SecondPose: SE(3)-Consistent Dual-Stream Feature Fusion for Category-Level Pose Estimation

Yamei Chen[1,*], Yan Di[1,*], Guangyao Zhai[1,2,†], Fabian Manhardt[3], Chenyangguang Zhang[4], Ruida Zhang[4], Federico Tombari[1,3], Nassir Navab[1] and Benjamin Busam[1,2,5]

[1] Technical University of Munich [2] Munich Center for Machine Learning
[3] Google [4] Tsinghua University [5] 3dwe.ai

https://github.com/NOrangeeroli/SecondPose.git

Abstract

Category-level object pose estimation, aiming to predict the 6D pose and 3D size of objects from known categories, typically struggles with large intra-class shape variation. Existing works utilizing mean shapes often fall short of capturing this variation. To address this issue, we present SecondPose, a novel approach integrating object-specific geometric features with semantic category priors from DINOv2. Leveraging the advantage of DINOv2 in providing SE(3)-consistent semantic features, we hierarchically extract two types of SE(3)-invariant geometric features to further encapsulate local-to-global object-specific information. These geometric features are then point-aligned with DINOv2 features to establish a consistent object representation under SE(3) transformations, facilitating the mapping from camera space to the pre-defined canonical space, thus further enhancing pose estimation. Extensive experiments on NOCS-REAL275 demonstrate that SecondPose achieves a 12.4% leap forward over the state-of-the-art. Moreover, on a more complex dataset HouseCat6D which provides photometrically challenging objects, SecondPose still surpasses other competitors by a large margin.

1. Introduction

Category-level pose estimation involves estimating the complete 9 degrees-of-freedom (DoF) object pose, encompassing 3D rotation, 3D translation, and 3D metric size, for arbitrary objects within a known set of categories. This task has garnered significant research interest due to its essential role in various applications, including the AR/VR industry [30, 38, 40, 55], robotics [51, 52, 54], and scene understanding [1, 7, 53]. In contrast to traditional instance-level pose estimation methods [6,21], which rely on specific 3D CAD models for each target object, the category-level approach necessitates greater adaptability to accommodate inherent shape diversity within each category. Effectively addressing intra-class shape variations has thus become a central focus, crucial for real-world applications where objects within a category may exhibit significant differences in shape while sharing the same general category label.

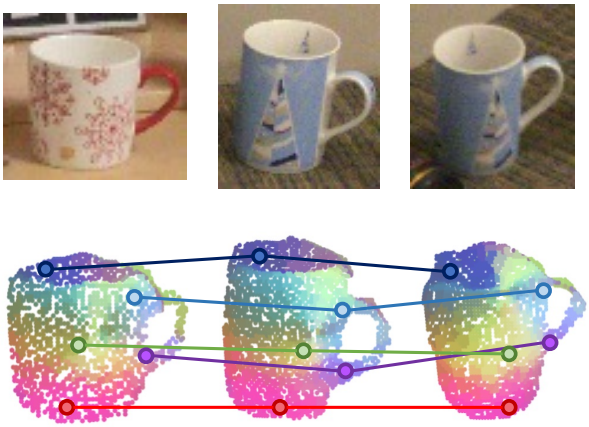


Figure 1. **Categorical SE(3)-consistent features.** We visualize our fused features by PCA. Colored points highlight the most corresponding parts, where our proposed feature achieves consistent alignment cross instances (left vs. middle) and maintains consistency on the same instance of different poses (middle vs. right).

Mean Shape *vs.* Semantic Priors. One common approach to handle intra-class shape variation involves using explicit mean shapes as prior knowledge [24, 39, 57]. These methods typically consist of two functional modules: one for reconstructing the target object by slightly deforming the mean shape and another for regressing the 9D pose based on the reconstructed object [24,39] or enhanced inter-

* Equal contributions.
† Corresponding author (e-mail: guangyao.zhai@tum.de).

mediate features [57]. These methods assume that the mean shape can perfectly encapsulate the structural information of objects within each category, thus achieving reconstruction of the target object with minimal deformation is feasible. However, this assumption does not hold in reality. Objects within the same category, such as chairs, may have fundamental structural differences, leading to the failure of such methods.

Recently, self-supervised learning with large vision models has experienced a significant leap forward, among which DINOv2 [32], due to its exceptional performance in providing semantically consistent patch-wise features, has gained great attention. In particular, various methods [56] utilize semantic features from DINOv2 as essential priors to understand the object. In the field of pose estimation, compared to category-specific mean shapes, DINOv2 demonstrates superior generalization capabilities in object representation across each category, thanks to its large-scale training data and advanced training strategy. ZSP [12] directly leverage DINOv2 features for zero-shot construction of semantic correspondences between objects under different camera viewpoints, and then estimates the pose with RANSAC. POPE [10] and CNOS [31] harness DINOv2 to refine the object detection, thus implicitly boosting the accuracy of pose estimation. However, to our knowledge, currently there exists no method that explores how to fuse DINOv2 features with object-specific features to directly enhance the performance of category-level pose estimation.

In this paper, we present **SecondPose**, a novel method that fuses **SE**(3)-**Con**sistent **D**ual-stream features to enhance category-level **Pose** estimation. Leveraging DINOv2's patch-wise SE(3)-consistent semantic features, we extract two types of SE(3)-invariant geometric features—pair-wise distance and pair-wise angles—to encapsulate object-specific cues. We hierarchically aggregate geometric features within support regions of increasing radius to encode local-to-global object structure information. These features are then point-aligned with DINOv2 features to establish a unified object representation that is consistent under SE(3) transformations. Specifically, given an RGB-D image capturing the target object, we first back-project the depth map to generate the respective point cloud, which is then fed into our *Geometric and Semantic Streams* (Fig. 2.A-B) to extract the corresponding features for our dual-stream fusion (Fig. 2.C). The fused features denoted as **SECOND** are finally fed into an off-the-shelf *pose estimator* [27] (Fig. 2.D) to regress the 9D pose.

SE(3)-Consistent Fusion *vs.* Direct Fusion. Alternatively, one could think of directly concatenating DINOv2 features with the back-projected point in a point-wise manner, without extracting SE(3)-invariant geometric features. However, our instead proposed SE(3)-consistent fusion holds two important advantages over such a straightforward approach. First, while DINOv2 is trained solely with RGB images, the incorporation of geometric features from the point cloud enriches it with valuable local-to-global 3D structural information. This enrichment proves particularly advantageous in handling diverse object shapes within a given category. Second, our SE(3)-consistent object representation modifies the underlying pose estimation process from $\{$*point cloud* $\longrightarrow$ *canonical space*$\}$ to $\{$*point cloud* $\longrightarrow$ *SE(3)-consistent representation* $\longrightarrow$ *canonical space*$\}$. In this optimized pipeline, the second stage – transitioning from our object representation to the human-defined canonical space – is consistent under SE(3) transformations. (approximately invariant, see Fig. 4) This consistency significantly simplifies the pose estimation process, as the pose estimator only needs to operate within the second stage. Further, this streamlined approach not only enhances the accuracy of pose estimation but also contributes to the efficiency of the overall method.

To summarize, our main contributions are threefold:

1. We present **SecondPose**, the first method to directly fuse object-specific hierarchical geometric features with semantic DINOv2 features for category-level pose estimation.
2. Our SE(3)-consistent dual-stream feature fusion strategy yields a unified object representation that is robust under SE(3) transformations, better suited for downstream pose estimation.
3. Extensive evaluation proves that our SE(3)-consistent fusion strategy significantly boosts pose estimation performance even under severe occlusion and clutter, enabling real-world applications.

2. Related Works

Instance-Level Pose Estimation Instance-level pose estimation focuses on determining the 3D rotation and 3D translation of known objects given their 3D CAD models. Recent methods can be mainly categorized into three types: direct pose regression [20, 48], methods that establish 2D-3D correspondences through keypoint detection or pixel-wise 3D coordinate estimation [33, 43, 50], and approaches that learn pose-sensitive embeddings for subsequent pose retrieval [37]. While most keypoints based approaches rely on the P*n*P algorithm [33, 36, 50] to solve for pose, some methods instead employ neural networks to learn the optimization step [43]. As for RGB-D input, traditional methodologies often rely on hand-crafted features [9,16]. Some more recent approaches [4,14,15,42,47] instead extract features independently from RGB images and point clouds, using dedicated CNNs and point cloud networks. These individual features are then fused for direct pose regression [4,42] or keypoint detection [14,15,47]. Despite significant progress, practical applications of these

methods remain limited due to their restriction to a few objects and the need for 3D CAD models.

Category-Level Pose Estimation In the domain of category-level pose estimation, the objective encompasses predicting the 9DoF pose for any object, regardless if previously seen or novel, from a predefined set of categories. This task is inherently more complex due to significant intra-class variations in shape and texture. To address these challenges, Wang et al. [44] developed the Normalized Object Coordinate Space (NOCS), offering a unified representation framework. This approach involves mapping the observed point cloud to the NOCS system, followed by pose recovery via the Umeyama algorithm [41]. Alternatively, CASS [2] introduces a learned canonical shape space, while FS-Net [5] advocates for a decoupled representation of rotation, focusing on direct pose regression. DualPoseNet [26] employs dual networks for both explicit and implicit pose prediction, ensuring consistency for refined pose estimation. GPV-Pose [8] and OPA-3D [35] leverage geometric insights in bounding box projection to augment the learning of pose-sensitive features specific to categories. HS-Pose [59] proposed the HS-layer, a simple network structure that extends 3D graph convolution to extract hybrid scope latent features from point cloud data. In contrast, 6-PACK [45] conducts pose tracking by means of semantic keypoints, and CAPTRA [46] combines coordinate prediction with direct regression for enhanced accuracy. SelfPose [49] utilizes optical flow to enhance the pose estimation accuracy.

To address the issue of intra-class shape variations, several works have focused on the incorporation of additional shape priors. SPD [39] utilizes a PointNet autoencoder to derive a prior point cloud for each category, representing the average shape. This model is then adapted to fit specific observed instances, assigning the observed point cloud to the reconstructed shape model. SGPA [3] dynamically adjusts the shape prior based on structural similarities of the observed instances. SAR-Net [22], while also employing shape priors, further leverages geometric attributes of objects to enhance performance. ACR-Pose [11], instead utilizes a shape prior-guided reconstruction network paired with a discriminator to achieve high-quality canonical representations.

Furthermore, recent research has introduced prior-free methods that demonstrate performance comparable to approaches relying on priors. VI-Net [27] attains high precision in object pose estimation by separating rotation into viewpoint and in-plane rotations. Additionally, IST-Net [28] achieves state-of-the-art performance on the REAL275 benchmark by implicitly transforming camera-space features to world-space counterparts without depending on priors.

3. Method

The objective of SecondPose is to estimate the 9DoF object pose from a single RGB-D image. In particular, given an RGB-D image capturing the target object from a set of known categories, our goal is to recover its full 9DoF object pose, including the $\boldsymbol{R} \in SO(3)$ and the 3D translation $t \in \mathbb{R}^3$ and the 3D metric size $s \in \mathbb{R}^3$.

3.1. Overview.

As illustrated in Fig 2, SecondPose mainly consists of 3 modules to predict object pose from a single RGB-D input, *i.e.* i) the extraction of relevant geometric features F_g and semantic features F_s, ii) the dual-stream feature fusion to build our SE(3)-consistent object representation F_f, iii) the final pose regression from the extracted representation.

3.2. Semantic Category Prior From DINOv2

DINOv2 is an implicit rotation learner We use DINOv2 [32] as our image feature extractor. As shown in [56], DINOv2 can extract semantic-aware information from RGB images that can be well leveraged to establish zero-shot semantic correspondences, rendering it an excellent method for rich semantic information extraction.

As for estimating the 3D rotation, such extra semantic-aware information can provide a noticeable boost in performance. Exemplary, imagine that the z-axis commonly points to the top side of the object in model space, the y-axis always points to the front side of the object, and the x-axis always points to the left side of the object. Harnessing the semantic information given by DINOv2, the model can more easily identify the top, front, and left sides of the object, thus turning rotation estimation into a much simpler task. Moreover, DINOv2 features additionally contain global information about the object, including the object category and pose. Such information can thus serve as a good global prior to our method.

Deeper DINOv2 features We use the "token" facet from the last (11th.) layer as our extracted semantic feature. Essentially, [56] has demonstrated that the features of deeper layers exhibit optimal semantic matching performance, thus providing improved consistency in terms of semantic correspondence across different objects. In addition, features from deeper layers also possess more holistic semantic information. A visualization piece is shown in Fig. 2.A.

Direct pose estimation from DINOv2 As aforementioned, the ad-hoc fusion of DINOv2 features with the back-projected points exhibits several downsides. First, DINOv2 extracts information only from RGB images; hence, the contained geometric information is limited. Second, as we make use of deeper-layer features from DINOv2 for a

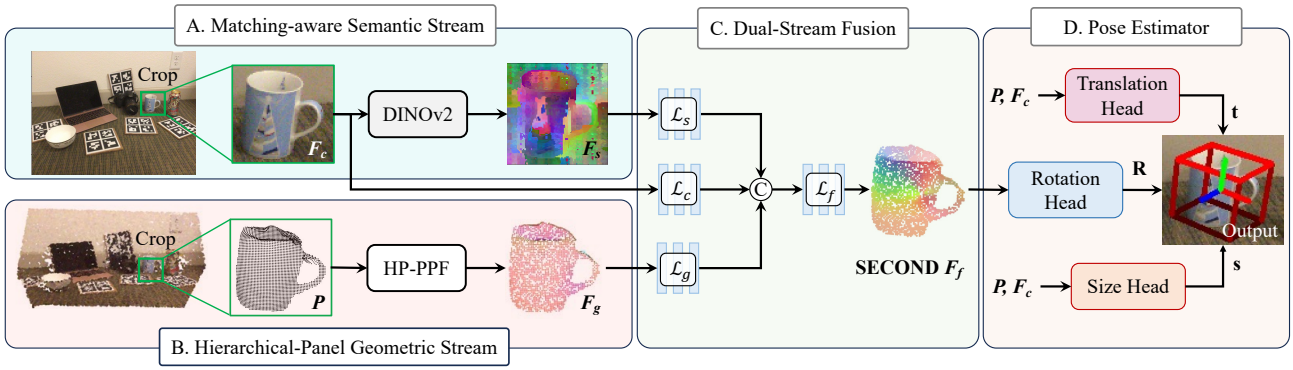
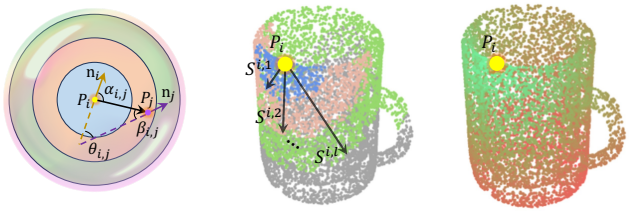


Figure 2. **Illustration of SecondPose.** Semantic features are extracted using the DINOv2 model (A), and the HP-PPF feature is computed on the point cloud (B). These features, combined with RGB values, are fused into our SECOND feature F_f (C) using stream-specific modules L_s, L_g, L_c, and a shared module L_f for concatenated features. The resulting fused features, in conjunction with the point cloud, are utilized for pose estimation (D).



a) Point Pair Features b) Hierarchical Panels c) HP-PPF

Figure 3. **Hierarchical panel-based geometric features.** The inner panel contains points that are close to the point of interest, and outer panels contain points far from the point of interest.

more holistic representation, the local detailed information is blurred to some extent. To complement DINOv2 features in these aspects, we thus need to combine them with geometric features containing local information for better descriptive power.

3.3. Hierarchical Geometric Features

The stream pipeline is shown in Fig. 2.B. Our geometric embedding in this stream is based on the calculation of pair-wise SE(3)-equivariant Point Pair Features (PPFs) [9]. We construct our SE(3)-invariant coordinate representation by aggregating the PPFs between the point of interest and neighborhood points in the multiple panels centered on it. We hierarchically concatenate the corresponding SE(3)-invariant coordinate representations in each panel to enrich the representation power of our geometric features HP-PPF. Fig. 3.c provides a visualization of HP-PPF.

Point Pair Features PPFs A comprehensive example is shown in Fig. 3.a. Given an object point cloud denoted as $\boldsymbol{P}$, we consider each pair of points (p_i, p_j) where $p_i, p_j \in \boldsymbol{P}$. Associated with each point, local normal vectors $\boldsymbol{n_i}$ and $\boldsymbol{n_j}$ are computed at each p_i and p_j, respectively. The final pairwise feature between p_i and p_j is defined as

$$f_{i,j} = [d_{i,j}, \alpha_{i,j}, \beta_{i,j}, \theta_{i,j}], \tag{1}$$

where $d_{i,j} = \|p_j - p_i\|$ describes the Euclidean distance between points p_i and p_j. $\alpha_{i,j} = \angle(\boldsymbol{n_i}, p_j - p_i)$ represents the angular deviation between the normal vector $\boldsymbol{n_i}$ at point p_i and the vector extending from p_i to p_j. $\beta_{i,j} = \angle(\boldsymbol{n_j}, p_j - p_i)$ denotes the angle subtended by the normal vector $\boldsymbol{n_j}$ at point p_j with the aforementioned vector from p_i to p_j. $\theta_{i,j} = \angle(\boldsymbol{n_j}, \boldsymbol{n_i})$ denotes the angular disparity between the normal vectors $\boldsymbol{n_j}$ and $\boldsymbol{n_i}$ at points p_j and p_i, respectively. Notice that thanks to its locality, this descriptor is invariant under $SE(3)$.

Geometric Feature Panel Based on PPFs, we propose panel-based PPFs to construct our geometric representation, which increases the perception field while maintaining the merit of the locality. For each point p_i in the point cloud $\boldsymbol{P}$, there is a support panel $\mathcal{S}^i \subseteq \boldsymbol{P}$ whose cardinality $s_i = |\mathcal{S}^i|$. For all points $p_j \in \mathcal{S}^i$, we calculate the PPF $f_{i,j}$ between p_i, p_j and the local coordinate representation f_l^i of p_i is then obtain as average the average according to

$$f_l^i = \frac{1}{s_i}\left(\sum_j d_{i,j}, \sum_j \alpha_{i,j}, \sum_j \beta_{i,j}, \sum_j \theta_{i,j}\right). \tag{2}$$

From Single to Hierarchical Panels Even though the mean aggregation in the panel can take the neighboring points into account, the inherent local representation limits its representational power, as the features brought by normals $\boldsymbol{n_i}, \boldsymbol{n_j}$ are noisy when constraining the perception field. Inspired by CNNs, which extract hierarchical features from local to global, we hierarchically sample multiple panels from local to global, as shown in Fig. 3.b. Specifically, for a point set $\boldsymbol{P}$ with cardinality $|\boldsymbol{P}|$, for integers

$(k_0, k_1, k_2, ..., k_l)$ satisfying $0 = k_0 < k_1 < k_2 < ... < k_l = |\boldsymbol{P}| - 1$, for each point $p_i \in \boldsymbol{P}$ we first rank its distance to any other points in $\boldsymbol{P}$ from smallest to largest:

$$r_{i,j} = sort(d_{i,j}) \tag{3}$$

and construct support panels:

$$\mathcal{S}^{i,m} = \{p_j \in \boldsymbol{P} | k_{m-1} < r_{i,j} \leq k_m\}, 1 \leq m \leq l, \tag{4}$$

with l being the number of employed panels. We then calculate the corresponding pose-invariant coordinate representations $f^{i,m}$ for each panel $\mathcal{S}^{i,m}$ and concatenate them to get the point-wise geometric features with

$$f_g^i = f_l^{i,1} \oplus f_l^{i,2} \oplus ... \oplus f_l^{i,l}. \tag{5}$$

Thereby, for smaller k, the support panel is composed of points that are closer to the point of interest, whereas for larger k, the support panel consists of points that are farther from the point of interest. By concatenating features calculated by panels of different scales, we can harness geometric features in a way that balances details of local geometric landscapes and global instance-wise shape information. We experimentally show in Sec. 4 that our design performs better than the usual single-panel descriptor.

3.4. SE(3)-Consistent Feature Fusion

Fusion Strategy We fuse the DINOv2 features, the geometric feature and RGB values, as shown in Fig. 2.C. In particular, we use VI-Net [27] as an example of the pose estimator, first projecting each feature to each feature stream $\mathcal{F}$ and 3D point cloud $P = \{p_i\}$ to a spherical feature map F. To this end, we divide the sphere uniformly into $W \times H$ along the azimuth and elevation axes, following VI-Net [27]. We assign the feature of the point with the largest distance to each bin. When there is no point in the region, we set 0 in the bin. For each feature map $F_i \in \{F_g, F_s, F_c\}$ representing the geometric feature, the DINOV2 feature, and the respective RGB value, we employ a separate ResNet model $\mathcal{L}_i$ as feature extractor. The outputs of these individual feature extractors are then concatenated to form the input to another ResNet for feature fusion, obtaining F_f also denoted as **SECOND**,

$$F_f = \mathcal{L}_f \left(\mathcal{L}_g(F_g) \oplus \mathcal{L}_s(F_s) \oplus \mathcal{L}_c(F_c) \right). \tag{6}$$

Advantages of SE(3)-Consistent Fusion The Design of a SE(3)-consistent fusion is an integral part of the improved quality of our method. As for the 3D rotation, we are learning a mapping from the space of point clouds and its features $(P, F) \in \mathbb{R}^{n \times 3} \times \mathbb{R}^{n \times C}$ to space of 3D rotations $\boldsymbol{R} \in SO(3)$

$$\Phi : \mathbb{R}^{n \times 3} \times \mathbb{R}^{n \times C} \mapsto SO(3). \tag{7}$$

This mapping Φ should ensure rotation-equivariance, meaning that

$$\Phi(R_x P, \psi_{R_x}(F)) = R_x \Phi(P, F), \forall R_x \in SO(3), \tag{8}$$

where ψ_{R_x} is the transformation applied to the feature when rotating the point cloud by R_x. This rotation-equivariance relation is essential for the learned model to generalize well on unseen data. Without such equivariance embedded in the model structure, these relation needs to be learned through large amounts of data, which is limited by the scale of the data. Our design of SE(3)-consistent features are approximately rotation-invariant, hence

$$\psi_{R_x}(F) \approx F, \forall R_x \in SO(3), \tag{9}$$

eliminating the effect of ψ_{R_x} in Eq. (8), and thus making learning of the rotation-equivariance relationship easier.

3.5. SecondPose Training and Inference

Following [27], we leverage a lightweight PointNet++ [34] as the translation and size estimation heads. Given an RGB-D image, we first segment the object of interest using Mask-RCNN [13], similar to [8, 27]. We then randomly select N points from the back-projected 3D point clouds $\boldsymbol{P} \in \mathbb{R}^{n \times 3}$ with RGB features F_c and use them to estimate the translation and size, as shown in Fig. 2.D.

The core of our method is thus developed to focus on the more challenging task of 3D rotation estimation. We essentially train a separate translation-size network and rotation network. For the translation-size network, we adopt the L1 loss for both size and translation with

$$L_{ts} = \lambda_t |t_{pred} - t_{gt}| + \lambda_s |s_{pred} - s_{gt}|. \tag{10}$$

For the 3D rotation, we instead directly predict the 9D rotation matrix, which we optimize via the L1-loss according to

$$L_R = |R_{pred} - R_{gt}|. \tag{11}$$

During training, the ground truth translation and size are used to center and normalize the point cloud before rotation estimation, while during inference the predicted size and translation are instead utilized for normalization.

4. Experiment

4.1. Experimental Setup.

Datasets We conduct our experiments on the common 9D pose estimation benchmarks NOCS-REAL275 [44], NOCS-CAMERA25 [44] as well as HouseCat6D [19] datasets. NOCS-REAL275 is a real-world dataset with 13 scenes containing objects from 6 different categories; 4,300 images of 7 scenes are used as a training set, while the other 2,750 images of 6 scenes form the test set. NOCS-CAMERA25 is a synthetic dataset containing 300k images

Method Name	Mean Shape Priors	REAL275				
		IoU $_{75}$*	5° 2 cm	5° 5 cm	10° 2 cm	10° 5 cm
SPD [39]	✓	27.0	19.3	21.4	43.2	54.1
CR-Net [45]	✓	33.2	27.8	34.3	47.2	60.8
CenterSnap-R [17]	✓	-	-	29.1	-	64.3
ACR-Pose [11]	✓	-	31.6	36.9	54.8	65.9
SAR-Net [22]	✓	-	31.6	42.3	50.3	68.3
SSP-Pose [58]	✓	-	34.7	44.6	-	77.8
SGPA [3]	✓	37.1	35.9	39.6	61.3	70.7
RBP-Pose [57]	✓	-	38.2	48.1	63.1	79.2
SPD + CATRE [29]	✓	43.6	45.8	54.4	61.4	73.1
DPDN [25]	✓	-	46.0	50.7	70.4	78.4
FS-Net [5]	×	-	-	28.2	-	60.8
DualPoseNet [26]	×	30.8	29.3	35.9	50.0	66.8
GPV-Pose [8]	×	-	32.0	42.9	-	73.3
SS-ConvNet [23]	×	-	36.6	43.4	52.6	63.5
HS-Pose [59]	×	-	46.5	55.2	68.6	82.7
IST-Net [28]	×	-	47.5	53.4	72.1	80.5
VI-Net [27]	×	48.3	50.0	57.6	70.8	82.1
SecondPose (Ours)	Semantic Priors	**49.7**	**56.2**	**63.6**	**74.7**	**86.0**

Table 1. Quantitative comparisons of different methods for category-level 6D object pose estimation on REAL275 [44]. '*' denotes the CATRE [29] IoU metrics. The best results are in bold, and the second best results are underlined.

with objects from the same categories as NOCS-REAL275. HouseCat6D is a comprehensive multi-modal real-world dataset, featuring 194 high-fidelity 3D models of household items of 10 categories. The collection encompasses transparent and reflective objects situated in 41 scenes, presenting a wide range of viewpoints, challenging occlusions, and devoid of markers.

Evaluation Metrics As for the NOCS-REAL275 dataset, we report the mean Average Precision (mAP) of $5°2\text{cm}, 5°5\text{cm}, 10°2\text{cm}, 10°5\text{cm}$ metrics. $n°m\text{cm}$ denotes the percentage of prediction with rotation prediction error within n degrees and translation prediction error within m centimeters. We also report mAP of 3D Intersection over Union (IoU) at the threshold of 75%. For the HouseCat6D dataset, we again report the mAP of 3D IoU under thresholds of 25% and 50%.

Efficiency Our method achieves an inference speed of 9 FPS. Excluding the running time of DINOv2, our inference speed increases to 10 FPS.

Implementation Details We use MaskRCNN [13] to segment the objects of interest from the input image. We then combine point-wise radial distances, RGB values, and semantic-aware features from DINOv2 together with our proposed local-to-global SE(3)-invariant geometric features as input for further processing. Next, for the RGB values and the point-wise radial distances, we sample 2048 points from the point cloud. For DINOv2 features, we first crop the image by the bounding box around the object of interest and then resize the image to a resolution of 210×210. Finally, for our geometric features, we sample 300 points from previously sampled 2048 points and estimate point-wise normal vectors using the 10 nearest neighbors. To train our model on the NOCS dataset, we use a mixture of 25% real-world images from the training set of REAL 275 and 75% synthetic images from the CAMERA25 training set, similar to [44]. For all experiments, we train our models with batch size 48 on a single NVIDIA 3090 GPU to the 40th. epoch.

4.2. Comparison with State-of-the-Art Methods

In Tab. 1, we compare SecondPose with the state-of-the-art on NOCS-REAL275 dataset. As can be easily observed, our method outperforms all state-of-the-art approaches, including the recent VI-Net [27], by a large margin on all metrics. More specifically, our method respectively exceeds VI-Net for $5°2\text{ cm}$ and $10°2\text{ cm}$ by 6.2% and 3.9%, demonstrating the effectiveness of our SE(3)-consistent feature fusion design. When comparing with DPDN [24], the best method using mean shape prior, our improvements in $5°2\text{ cm}$ and $5°5\text{ cm}$ metrics amount to 10.2% and 12.8%. We show qualitative results in Figure 4. It can be observed that SecondPose is more robust when handling objects with large intra-class variations, such as *camera*. In Tab. 2, we evaluate our method on the HouseCat6D dataset.

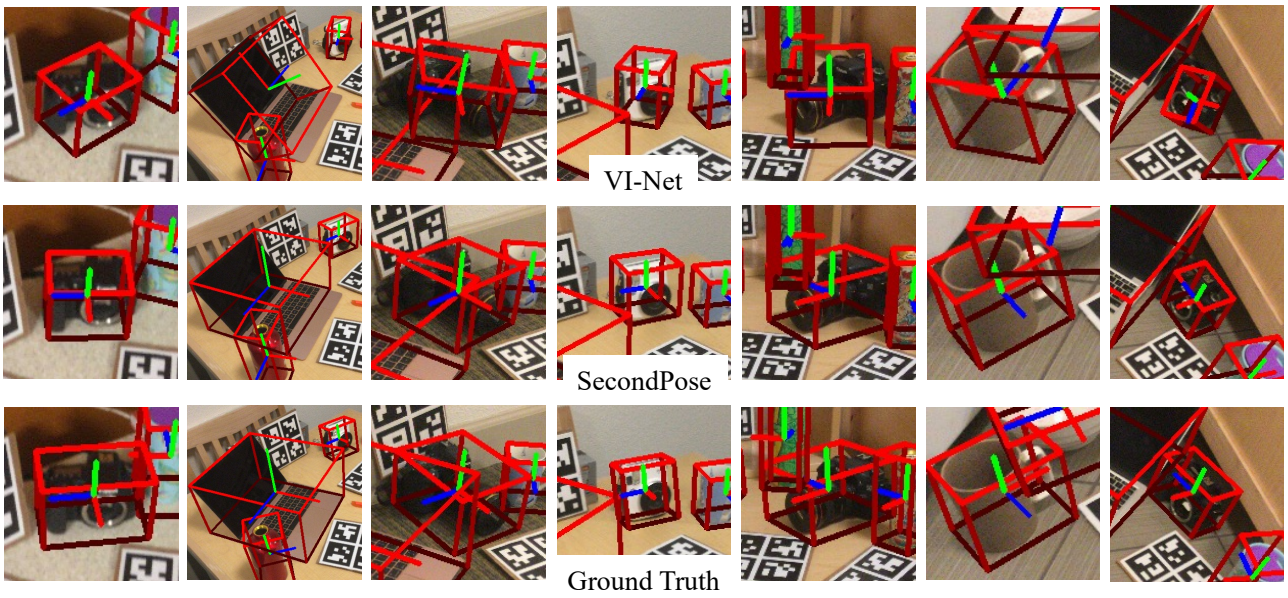


Figure 4. Qualitative comparison on REAL275 [44]. We compare our prediction with ground truth and the prediction of our baseline, VI-Net [27]. Our approach achieves significantly higher precision in rotation estimation.

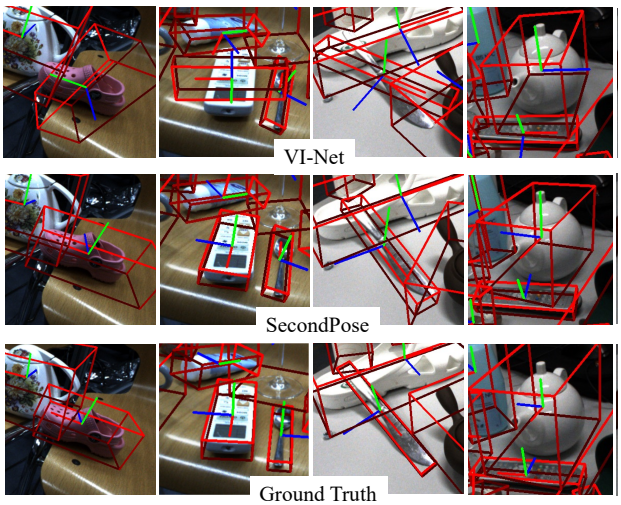


Figure 5. Qualitative comparison on HouseCat6D [18]. We compare our prediction with ground truth and the prediction of our baseline, VI-Net [27].

Our method can again exceed current state-of-the-art methods by a large margin. As for the IoU_{50} metric, our method outperforms the second-best method VI-Net by 9.7% on average. Additional qualitative results can be found in Fig.

4.3. Limitations

Our method's efficacy is restricted by the constraints of DINOv2 due to our utilization of its features. When DINOv2 is unable to provide meaningful semantic information for specific images, our approach is unable to surpass this limitation.

4.4. Ablation Studies

To confirm the efficacy of our design choices, we conduct several ablation studies on the NOCS-REAL275 [44] dataset.

[AS-1] Efficacy of employing semantic and geometric features. To show the effectiveness of our semantic-geometric-feature-fusion, we train the proposed model in 3 different variations: i) without semantic feature, ii) without geometric feature, and iii) without both semantic and geometric features. The results are presented in Tab. 3 (B0) - (B2). When considering the strict $5°2\text{cm}$ metric, it turns out that removing semantic features, geometric features or both always leads to a large decrease in performance. In particular, the performance respectively drops by 5.1%, 1.1% and 6.3%.

[AS-2] Efficacy of individual geometric feature. We further run ablations on the four geometric features, d, α, β, θ. The corresponding results are presented again in Table 3 under (C0) - (C3). As can be observed, removing any component from the geometric feature leads to a strict drop in performance. Exemplary, for the $5°2\text{cm}$ metric the performance drops by at least 1.1%. To summarize, each geometric feature contributes to the expressiveness of the geometric representation.

[AS-3] Efficacy of hierarchical panel construction. As shown in Tab. 3 (D0), when the hierarchical panel is substituted by KNN with 10 nearest neighbors, the $5°2\text{cm}$ metric undergoes a decrease by 0.8%. This demonstrates the im-

Approach	IoU_{25} / IoU_{50}	Bottle	Box	Can	Cup	Remote	Teapot	Cutlery	Glass	Tube	Shoe
NOCS [44]	50.0 / 21.2	41.9 / 5.0	43.3 / 6.5	81.9 / 62.4	68.8 / 2.0	81.8 / 59.8	24.3 / 0.1	14.7 / 6.0	95.4 / 49.6	21.0 / 4.6	26.4 / 16.5
FS-Net [5]	74.9 / 48.0	65.3 / 45.0	31.7 / 1.2	98.3 / 73.8	96.4 / 68.1	65.6 / 46.8	69.9 / 59.8	71.0 / 51.6	99.4 / 32.4	79.7 / 46.0	71.4 / 55.4
GPV-Pose [8]	74.9 / 50.7	66.8 / 45.6	31.4 / 1.1	98.6 / 75.2	96.7 / 69.0	65.7 / 46.9	75.4 / 61.6	70.9 / 52.0	99.6 / 62.7	76.9 / 42.4	67.4 / 50.2
VI-Net [27]	80.7 / 56.4	90.6 / 79.6	44.8 / 12.7	**99.0** / 67.0	96.7 / 72.1	**54.9** / 17.1	52.6 / 47.3	89.2 / **76.4**	99.1 / **93.7**	**94.9 / 36.0**	85.2 / 62.4
SecondPose (Ours)	**83.7 / 66.1**	**94.5 / 79.8**	**54.5 / 23.7**	98.5 / **93.2**	**99.8 / 82.9**	53.6 / **35.4**	**81.0 / 71.0**	**93.5** / 74.4	**99.3** /92.5	75.6 / 35.6	**86.9 / 73.0**

Table 2. Overall and class-wise evaluation of 3D IoU(at 25%, 50%) on the dataset HouseCat6D [18]. The best results are in bold.

Row	Method	IoU_{75}*	5° 2 cm	5° 5 cm	10° 2 cm	10° 5 cm
A0	SecondPose (baseline)	49.7	56.2	63.6	74.7	86.0
B0	w/o semantic	48.0	51.1	58.9	71.6	82.4
B1	w/o geometric	49.5	55.1	62.3	73.7	84.8
B2	w/o semantic+geometric	48.5	49.9	57.4	70.4	80.8
C0	w/o d in Eq. 1	49.1	55.1	63.1	73.7	85.0
C1	w/o α in Eq. 1	49.3	54.7	62.8	73.1	84.7
C2	w/o β in Eq. 1	49.6	54.8	62.7	74.6	86.7
C3	w/o θ in Eq. 1	49.5	55.1	63.1	74.2	85.6
D0	KNN Panel (10 nearest neighbors)	49.4	55.4	63.1	73.7	85.5
E0	random rotation 5°	49.7	56.1	63.4	74.6	85.9
E1	random rotation 10°	49.4	55.8	63.5	74.4	85.8
E2	random rotation 15°	48.5	55.4	63.0	73.9	85.4
E3	random rotation 20°	47.9	54.5	62.4	73.2	85.1
F0	manual occlusion n = 16	49.7	56.0	63.6	74.8	86.2
F1	manual occlusion n = 8	49.5	55.7	63.2	74.3	85.6
F2	manual occlusion n = 4	46.7	52.5	60.9	71.5	84.6
G0	random perturbation s = 0.002	49.7	56.1	63.6	74.6	85.8
G1	random perturbation s = 0.005	49.6	55.8	63.4	74.4	86.0
G2	random perturbation s = 0.01	45.9	53.7	62.6	73.4	86.1

Table 3. Ablation Study on REAL275 [44]. '*' denotes the CATRE [29] IoU metrics.

portance of our hierarchical panel construction, as it better captures finer-grained local and global information.

[AS-4] Robustness under random rotation. To show the robustness of our method under random rotation applied on point cloud, we perform experiments on test images when randomly rotating the entire point cloud by rotation angle A $[0°, n°]$, n = 5, 10, 15, 20, see Table 3 (E0) - (E3). The results show that our method performs well under these circumstances.

[AS-5] Robustness under manual occlusions We also perform an additional experiment to show the robustness of our method under various levels of occlusions. We manually mask out the object with different scale of rectangular masks, whose length and width is set to 1/n of the length and width of the original object bounding box. We further run tests with n = 16, 8, 4 in Tab. 3 (F0) - (F2). When undergoing only mild occlusion, *i.e.* $n = 16$, the performance is almost identical to the original result. Moreover, even when dealing with very large occlusions of 1/4 of the size of the object, the performance is still fairly strong with only a small decrease of 3% for IoU_{75}.

[AS-6] Robustness Under Perturbation on Point Cloud. Next, we also evaluate the robustness of our method under random perturbations applied to the point clouds. To this end, we add random noise sampled from a uniform distribution ranging from $-0.5sr$ to $0, 5sr$, where s is the scale factor and r is the average distance of the point cloud to the object center. We test our model with s = 0.002, 0.005,0.01 in Table 3 G0-G2. We observe again that with mild perturbation of s = 0.002, the performance is almost identical to the original result, while with relatively large perturbation of s = 0.01, the performance is still fairly strong with only a small decrease of 3.8% for IoU_{75}.

5. Conclusion

In this paper, we propose SecondPose designing SE(3)-consistent fusion of semantic and geometric features for pose estimation. The two feature streams are proven to complement each other and jointly contribute to improving our method. To confirm the efficacy of our method, we apply our method on the challenging real-world category-level 6D object pose estimation datasets REAL275 and HouseCat6D and exceed the current SOTA by a large margin.

References

[1] Benjamin Busam, Marco Esposito, Simon Che'Rose, Nassir Navab, and Benjamin Frisch. A stereo vision approach for cooperative robotic movement therapy. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 127–135, 2015. 1

[2] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11973–11982, 2020. 3

[3] 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*, pages 2773–2782, 2021. 3, 6

[4] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, and Ales Leonardis. G2l-net: Global to local network for real-time 6d pose estimation with embedding vector features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[5] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1581–1590, 2021. 3, 6, 8

[6] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12396–12405, October 2021. 1

[7] Yan Di, Chenyangguang Zhang, Ruida Zhang, Fabian Manhardt, Yongzhi Su, Jason Rambach, Didier Stricker, Xiangyang Ji, and Federico Tombari. U-red: Unsupervised 3d shape retrieval and deformation for partial point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8884–8895, 2023. 1

[8] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6781–6791, 2022. 3, 5, 6, 8

[9] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *CVPR*, pages 998–1005. Ieee, 2010. 2, 4

[10] Zhiwen Fan, Panwang Pan, Peihao Wang, Yifan Jiang, Dejia Xu, Hanwen Jiang, and Zhangyang Wang. Pope: 6-dof promptable pose estimation of any object, in any scene, with one reference. *arXiv preprint arXiv:2305.15727*, 2023. 2

[11] Zhaoxin Fan, Zhengbo Song, Jian Xu, Zhicheng Wang, Kejian Wu, Hongyan Liu, and Jun He. Acr-pose: Adversarial canonical representation reconstruction network for category level 6d object pose estimation, 2021. 3, 6

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

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 5, 6

[14] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3003–3013, June 2021. 2

[15] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[16] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 858–865. IEEE, 2011. 2

[17] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone, and Zsolt Kira. Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation, 2022. 6

[18] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Guangyao Zhai, Hannah Schieber, Giulia Rizzoli, Pengyuan Wang, Hongcheng Zhao, Lorenzo Garattoni, Sven Meier, Daniel Roth, Nassir Navab, and Benjamin Busam. Housecat6d – a large-scale multi-modal category level 6d object pose dataset with household objects in realistic scenarios, 2023. 7, 8

[19] HyunJun Jung, Guangyao Zhai, Shun-Cheng Wu, Patrick Ruhkamp, Hannah Schieber, Pengyuan Wang, Giulia Rizzoli, Hongcheng Zhao, Sven Damian Meier, Daniel Roth, Nassir Navab, et al. Housecat6d–a large-scale multi-modal category level 6d object pose dataset with household objects in realistic scenarios. *arXiv preprint arXiv:2212.10428*, 2022. 5

[20] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017. 2

[21] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020. 1

[22] Haitao Lin, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue. Sar-net: Shape alignment and recovery network for category-level 6d object pose and size estimation, 2022. 3, 6

[23] Jiehong Lin, Hongyang Li, Ke Chen, Jiangbo Lu, and Kui Jia. Sparse steerable convolutions: An efficient learning of se(3)-equivariant features for estimation and tracking of ob-

ject poses in 3d space. In *Neural Information Processing Systems*, 2021. 6

[24] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks. In *European Conference on Computer Vision*, pages 19–34. Springer, 2022. 1, 6

[25] Jiehong Lin, Zewei Wei, Changxing Ding, and Kui Jia. Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks, 2022. 6

[26] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3560–3569, 2021. 3, 6

[27] Jiehong Lin, Zewei Wei, Yabin Zhang, and Kui Jia. Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14001–14011, 2023. 2, 3, 5, 6, 7, 8

[28] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation, 2023. 3, 6

[29] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. Catre: Iterative point clouds alignment for category-level object pose refinement, 2022. 6, 8

[30] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: a hands-on survey. *IEEE transactions on visualization and computer graphics*, 22(12):2633–2651, 2015. 1

[31] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2134–2140, 2023. 2

[32] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 2, 3

[33] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[34] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 5

[35] Yongzhi Su, Yan Di, Guangyao Zhai, Fabian Manhardt, Jason Rambach, Benjamin Busam, Didier Stricker, and Federico Tombari. Opa-3d: Occlusion-aware pixel-wise aggregation for monocular 3d object detection. *IEEE Robotics and Automation Letters*, 8(3):1327–1334, 2023. 3

[36] Yongzhi Su, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6738–6748, 2022. 2

[37] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018. 2

[38] David Joseph Tan, Nassir Navab, and Federico Tombari. 6d object pose estimation with depth images: A seamless approach for robotic interaction and augmented reality. *arXiv preprint arXiv:1709.01459*, 2017. 1

[39] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *European Conference on Computer Vision*, pages 530–546. Springer, 2020. 1, 3, 6

[40] Henning Tjaden, Ulrich Schwanecke, and Elmar Schomer. Real-time monocular pose estimation of 3d objects using temporally consistent local color histograms. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 1

[41] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 3

[42] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, 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 (CVPR)*, June 2019. 2

[43] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 2

[44] 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 Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 3, 5, 6, 7, 8

[45] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks, 2021. 3, 6

[46] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13209–13218, 2021. 3

[47] Yangzheng Wu, Mohsen Zand, Ali Etemad, and Michael Greenspan. Vote from the center: 6 dof pose estimation in

rgb-d images by radial keypoint voting. In *European Conference on Computer Vision*, pages 335–352. Springer, 2022. 2

[48] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter . Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems*, 2018. 2

[49] Michela Zaccaria, Fabian Manhardt, Yan Di, Federico Tombari, Jacopo Aleotti, and Mikhail Giorgini. Self-supervised category-level 6d object pose estimation with optical flow consistency. *IEEE Robotics and Automation Letters*, 8(5):2510–2517, 2023. 3

[50] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[51] Guangyao Zhai, Xiaoni Cai, Dianye Huang, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. Sg-bot: Object rearrangement via coarse-to-fine robotic imagination on scene graphs. *arXiv preprint arXiv:2309.12188*, 2023. 1

[52] Guangyao Zhai, Dianye Huang, Shun-Cheng Wu, HyunJun Jung, Yan Di, Fabian Manhardt, Federico Tombari, Nassir Navab, and Benjamin Busam. Monograspnet: 6-dof grasping with a single rgb image. In *IEEE International Conference on Robotics and Automation*. IEEE, 2023. 1

[53] Guangyao Zhai, Evin Pinar Örnek, Shun-Cheng Wu, Yan Di, Federico Tombari, Nassir Navab, and Benjamin Busam. Commonscenes: Generating commonsense 3d indoor scenes with scene graphs. In *NeurIPS*, 2023. 1

[54] Guangyao Zhai, Yu Zheng, Ziwei Xu, Xin Kong, Yong Liu, Benjamin Busam, Yi Ren, Nassir Navab, and Zhengyou Zhang. Da^2 dataset: Toward dexterity-aware dual-arm grasping. *RA-L*, 7(4):8941–8948, 2022. 1

[55] Chenyangguang Zhang, Yan Di, Ruida Zhang, Guangyao Zhai, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ddf-ho: Hand-held object reconstruction via conditional directed distance field. *arXiv preprint arXiv:2308.08231*, 2023. 1

[56] Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *arXiv preprint arXiv:2305.15347*, 2023. 2, 3

[57] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation, 2022. 1, 2, 6

[58] Ruida Zhang, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Ssp-pose: Symmetry-aware shape prior deformation for direct category-level object pose estimation, 2022. 6

[59] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Ales Leonardis, Wei Zhang, and Hyung Jin Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation, 2023. 3, 6