category-level 6D pose estimation datasets, such as limited category numbers, lack of instance diversity within categories, and overly simplistic scenes, this paper presents a novel category-level dataset dubbed ***Omni6D*** for 6D pose estimation. ***Omni6D*** significantly extends the number of object categories to **166**, and includes **4,688** real-scanned and well-annotated instance objects with a diverse range of shapes, sizes, and textures. The constructed benchmark includes **0.8M** images featuring complex scenes with various occlusions, changing lighting conditions, complex backgrounds, and varying viewpoints. For each scene, we provide the rendered image, depth map, NOCS map, and instance mask. Also, considering the widespread rotational symmetry in objects, we examine three types of rotational invariance where an object maintains its original shape under following rotations: any degrees (Sym-1), multiples of 90 degrees (Sym-2) and 180 degrees (Sym-3). Additionally, we introduce a symmetry-aware metric to specifically address rotational invariance. Every object in Omni6D is adjusted to the canonical pose and annotated with rotational symmetry around three axes.

Table 1: Comparisons between Omni6D and existing datasets. Omni6D significantly extends the range of everyday object categories and instances.

Datasets	Mode	Realism	# Categories	# Instances	# Images
ShapeNet-SRN Cars [22]	RGB	Synthetic	1	3514	-
Sim2Real Cars [22]	RGB	Real	1	10	-
CAMERA [40]	RGBD	Synthetic	6	1085	0.3M
REAL [40]	RGBD	Real	6	42	8k
Wild6D [44]	RGBD	Real	5	1722	1M
Omni6D	RGBD	Real-Scanned	**166**	**4,688**	0.8M

Including a broader range of categories, our dataset offers a more comprehensive and challenging evaluation benchmark for category-level 6D object pose estimation. Utilizing Omni6D, we train and analyze existing algorithms, initiating a profound exploration of the challenges and vital elements involved in category-level estimation within large-vocabulary categories. Additionally, we assess these algorithms' capability to generalize across categories, and carry out a category-wise analysis. Experiments show that our dataset presents a more challenging benchmark for 6D pose estimation, highlighting the need for more robust and generalized pose estimation approaches. As an initial attempt, we present a finetuning strategy that assists in broadening the scope of existing approaches from a limited range of categories to a broader vocabulary. Moreover, we conduct an analysis of the domain gap between our dataset and real-world dataset, emphasizing the benefits of their combined use.

Our dataset will be publicly available to the research community, which will foster future research on more practical and robust 6D pose estimation algorithms and pave the way for broader applications.

2 Related Work

Existing work on category-level 6D object pose estimation can be generally divided into two types. After extracting features from images or point clouds, they compute Rotation, Translation, and Size (RTS) either through implicit point correspondence or explicit regression.

Existing Datasets. The most commonly used dataset for category-level 6D object pose estimation is NOCS [40], comprising both the synthetic CAMERA dataset and the real-world REAL dataset. CAMERA includes 300k RGBD images of 31 indoor scenes with 1,085 object instances across 6 categories, while REAL mirrors the categories in CAMERA and includes 8k RGBD images capturing 42 instances in 18 real scenes. Wild6D [44] consists of 5,166 videos with 1.1 million images over 1,722 object instances in 5 categories. ShapeNet-SRN Cars dataset and Sim2Real Cars dataset proposed in iNerf [22] both exclusively include a single car category. The former includes 3,514 instances derived from ShapeNet cars, while the latter is extracted from videos capturing 10 distinct unseen car models. These datasets are limited by their narrow range of categories, hindering their ability to generalize broadly. Additionally, most training images

are synthetic and lack realism, and their scenes are overly simplified, failing to account for common real-world challenges like occlusions.

Implicit Methods. Implicit methods are based on point correspondence [6, 20, 24, 34, 37, 40, 44, 46] . NOCS [40], one of the pioneering works in this area, introduced the concept of Normalized Object Coordinate Space (NOCS). The final pose and size of the object are obtained by matching the predicted NOCS map with the observed depth input using the Umeyama algorithm [36] and RANSAC algorithm [12].

Subsequent algorithms such as DualPoseNet, RBP-Net and RePoNet [20, 44, 46] have continued to develop along the vein of NOCS, implicitly solving for pose after predicting the NOCS map. SPD [34] proposed a category-level shape prior, subsequently deforming this shape prior (i.e., average shape) to fit observed point cloud. SGPA, RePoNet, and CATRE [6, 24, 44] continue to develop along SPD's category-level shape prior approach. Algorithms like 6-PACK and SGPA [6, 37] extract low-rank structure points, i.e., keypoints, from dense observed point clouds. 6-PACK [37] predicts interframe motion of target instances through keypoint matching, while SGPA [6] employs keypoints for more effective incorporation of sparse structural information during prior adaptation. These methods rely heavily on the RANSAC process to eliminate outliers, making them non-differentiable and time-consuming.

Explicit Methods. Explicit methods are based on direct pose regression [7, 10, 20, 24, 46, 47]. DualPoseNet and RBP-Net [20, 46] conduct both explicit and implicit training,, where one parallel pose decoder explicitly regresses the pose. CATRE [24], recognizes the inherent difference between estimations of rotation and translation/size, explicitly regressing their residuals and carrying out an iterative pose estimation process. FS-Net [7] designs an autoencoder with 3D Graphic Convolution for latent feature extraction and separates the predictions for rotation and translation/size into two distinct networks: one estimates translation/size through two residuals, while the other handles rotation prediction by estimating deflections on two orthogonal axes. GPV-Pose and HS-Pose [10, 47] utilize the same foundational mechanism introduced by FS-Net [7]. GPV-Pose [10] proposes a decoupled confidence-driven rotation representation that facilitates geometrically-aware recovery of correlated rotation matrices and introduces a new geometry-guided point-by-point voting paradigm for robust retrieval of 3D object bounding boxes. Meanwhile, HS-Pose [47] extends 3D-GC to extract mixed-range latent features from point cloud data through a simple network structure known as the HS layer.

3 Omni6D Dataset

3.1 Construction

Dataset Collection. As shown in Tab. 1, Omni6D comprises 4,688 instances across an impressive span of 166 categories. Each instance is a high-resolution

textured mesh, obtained using Shining 3D scanner[1] and Artec Eva 3D scanner[2], collected from OmniObject3D [41]. We normalize object models to fit within a $(-1, 1)^3(m^3)$ three-dimensional space, and align objects within each category to a consistent canonical pose. In the latest dataset, Omni6D-xl builds upon and extends Omni6D, comprising 13,161 instances across an impressive span of 339 categories. For more details, please refer to Appendix Section C.

Rendering. We employ stratified sampling to split instances within each category, subsequently dividing them into training, validation, and test sets in a 7:2:1 ratio. In the construction of our dataset, we utilize 9 room models from the Replica dataset as backdrops. For each scenery setup, we randomly select a room model to act as the background, along with $6 - 8$ object instance models, which are allowed to perform free-fall motion within the room model, resulting in random scattering in a specific section of the room. Each object model is scaled by a random factor ranging from 0.8 to 1.2 as part of our data augmentation strategy. Considering the attention center of the combined instance models as the origin point, the camera randomly selects ten positions within a radius of $8-9$ m and an elevation angle range between $30-90°$. The camera then performs rendering at these selected positions while facing towards the attention center.

Setting. We utilize BlenderProc 2.5.0 [9] to implement the aforementioned rendering process. The intrinsic parameters of the camera are set to [577.5, 577.5, 319.5, 239.5], with an image size specified as 640×480. Our approach ensures the diversity and breadth of the dataset, making it suitable for rigorous testing and yielding accurate results.

3.2 Data Annotations

Rich Annotations. Each rendered output includes the ground truth class label, instance mask, NOCS mapping [40], depth map, as well as 6D pose and size. Fig. 1 exhibits a selection of rendered outputs. To reduce the storage size of the dataset, we encode high-precision depth maps into RGB images by multiplying depth by 10,000, rounding to nearest integer, and converting to base 256. The resulting three digits represent RGB channels.

Rotational Invariance. Rotational invariance implies that a symmetric object can retains its original shape after rotation by certain angles. Many common objects have this property. As shown in Fig. 6, we define the coordinate system as a right-handed system with the x-axis pointing outwards and the y-axis oriented upwards. We contemplate three cases of rotational invariance where an object maintains its original shape after following rotations: any degrees (**Sym-1**), multiples of 90 degrees (**Sym-2**) and 180 degrees (**Sym-3**). Additionally, we denote the case of no rotational invariance around the axis as **Sym-0**. According to these definitions, all objects in Omni6D are annotated for their rotational symmetry around the xyz-axes. It's worth noting that symmetry attributes may differ among instances within the same category, requiring instance-level rather

[1] https://www.einscan.com/
[2] https://www.artec3d.cn/

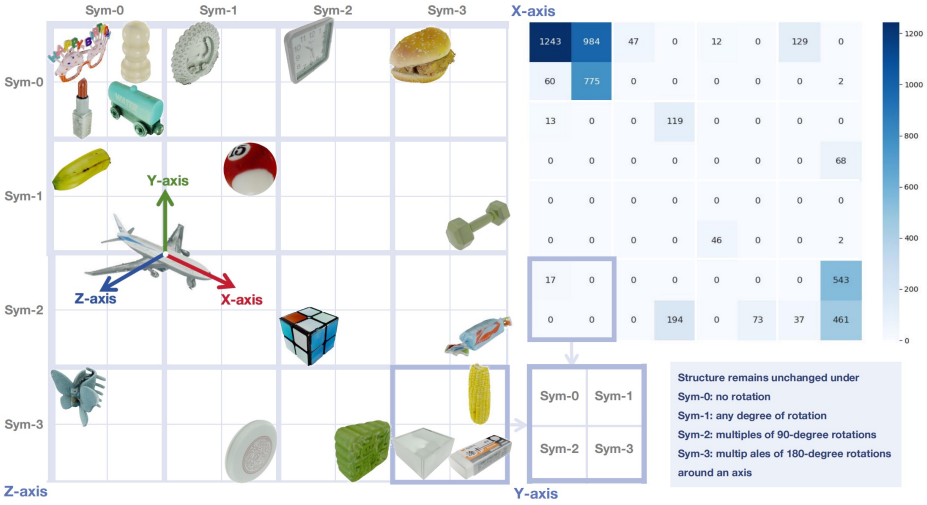


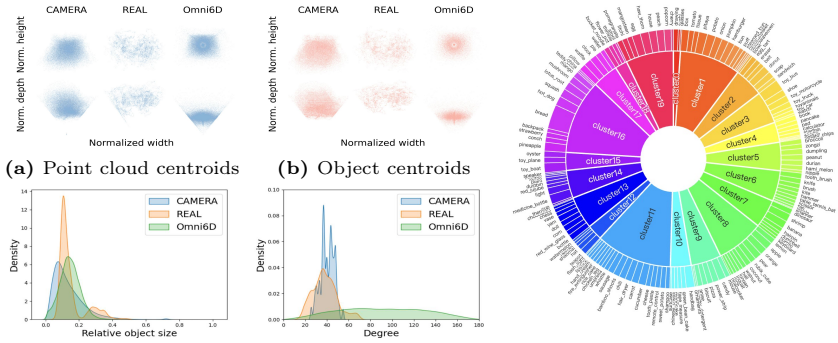
Fig. 2: Symmetry statistics. The figure demonstrates different symmetry cases using object instances and provides a quantitative representation of the occurrence frequency for various combinations of distinct symmetry cases across the xyz-axes.

than category-level annotations. Fig. 6 illustrates all kinds of symmetry cases using object instances and quantifies their occurrence frequency. Fig. 1 selects several examples to provide a more visual explanation of rotational invariance. These considerations are then integrated into our evaluation protocols in Sec. 4.2.

3.3 Dataset Statistics

Spatial Statistics. Omni6D aims to overcome challenges in estimating poses for occluded object instances. Fig. 3a and Fig. 3b shows the spatial distribution of point clouds and objects by projecting their centroids on the XY-plane (top) and XZ-plane (bottom) [4]. Fig. 3c depicts the relative object size distribution, defined as the square root of the object-to-image area ratio. We observe that the spatial distribution of Omni6D is similar to that of CAMERA and REAL, with a greater resemblance to CAMERA despite having a closer depth range. However, a more pronounced discrepancy between the spatial distribution of point clouds and objects is evident in Omni6D compared to CAMERA and REAL. This observation suggests a higher occurrence of occlusion scenes in Omni6D, highlighting the intricate challenges it presents to 6D object pose estimation. Nonetheless, as depicted in Fig. 4a, algorithms trained on Omni6D demonstrate their robustness in tackling these complexities.

Angular Deviation. Omni6D enables accurate pose estimation using only the lower half or bottom appearance of objects. Fig. 3d depicts the density of angular deviations from the upward direction, *i.e.* y-axis. Our dataset displays a more uniform distribution of object angles relative to the upward axis and exhibits greater deviation from the canonical pose angles. Unlike NOCS, which primarily



(a) Point cloud centroids **(b)** Object centroids **(c)** Relative 2D object size **(d)** Angular deviation **(e)** Clustering results

Fig. 3: Omni6D analysis. (a) distribution of point cloud centroids, **(b)** distribution of object centroids on (top) normalized image, XY-plane, and (bottom) normalized depth, XZ-plane, **(c)** density of relative 2D object size, **(d)** density of angular deviation from the upward direction, **(e)** Omni6D dataset clustering results. The angle of each sector in the chart reflects the relative size of the instance count within that category.

uses upright object placement, Omni6D utilizes physical simulations for free-fall object positioning [9]. As a result, it presents more challenging and diverse pose estimation scenes. Training on Omni6D enhances algorithms' robustness to object rotation angles, as evidenced by the image in Fig. 4b.

Shape Priors. We obtain the mean latent embedding and shape prior for each category from the variational autoencoder [5]. Fig. 1 showcases categorical shape priors, each displaying unique characteristics, facilitating an intuitive association between point cloud shapes and corresponding real-world entities. Meanwhile, Fig. 3e explains clustering results based on categorical latent embeddings, where we employ agglomerative clustering [27] to group categories into 20 clusters. It highlights the geometric coherence among semantically identical objects (especially man-made ones) in Omni6D dataset and further confirms that these categorical shape priors can effectively leverage the wealth of shape information from numerous similar objects to elucidate category features. These insights provide a theoretical basis for applications of category-level 6D object pose estimation using our Omni6D dataset.

4 Evaluation and Analysis

4.1 Experimental Setup

Datasets. Our experimentation utilized two datasets, namely Omni6D and Omni6D$_{out}$. Omni6D are partitioned into training, validation, and test sets in a 7:2:1 ratio, denoted as Omni6D$_{train}$, Omni6D$_{val}$ and Omni6D$_{test}$ respectively. These sets are further subdivided into subsets with increasing category sizes of 3, 6, 12, 24, and 48. We denote the subset containing n categories as cls$n$. Each subset includes all classes present in the previous subset with additional

(a) Challenges from occluded object **(b)** Challenges from bottom views

Fig. 4: Challenges from Omni6D. (a) Algorithms trained on Omni6D can overcome challenges in estimating poses for occluded object instances. The **left** shows an occluded object instance at the edge of the image, while the **right** image shows an object instance obstructed by other objects. **(b)** Algorithm trained on Omni6D can accurately estimate poses with only the lower half or bottom appearance of an object. The green and red colors respectively denote the ground truth and predicted 3D bounding boxes. The blue and orange lines on the boxes separately highlight the intersecting lines of the frontal face and the top face of the two 3D bounding boxes, while the darker lines indicate the bottom of the bounding boxes.

classes included to meet the desired total. Fig. 6a presents the specific categories included in cls$n$ and their respective sizes relative to each other. Omni6D$_{out}$ is utilized as an additional test set to measure our algorithm's inter-category generalization. This dataset, constructed similarly to Omni6D, encompasses 52 models spanning 17 categories unseen in Omni6D, along with 4762 images. For additional details on datasets, please refer to the appendix.

Details. All experiments are carried out on a server equipped with an Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz and an NVIDIA A100-SXM4-80GB GPU. We maintain consistency in parameters and strategies throughout training, ensuring uniformity in our experiment environment. Given the challenges of semantic classification with a large vocabulary, we use ground truth masks to mitigate the impact of low-quality classification on pose estimation results.

4.2 Symmetry-Aware Evaluation

Basic Evaluation Metrics. We utilize the average accuracy of Intersection over 3D Union (IoU) [14] in object detection, and $n°m\ cm$ in pose estimation. We further decompose $n°m\ cm$ [19,31] to individually evaluate the model's predictive error $n°$ for pose and $m\ cm$ for translation. For these three types of errors, the thresholds considered are $\{50\%, 75\%\}$, $\{5°, 10°\}$ and $\{2\ cm, 5\ cm\}$ [3, 30, 42]. Additionally, we set a detection threshold for objects requiring at least a 10% overlap between predicted and ground-truth bounding boxes.

Our Symmetry-Aware Metrics. Due to NOCS's limited categories, traditional algorithms mainly handle basic symmetry cases, such as rotational symmetry around the y-axis. However, Omni6D has a wider range of objects with different rotational invariances across multiple axes. Fig. 6 provides symmetry statistics for Omni6D objects. To alleviate this issue, we propose a symmetry-

Algorithm 1 Compute Our Symmetry-Aware Metric L_s

procedure SYMMETRIC_METRIC(L, R, n_x, n_y, n_z)
  $\Theta_0 = \{0^\circ\}$
  $\Theta_2 = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$
  $\Theta_3 = \{0^\circ, 180^\circ\}$ *// Rotations around Sym-1 axis need not be considered.*
  $c = \text{count}(1 \text{ occurrences in } \{n_x, n_y, n_z\})$
  if $c \geq 2$ **then** *// The object is a sphere.*
    $L_s = L(R^*, R)$
  else if $c == 1$ **then** *// Rotations around Sym-1 axis can be disregarded.*
    Without loss of generality, assume $n_x == 1$.
    $L_s = \min_{\theta_y \in \Theta_{n_y}, \theta_z \in \Theta_{n_z}} L(R^*_{\theta_y, \theta_z}, R)$
  else if $c == 0$ **then** *// Simply enumerate all cases.*
    $L_s = \min_{\theta_x \in \Theta_{n_x}, \theta_y \in \Theta_{n_y}, \theta_z \in \Theta_{n_z}} L(R^*_{\theta_x, \theta_y, \theta_z}, R)$
  end if
  return L_s
end procedure

aware metric. Unlike prior works focusing solely on the y-axis, our method considers rotation symmetry around all three axes.

We define the relevant variables as follows: L_s denotes our symmetry-aware metric, L denotes the original metric. R stands for the ground truth rotation matrix, while R^* represents the predicted rotation matrix. $R^*_{\theta_x, \theta_y, \theta_z}$ corresponds to the predicted rotation matrix after sequentially rotating by θ_x, θ_y, and θ_z degrees around the xyz axes. The rotational invariance cases around the x, y, and z axes are denoted as Sym-n_x, Sym-n_y, and Sym-n_z, where n_x, n_y, and n_z are the respective rotation parameters. Objects that align with Sym-n around an axis maintain their original shape when rotated by an angle from Θ_n.

Since the Euler angles are compact [13], the most straightforward approach is to determine the category of rotational invariance for each axis {x, y, z} sequentially, as mentioned in 3.2. To simplify computations, we set $\Theta_0 = \{0^\circ\}$, $\Theta_1 = \{0^\circ, 1^\circ, ..., 359^\circ\}$, $\Theta_2 = \{0^\circ, 90^\circ, 180^\circ, 270^\circ\}$, $\Theta_3 = \{0^\circ, 180^\circ\}$. We can define L_s as $L_s = \min_{\theta_x \in \Theta_{n_x}, \theta_y \in \Theta_{n_y}, \theta_z \in \Theta_{n_z}} L(R^*_{\theta_x, \theta_y, \theta_z}, R)$.

However, due to the singularity of Euler angles [13], we can simplify the above rotation transformation. The pseudo-code implementation of our Symmetry-Aware Evaluation is provided in Algorithm 1. It allows us to simplify what was originally at most 360^3 computations to a maximum of only 4^3 computations.

4.3 Large-Vocabulary 6D Pose and Size Estimation

Performance on Omni6D. We present results of algorithms [6, 10, 34, 46, 47] trained on Omni6D$_{train}$ and tested on Omni6D$_{test}$. We compare their quantitative results in Tab. 2 and their qualitative results in Fig. S10 in Appendix. The performance disparity among algorithms for category-level 6D object pose estimation becomes markedly pronounced when applied to large-vocabulary datasets, in contrast to the more consistent performance previously observed on the Real and CAMERA datasets [40]. This highlights the inherent strengths and weaknesses across various model structures.

This observation suggests the potential importance of our large-vocabulary dataset in uncovering the relative performance of different models. It appears

Table 2: Category-level performance on Omni6D dataset. Models are trained on Omni6D$_{train}$ and tested on Omni6D$_{test}$. Instances within each category in the test set are unseen during training, substantiating the algorithms' capacity to generalize within individual categories under large-vocabulary settings. **Bold** and underlined results indicate the best and second-best performers.

Methods	Network	IoU_{50}	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$5°$	$10°$	$2cm$	$5cm$
SPD [34]	implicit	44.56	20.37	7.55	**9.56**	14.76	**19.23**	**10.68**	**21.02**	37.49	70.09
SGPA [6]	implicit	36.34	14.44	4.78	6.84	10.13	15.03	8.49	17.73	25.57	59.18
DualPoseNet [20]	hybrid	58.84	**25.49**	**8.28**	9.30	**17.26**	19.05	9.38	19.18	73.82	96.37
RBP-Pose [46]	hybrid	35.92	4.66	0.37	0.60	0.53	0.80	0.75	0.96	39.73	83.55
GPV-Pose [10]	explicit	15.28	0.26	0.10	0.70	0.14	0.96	2.25	2.96	5.31	33.70
HS-Pose [47]	explicit	**62.65**	23.02	4.26	4.85	10.49	11.61	4.96	11.75	**80.93**	**97.78**

Table 3: Category-level performance on unseen categories. Models are trained on Omni6D$_{train}$ and tested on Omni6D$_{out}$. Categories in the test set never appear in the training set, validating the algorithms' ability to generalize across categories.

Methods	Network	IoU_{50}	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$5°$	$10°$	$2cm$	$5cm$
SPD [34]	implicit	7.56	0.95	0.18	0.40	0.80	1.65	0.65	2.36	8.88	40.59
SGPA [6]	implicit	7.05	0.60	0.07	0.28	0.19	0.82	0.53	1.69	3.87	28.28
DualPoseNet [20]	hybrid	**36.85**	**12.06**	**3.24**	**3.37**	**8.04**	**8.51**	**3.39**	**8.64**	78.00	98.60
RBP-Pose [46]	hybrid	26.18	1.95	0.01	0.02	0.02	0.03	0.02	0.03	16.74	43.06
GPV-Pose [10]	explicit	10.97	0.14	0.03	0.18	0.12	0.57	0.30	1.07	7.14	41.30
HS-Pose [47]	explicit	36.75	8.92	1.54	1.66	4.67	5.16	1.75	5.38	**79.95**	**98.27**

that the increased complexity of the dataset could push model architectures to their theoretical limits, possibly revealing intrinsic characteristics otherwise obscured in less complex scenarios. For example, SPD, SGPA is particularly proficient in predicting rotation, and SPD achieves the highest score in $n°m\ cm$. This could be due to its implicit network's propensity for generating more reliable rotational forecasts. Meanwhile, DualPoseNet and HS-Pose provide more accurate predictions for translation and score higher in IoU. This could be associated with the characteristic of models with explicit networks to produce better translations and size estimates.

Our large-vocabulary dataset, encompassing a broad spectrum of shapes and appearances, enables a comprehensive evaluation of diverse category-level pose estimation methods. This serves not only as a robust test of an algorithm's generalizability but also as a valuable tool in understanding the advantages offered by different algorithmic structures.

Generalization Performance. We evaluate algorithms on Omni6D$_{out}$ to assess their inter-category generalization capabilities. The outcomes are presented in Tab. 3. Notably, DualPoseNet and HS-Pose emerged as superior performers, outclassing others across all metrics, thereby demonstrating excellent generalization abilities. Contrastingly, implicit methods including SPD and SPGA exhibited marked limitations. Qualitative results are shown in Fig. S11 in Appendix.

Drawing parallels with the observations from Tab. 2, we found that metrics such as translation and IoU were relatively easier to excel in, suggesting superior

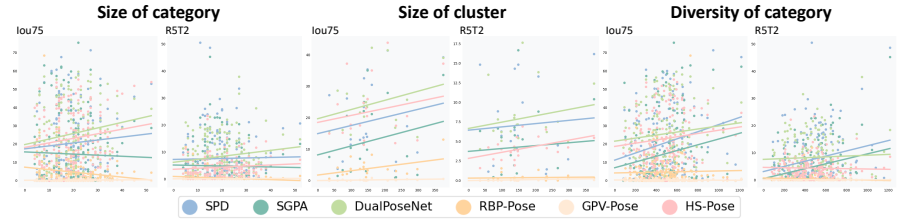
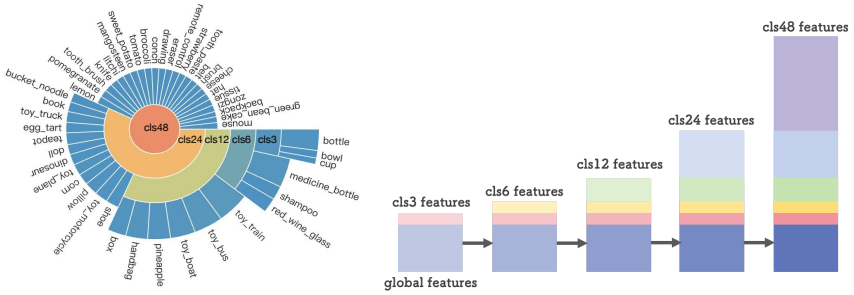


Fig. 5: Category-Wise Performance on Omni6D Dataset. The x-axis, moving from left to right, sequentially represents: the number of objects within a category (**Semantic Category**), the number of objects within a cluster clustered based on shape priors (**Shape Category**) and the diversity of instances within a category. The y-axis depicts category or clustered group results for IoU_{75} and $5°2\ cm$ metrics. Each plotted point illustrates the algorithm's result for a specific category or cluster, while the line showcases the trend of the linear fit for the scattered points.

generalization abilities in translation and size prediction. Conversely, the generalization of rotation emerges as a considerable challenge in category-level 6D object pose estimation, especially within large-vocabulary scenes.

Category-wise Analysis. Based on the IoU_{75} and $5°2\ cm$ metrics, we conducted a detailed category-wise analysis of the results from Tab. 2. Left columns in Fig. 5 illustrate the correlation between category-level 6D pose estimation performance and the number of instances within each category in Omni6D_{train}. Middle columns in Fig. 5 analyze the correlation between cluster-level average performance and cluster size based on the clustering results described in Fig. 3e. We found that the performance of pose estimation for each category is more strongly correlated with the number of instances within clusters than with semantic categories, showing a positive correlation. This suggests that shape categories have a greater impact on training than semantic categories do. Notably, algorithms like SPD, SGPA, and RBP-Pose that utilize shape prior structures are particularly sensitive to this influence.

Right columns in Fig. 5 reveal the correlation of pose estimation performance relative to instance diversity within each category in the training set. We measured instance diversity by calculating the mean chamfer distance [1] among all pairs of instances in each category. The results show that as diversity within a category increases, pose estimation performance tends to improve. This observation aligns with the assertion made by [23]: The key to the success of prior-based methods lies in the deformation modules, which learns to synthesize world-space target objects and explicitly builds the correspondence between camera and world-space. As the number of instances increases and the diversity within a shape category expands, the model's capacity to learn deformation from priors to actual instance shapes is strengthened, leading to improved results.



(a) Category inventory of cls$n$ **(b)** Our finetune strategy

Fig. 6: Our finetune strategy. (a) Category inventory of cls$n$ within Omni6D dataset. The angle of each sector in the chart reflects the relative size of the instance count within that category. **(b)** In each fine-tuning step, we double the category count, copying trained global features and old category parameters into the new network while initializing the new category parameters. An observable deepening of color is indicative of the escalating count of training iterations.

4.4 Fine-Tuning from Limited Categories

We propose a finetuning strategy that helps extend methods from a limited set of categories to large-vocabulary. We take SPD [34], DualPoseNet [20], and HS-Pose [47] as examples which belong to three different network architectures and show good performance on Omni6D$_{test}$. We respectively take their best models on CAMERA as our pre-trained models.

Initiating the fine-tuning process, we utilize three categories: bottle, bowl, and cup, which are concurrently present in both Omni6D and CAMERA datasets, aligning with the cls3 category. By facilitating the training on Omni6D-cls3, we enable a transfer of the model from CAMERA to Omni6D. Following the method illustrated in Fig. 6b, we engage in an iterative fine-tuning process on a progressively expanded category dataset until it reaches our desired number. In our experiments, we set this target number to be 48 categories.

In parallel, we conduct training from scratch separately on cls3, cls6, ..., and cls48 as a comparison, employing the same number of training iterations. As shown in Fig. 7, even with an exponential increase in the number of categories, pre-trained models remain pivotal in our fine-tuning strategy. The performance of fine-tuning consistently outperforms that of training from scratch.

However, regardless of whether the training approach is finetuning or training from scratch, a decline in performance is observed as the number of categories increases. The decline rates for SPD and DualPoseNet are slower, coupled with an initial augmentation in performance due to increased training data and iterations. In contrast, HS-Pose experiences a more rapid decline, with fine-tuned $5°2\,cm$ results dropping from initial 62.52% to 14.42%. Models that excel in tasks involving a limited number of categories may not necessarily maintain their superiority in large-vocabulary tasks, they might be surpassed by models that are more robust and easier to train.

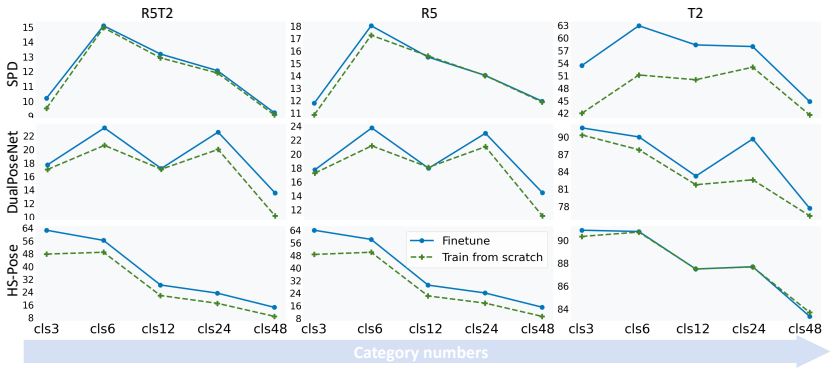


Fig. 7: Finetuned results. Each figure's x-axis represents the number of categories in the training and test set, while the y-axis displays the outcomes of $5°2\ cm$, $5°$ and $2\ cm$ metrics. Each row, from top to bottom, sequentially employs three methods: SPD [34], DualPoseNet [20], and HS-Pose [47]. The figures depict the outcomes derived from two training strategies as the number of training categories increases, accompanied by the gradual expansion of corresponding test sets.

4.5 Visual Realism

Due to the complexity of collecting and annotating real-world data, contemporary datasets like NOCS [40] are composed of a large amount of synthetic data and a small portion of real-world data. While collecting real data is relatively straightforward when the number of categories is limited, gathering well-annotated real-world data for pose estimation tasks involving large vocabulary categories becomes a monumental task.

Our Omni6D dataset, which includes large vocabulary objects, is also derived from rendering. However, the incorporation of real-scanned objects significantly enhances the realism of the rendered images. As depicted in Fig. 8, Omni6D receives a score of 2.69 ± 0.39, surpassing the results obtained by CAMERA.

Given these significant advantages, our dataset excels not only in large-vocabulary scenarios but also in real-world scenes. As depicted in Tab. 4, We use DualPoseNet [20] to train on the common categories in REAL [40] and Omni6D, namely bottles, bowls, and mugs. We train separately on the two datasets and their mix. The results show that Omni6D models perform well on REAL275, and training on the mixed dataset outperforms using REAL or Omni6D datasets alone. This demonstrates that our dataset enables the direct transfer of models to real-world scenes. Moreover, it seamlessly supplements the existing real-world dataset, enabling joint training of models on our dataset and the real-world data.

To further validate the sim2real capability of models trained with Omni6D, we constructed a real-world dataset, ***Omni6D-Real***, comprising 30 scenes, 39 categories, 73 instances, and 1k images. We captured RGBD images with Azure

Fig. 8: Comparison of Visual Realism. We evaluated the visual realism of Omni6D in comparison to other datasets through a survey involving 70 human subjects. We randomly selected 10 images from each dataset and introduced noise by blending in 5 images from COCO [21], which included captured photos, and SKETCH[4], which comprised rendered images. Subjects were asked to rate the realism of sampled images on a scale from 1 (least realistic) to 5 (most realistic). We report the mean and standard deviation and include a sampled image from the study.

Table 4: Performance on REAL275 with Different Training Sets. It compares how different training sets influence DualPoseNet's performance on REAL275 [40], providing insights into the model's ability to generalize in real-world tasks using Omni6D.

Train data	Realism	IoU_{50}	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$5°$	$2cm$
Omni6D	Real-Scanned	78.76	32.69	6.55	8.80	15.00	21.38	11.20	49.54
REAL [40]	Real	84.51	43.43	8.76	10.40	21.24	25.39	13.01	69.46
REAL+Omni6D	Mixed	**85.28**	**58.59**	**14.10**	**17.83**	**30.10**	**38.97**	**20.96**	**71.00**

Kinect DK[3] and preprocessed them using SAM [18] for object masks and ICP [2] for point cloud registration. Details are provided in Appendix Section D.

5 Conclusion

In conclusion, this paper introduces ***Omni6D***, a novel 6D pose estimation dataset with large-vocabulary categories and intricate scenes. We evaluate existing category-level 6D object pose estimation methods on this benchmark, analyze its challenges, and propose a fine-tuning strategy for large-vocabulary scenarios.
Limitations. Our dataset, though more complex, doesn't fully encompass all real-world challenges. Additionally, our fine-tuning strategy effectively extends methods from a small set to a larger one, but its efficacy may decrease with growing category diversity.
Future Work. Our study paves the way for diverse research avenues. An immediate next step is expanding the Omni6D dataset with more object types and scenes for comprehensive coverage. Additionally, annotating videos for scanned objects will validate algorithms' large-vocab pose estimation in real-world scenarios. Designing new training strategies for coping with increasing category diversity presents an intriguing challenge.

[3] https://learn.microsoft.com/azure/kinect-dk/
[4] https://sketchfab.com/

Acknowledgements

This research is supported by Shanghai Artificial Intelligence Laboratory, the Centre for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission (ITC)'s InnoHK, the Ministry of Education, Singapore, under its MOE AcRF Tier 2 (MOET2EP20221- 0012), NTU NAP, and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative.

References

1. Barrow, H.G., Tenenbaum, J.M., Bolles, R.C., Wolf, H.C.: Parametric correspondence and chamfer matching: Two new techniques for image matching. In: IJCAI. pp. 659–663. William Kaufmann (1977)
2. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**(2), 239–256 (1992)
3. Brachmann, E., Michel, F., Krull, A., Yang, M.Y., Gumhold, S., Rother, C.: Uncertainty-driven 6d pose estimation of objects and scenes from a single RGB image. In: CVPR. pp. 3364–3372 (2016)
4. Brazil, G., Kumar, A., Straub, J., Ravi, N., Johnson, J., Gkioxari, G.: Omni3d: A large benchmark and model for 3d object detection in the wild. In: CVPR. pp. 13154–13164 (2023)
5. Chen, D., Li, J., Wang, Z., Xu, K.: Learning canonical shape space for category-level 6d object pose and size estimation. In: CVPR. pp. 11970–11979. Computer Vision Foundation / IEEE (2020)
6. Chen, K., Dou, Q.: SGPA: structure-guided prior adaptation for category-level 6d object pose estimation. In: ICCV. pp. 2753–2762 (2021)
7. Chen, W., Jia, X., Chang, H.J., Duan, J., Shen, L., Leonardis, A.: Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In: CVPR. pp. 1581–1590 (2021)
8. Chen, X., Dong, Z., Song, J., Geiger, A., Hilliges, O.: Category level object pose estimation via neural analysis-by-synthesis. In: ECCV (26). pp. 139–156 (2020)
9. Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K.H., Humt, M., Triebel, R.: Blenderproc2: A procedural pipeline for photorealistic rendering. J. Open Source Softw. **8**(83), 4901 (2023)
10. Di, Y., Zhang, R., Lou, Z., Manhardt, F., Ji, X., Navab, N., Tombari, F.: Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In: CVPR. pp. 6771–6781 (2022)
11. Du, G., Wang, K., Lian, S., Zhao, K.: Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. Artif. Intell. Rev. **54**(3), 1677–1734 (2021)
12. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24**(6), 381–395 (1981)
13. Gao, X., Zhang, T.: Introduction to Visual SLAM - From Theory to Practice. Springer (2021)
14. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The KITTI dataset. Int. J. Robotics Res. **32**(11), 1231–1237 (2013)

15. Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y.N., Zhu, S.: Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In: NeurIPS. pp. 206–217 (2018)
16. Irshad, M.Z., Kollar, T., Laskey, M., Stone, K., Kira, Z.: Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation. In: ICRA. pp. 10632–10640. IEEE (2022)
17. Irshad, M.Z., Zakharov, S., Ambrus, R., Kollar, T., Kira, Z., Gaidon, A.: Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In: ECCV (2). Springer (2022)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. arXiv:2304.02643 (2023)
19. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: Deepim: Deep iterative matching for 6d pose estimation. Int. J. Comput. Vis. **128**(3), 657–678 (2020)
20. Lin, J., Wei, Z., Li, Z., Xu, S., Jia, K., Li, Y.: Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In: ICCV. pp. 3540–3549 (2021)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755 (2014)
22. Lin, Y., Florence, P., Barron, J.T., Rodriguez, A., Isola, P., Lin, T.: inerf: Inverting neural radiance fields for pose estimation. In: IROS. pp. 1323–1330 (2021)
23. Liu, J., Chen, Y., Ye, X., Qi, X.: Prior-free category-level pose estimation with implicit space transformation. CoRR **abs/2303.13479** (2023)
24. Liu, X., Wang, G., Li, Y., Ji, X.: CATRE: iterative point clouds alignment for category-level object pose refinement. In: ECCV (2). pp. 499–516 (2022)
25. Lunayach, M., Zakharov, S., Chen, D., Ambrus, R., Kira, Z., Irshad, M.Z.: FSD: fast self-supervised single RGB-D to categorical 3d objects. CoRR **abs/2310.12974** (2023)
26. Marchand, É., Uchiyama, H., Spindler, F.: Pose estimation for augmented reality: A hands-on survey. IEEE Trans. Vis. Comput. Graph. **22**(12), 2633–2651 (2016)
27. Murtagh, F., Legendre, P.: Ward's hierarchical agglomerative clustering method: Which algorithms implement ward's criterion? J. Classif. **31**(3), 274–295 (2014)
28. Nie, Y., Han, X., Guo, S., Zheng, Y., Chang, J., Zhang, J.: Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In: CVPR. pp. 52–61 (2020)
29. Peng, W., Yan, J., Wen, H., Sun, Y.: Self-supervised category-level 6d object pose estimation with deep implicit shape representation. In: AAAI. pp. 2082–2090. AAAI Press (2022)
30. Rad, M., Lepetit, V.: BB8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: ICCV. pp. 3848–3856 (2017)
31. Shotton, J., Glocker, B., Zach, C., Izadi, S., Criminisi, A., Fitzgibbon, A.W.: Scene coordinate regression forests for camera relocalization in RGB-D images. In: CVPR. pp. 2930–2937 (2013)
32. Song, C., Song, J., Huang, Q.: Hybridpose: 6d object pose estimation under hybrid representations. In: CVPR. pp. 428–437 (2020)
33. Su, Y., Rambach, J.R., Minaskan, N., Lesur, P., Pagani, A., Stricker, D.: Deep multi-state object pose estimation for augmented reality assembly. In: ISMAR Adjunct. pp. 222–227. IEEE (2019)

34. Tian, M., Ang, M.H., Lee, G.H.: Shape prior deformation for categorical 6d object pose and size estimation. In: ECCV (21). pp. 530–546 (2020)
35. Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. In: CoRL. pp. 306–316 (2018)
36. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Trans. Pattern Anal. Mach. Intell. **13**(4), 376–380 (1991)
37. Wang, C., Martín-Martín, R., Xu, D., Lv, J., Lu, C., Fei-Fei, L., Savarese, S., Zhu, Y.: 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In: ICRA. pp. 10059–10066 (2020)
38. Wang, G., Manhardt, F., Liu, X., Ji, X., Tombari, F.: Occlusion-aware self-supervised monocular 6d object pose estimation. CoRR **abs/2203.10339** (2022)
39. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: CVPR. pp. 16611–16621 (2021)
40. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: CVPR. pp. 2642–2651 (2019)
41. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., Lin, D., Liu, Z.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: CVPR. pp. 803–814 (2023)
42. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In: Robotics: Science and Systems (2018)
43. Zakharov, S., Shugurov, I., Ilic, S.: DPOD: dense 6d pose object detector in RGB images. CoRR **abs/1902.11020** (2019)
44. Ze, Y., Wang, X.: Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and A new dataset. In: NeurIPS (2022)
45. Zhang, K., Fu, Y., Borse, S., Cai, H., Porikli, F., Wang, X.: Self-supervised geometric correspondence for category-level 6d object pose estimation in the wild. In: ICLR. OpenReview.net (2023)
46. Zhang, R., Di, Y., Lou, Z., Manhardt, F., Tombari, F., Ji, X.: Rbp-pose: Residual bounding box projection for category-level pose estimation. In: ECCV (1). pp. 655–672 (2022)
47. Zheng, L., Wang, C., Sun, Y., Dasgupta, E., Chen, H., Leonardis, A., Zhang, W., Chang, H.J.: Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In: CVPR. pp. 17163–17173 (2023)