Rethinking Correspondence-based Category-Level Object Pose Estimation

Huan Ren[1,2] Wenfei Yang[1,2,3] Shifeng Zhang[4] Tianzhu Zhang[1,2*]
[1]University of Science and Technology of China
[2]National Key Laboratory of Deep Space Exploration, Deep Space Exploration Laboratory
[3]State Key Laboratory of General Artificial Intelligence [4]Sangfor Technologies
rh_hr_666@mail.ustc.edu.cn zhangshifeng@sangfor.com.cn {yangwf,tzzhang}@ustc.edu.cn

Abstract

Category-level object pose estimation aims to determine the pose and size of arbitrary objects within given categories. Existing two-stage correspondence-based methods first establish correspondences between camera and object coordinates, and then acquire the object pose using a pose fitting algorithm. In this paper, we conduct a comprehensive analysis of this paradigm and introduce two crucial essentials: 1) shape-sensitive and pose-invariant feature extraction for accurate correspondence prediction, and 2) outlier correspondence removal for robust pose fitting. Based on these insights, we propose a simple yet effective correspondence-based method called SpotPose, which includes two stages. During the correspondence prediction stage, pose-invariant geometric structure of objects is thoroughly exploited to facilitate shape-sensitive holistic interaction among keypoint-wise features. During the pose fitting stage, outlier scores of correspondences are explicitly predicted to facilitate efficient identification and removal of outliers. Experimental results on CAMERA25, REAL275 and HouseCat6D benchmarks demonstrate that the proposed SpotPose outperforms state-of-the-art approaches by a large margin.

1. Introduction

Category-level object pose estimation is one of the fundamental tasks in robotic vision, which aims to predict the 6D pose and 3D size of arbitrary objects in the given categories. This task has received increasing attention from the research community due to its broad applications in augmented reality [18, 26], robotic manipulation [17, 35], and hand-object interaction [15, 25]. Compared to traditional instance-level approaches [20, 30, 31], category-level approaches do not require a CAD model for each object instance, providing stronger generalization in real-world scenarios.

*Corresponding author. Website: renhuan1999.github.io/SpotPose.

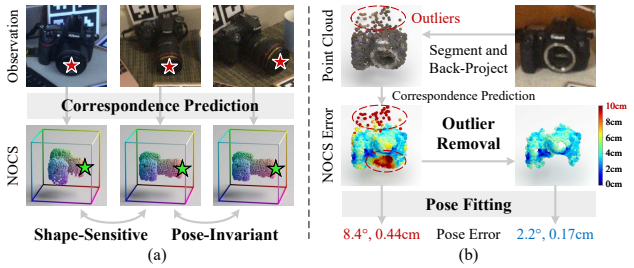

Figure 1. Reconsideration of the two-stage correspondence-based paradigm. (a) Essential of shape-sensitive and pose-invariant features during the correspondence prediction stage. (b) Essential of outlier correspondence removal during the pose fitting stage.

Most existing category-level methods [1, 12, 14, 24, 27] adopt a two-stage correspondence-based paradigm, which first establishes correspondences between the camera coordinate space and the Normalized Object Coordinate Space (NOCS) [32], and subsequently determines the object pose through a pose fitting algorithm. We conduct an in-depth analysis of this paradigm and introduce two crucial insights. **(1) During the correspondence prediction stage, shape-sensitive and pose-invariant features are fundamental.** Since there are significant shape variations among diverse objects within the same category, shape-sensitive features should be exploited to learn object-specific transformation. As shown in Figure 1(a), points on camera lenses of varying lengths ought to be mapped to different NOCS coordinates. Besides, since the specific object might be observed from arbitrary poses, pose-invariant features should be extracted to learn pose-irrelevant transformation. As illustrated in Figure 1(a), a static point on the camera lens under varying poses ought to be mapped to the same NOCS coordinates. **(2) During the pose fitting stage, the removal of outlier correspondences is essential**. Noise in the object segmentation mask and depth camera sensor can yield outlier points within the observed object point cloud. These outlier points, combined with inaccurate correspondence predictions, lead to outlier correspondences that have a detrimental impact on the pose fitting process. Therefore, a robust pose fitting

algorithm is desired to eliminate the interference of outliers. As illustrated in Figure 1(b), the rotation fitting error is reduced from 8.4° to 2.2° after an outlier removal process.

Despite the considerable progress achieved by previous methods, they either overlook these two crucial essentials or instead resort to complicated and inefficient designs. (1) During the correspondence prediction stage, point-wise features of observed objects are extracted and then mapped to the NOCS coordinates. However, existing approaches typically employ general-purpose point cloud backbones without pose-invariance, *e.g.*, the widely deployed PointNet++ [22] incorporates absolute coordinates and yields inherently pose-sensitive features. Furthermore, point-wise features lack sufficient interaction with each other from a holistic perspective, thereby limiting their shape-sensitivity. Consequently, these methods struggle with transformation learning and instead depend on complicated and inefficient designs, *e.g.*, category-level shape priors [1, 12, 27], shape augmentation [34], and auxiliary feature enhancers [16]. (2) During the pose fitting stage, early approaches [27, 32] employ the Umeyama algorithm [28] to solve the object pose with the Random Sample Consensus (RANSAC) algorithm [6] for outlier removal. However, RANSAC requires numerous iterations to find an accurate pose while excluding outliers, leading to slow convergence. Instead, recent methods [12, 14, 16] utilize an MLP-based deep estimator to directly regress the object pose from correspondences, which can benefit from end-to-end supervised training on the final objectives. Nevertheless, they exhibit limited robustness to outlier correspondences, as all correspondences are taken for regression without any filtering.

Motivated by the above discussions, we propose a simple yet effective two-stage correspondence-based approach for category-level object pose estimation, named as **SpotPose**, where **S**hape-sensitive and **P**ose-invariant features are first inherently extracted to facilitate correspondence prediction, followed by efficient identification and removal of **O**u**T**lier correspondences to enable robust and accurate pose fitting. (1) During the correspondence prediction stage, point-wise features are densely extracted from RGB-D observations, where PointNet++ [22] is adapted to extract pose-invariant point cloud features by incorporating a T-Net [21] to align the input point cloud while eliminating the injection of absolute coordinates. Afterwards, we represent object shapes with a set of sparse keypoints, and further leverage the pose-invariant geometric descriptor [23] to facilitate shape-sensitive holistic interaction among keypoint-wise features. Benefiting from the extraction of shape-sensitive and pose-invariant features, transformation between the camera and object coordinate spaces becomes more readily learned. (2) During the pose fitting stage, due to noise in the observations or predictions, there exist outliers in the derived correspondences that are harmful to the pose fitting process. To this end, we just utilize a lightweight outlier predictor to distinguish between inliers and outliers, and then solve the object pose based solely on inlier correspondences with the Umeyama algorithm [28]. Through this intuitive outlier removal process, we achieve comparable pose fitting performance without the need for the RANSAC algorithm [6], while maintaining a fast inference speed.

The main contributions of this work can be summarized as follows: (1) We conduct a comprehensive analysis of the existing two-stage correspondence-based paradigm for category-level object pose estimation and introduce two crucial essentials. (2) We propose a simple yet effective approach which extracts shape-sensitive and pose-invariant features during the correspondence prediction stage to facilitate transformation learning, and efficiently removes outlier correspondences during the pose fitting stage to ensure robust pose estimation. (3) Extensive experimental results on existing benchmarks demonstrate the superior performance of our method over state-of-the-art approaches.

2. Related Work

In this section, we briefly overview direct regression-based and correspondence-based category-level object pose estimation methods.

Direct Regression-based Methods. This group of methods adopts a single-stage paradigm, which directly regresses the object pose in an end-to-end manner after feature extraction from object observations. FS-Net [2] employs a 3D graph convolution (3DGC) autoencoder for orientation feature extraction and decouples the rotation into two orthogonal vectors to handle symmetric objects. GPV-Pose [4] and HS-Pose [37] enhance pose-sensitive feature extraction by focusing on geometric consistency and local-global geometric structure perception, respectively. VI-Net [13] proposes an innovative rotation estimation network that simplifies the task on the sphere by decoupling the rotation into viewpoint and in-plane components, while SecondPose [3] further improves it by leveraging SE(3)-consistent semantics and geometric feature extraction. Although conceptually simple, these methods struggle with pose-sensitive feature learning due to the non-linearity of the full SE(3) space.

Correspondence-based Methods. This group of methods adopts a two-stage paradigm, which first establishes correspondences between camera and object coordinates, and then derives the object pose through a pose fitting algorithm. During the correspondence prediction stage, the Normalized Object Coordinate Space (NOCS) is introduced in [32] as a shared canonical representation to align diverse object instances in a given category. SPD [27] and its subsequent works [1, 12, 33] reconstruct 3D object models by deforming a pre-learned categorical shape prior, and then predict correspondences between the observed and reconstructed object point clouds. However, acquiring the priors requires

collecting numerous CAD models, which is labor-intensive. Query6DoF [34] and IST-Net [16] eliminate the need for explicit shape priors through implicit queries and implicit space transformation, respectively, but still rely on complicated shape augmentation or auxiliary feature enhancers. These methods disregard shape-sensitive and pose-invariant feature extraction, struggling with transformation learning. During the pose fitting stage, early approaches [1, 27, 32] adopt the Umeyama algorithm [28] to solve the object pose, combined with the RANSAC algorithm [6] to mitigate the interference of outlier correspondences. Nevertheless, the RANSAC algorithm requires massive iterations to produce a fitting pose with enough inliers while excluding outliers, which suffers from slow convergence. DPDN [12] presents a pose and size estimator to directly regress the object pose and size from correspondences, which is composed of MLPs and average pooling operations. It can benefit from end-to-end supervised training on the final pose and size objectives, rather than solely on the surrogate correspondences. However, since all correspondences are employed for regression without any filtering, the deep estimator exhibits limited robustness to outlier correspondences.

3. Method

Task Definition. Given an RGB-D image, we first utilize an offline Mask R-CNN [7] to acquire instance segmentation masks, yielding the cropped RGB image $\boldsymbol{I}^{\mathrm{obj}} \in \mathbb{R}^{H\times W\times 3}$ and segmented depth image for each instance. The partially observed object point cloud $\boldsymbol{P}^{\mathrm{obj}} \in \mathbb{R}^{N^{\mathrm{obj}}\times 3}$ is then derived from the segmented depth image by back-projecting and downsampling, where N^{obj} is the number of points. Taking $\boldsymbol{I}^{\mathrm{obj}}$ and $\boldsymbol{P}^{\mathrm{obj}}$ as inputs, our method aims to predict the rotation $\boldsymbol{R} \in \mathrm{SO}(3)$, translation $\boldsymbol{t} \in \mathbb{R}^3$, and size $\boldsymbol{s} \in \mathbb{R}^3$ of the observed instance in the given categories.

Overview. As illustrated in Figure 2, the proposed method consists of three components, which include dense feature extraction (Section 3.1), sparse feature interaction (Section 3.2) and robust pose and size estimation (Section 3.3). In Section 3.4, we further discuss the distinctions between our SpotPose and the most relevant method, AG-Pose [14].

3.1. Dense Feature Extraction

We start with dense point-wise feature extraction with pose-invariance, which integrates the two modalities in a dense fusion manner [30]. For the point cloud $\boldsymbol{P}^{\mathrm{obj}}$, we extract pose-invariant geometric features via PoseInv-PointNet++, which is an adapted variant of PointNet++ [22]. In detail, a T-Net [21] is incorporated to predict an affine transformation matrix $\boldsymbol{T} \in \mathbb{R}^{3\times 3}$ from $\boldsymbol{P}^{\mathrm{obj}}$, which aligns the input point cloud before feeding it into PointNet++. Moreover, the injection of absolute coordinates is also excluded from the original PointNet++ to ensure pose-invariant feature extraction. As for the image $\boldsymbol{I}^{\mathrm{obj}}$, we follow SecondPose [3] to employ DINOv2 [19] to extract pose-consistent semantic features, followed by point-wise selection with $\boldsymbol{P}^{\mathrm{obj}}$. The final dense point-wise features $\boldsymbol{F}^{\mathrm{obj}} \in \mathbb{R}^{N^{\mathrm{obj}}\times D}$ are then derived by point-wise concatenation of the image and point cloud features, with feature dimension reduction to D.

3.2. Sparse Feature Interaction

After extracting the point-wise features, we further enhance their shape-sensitivity through holistic feature interaction. However, due to the large number of points, the interaction among dense point-wise features incurs significant computational overhead. Therefore, inspired by AG-Pose [14], we represent object shapes with a set of sparse keypoints, enabling more computationally friendly feature interaction. Concretely, Farthest Point Sampling (FPS) is applied on $\boldsymbol{P}^{\mathrm{obj}}$ to retrieve a set of sparse keypoints $\boldsymbol{P} \in \mathbb{R}^{N\times 3}$, and the corresponding keypoint-wise features $\boldsymbol{F}^{\mathrm{kpt}} \in \mathbb{R}^{N\times D}$ are indexed from $\boldsymbol{F}^{\mathrm{obj}}$, where N denotes the number of keypoints. The pose-invariant distance- and angle-based geometric descriptor [23] is then leveraged to perform shape-sensitive comprehensive interaction among $\boldsymbol{F}^{\mathrm{kpt}}$ and $\boldsymbol{F}^{\mathrm{obj}}$. This process involves L geometric interaction blocks, each consisting of a local geometric cross-attention layer and a global geometric self-attention layer, as detailed below.

Distance- and Angle-based Geometric Descriptor. Given a keypoint $\boldsymbol{P}_n$ and its attended points $\boldsymbol{P}^{\mathrm{att}} \in \mathbb{R}^{N^{\mathrm{att}}\times 3}$, the geometric structure embedding between $\boldsymbol{P}_n$ and $\boldsymbol{P}^{\mathrm{att}}_m$ consists of a pair-wise distance embedding and a K-wise angle embedding, where N^{att} is the number of the attended points. The distance embedding $\boldsymbol{E}^{\mathrm{dis}}_{n,m}$ is computed by applying a sinusoidal function [29] on $\|\boldsymbol{P}^{\mathrm{att}}_m - \boldsymbol{P}_n\|_2/\sigma^{\mathrm{dis}}$, where σ^{dis} is a distance hyper-parameter. As for the angle embedding, the K nearest neighbors $\boldsymbol{P}^{\mathrm{knn}} \in \mathbb{R}^{K\times 3}$ of $\boldsymbol{P}_n$ is first selected from $\boldsymbol{P}^{\mathrm{att}}$. The K-wise angle embedding $\boldsymbol{E}^{\mathrm{ang}}_{n,m,k}$ is then computed with a sinusoidal function on $\angle(\boldsymbol{P}^{\mathrm{knn}}_k - \boldsymbol{P}_n, \boldsymbol{P}^{\mathrm{att}}_m - \boldsymbol{P}_n)/\sigma^{\mathrm{ang}}$, where σ^{ang} is a angle hyper-parameter. Finally, the geometric structure embedding $\boldsymbol{E}_{n,m} \in \mathbb{R}^D$ is formulated as:

$$\boldsymbol{E}_{n,m} = \boldsymbol{E}^{\mathrm{dis}}_{n,m}\boldsymbol{W}^{\mathrm{dis}} + \max_k\{\boldsymbol{E}^{\mathrm{ang}}_{n,m,k}\boldsymbol{W}^{\mathrm{ang}}\}, \qquad (1)$$

where $\boldsymbol{W}^{\mathrm{dis}}, \boldsymbol{W}^{\mathrm{ang}} \in \mathbb{R}^{D\times D}$ are the projection matrices. Due to the pose-invariance of the pair-wise distance and K-wise angle, the geometric structure embedding $\boldsymbol{E}_{n,m}$ is thus pose-invariant. For more details, please refer to [23].

Local Geometric Cross-Attention. Since the sparse keypoints $\boldsymbol{P}$ are an abstract subset of the dense points $\boldsymbol{P}^{\mathrm{obj}}$ and lack local geometric details, we first aggregate local context information into keypoint features with a local geometric cross-attention layer in each geometric interaction block. Specifically, for the n-th keypoint $\boldsymbol{P}_n$, its K^{local} nearest neighbors $\boldsymbol{P}^{\mathrm{obj}}_{\mathrm{knn}(n)} \in \mathbb{R}^{K^{\mathrm{local}}\times 3}$ are first selected within $\boldsymbol{P}^{\mathrm{obj}}$, and the corresponding features $\boldsymbol{F}^{\mathrm{obj}}_{\mathrm{knn}(n)} \in \mathbb{R}^{K^{\mathrm{local}}\times D}$

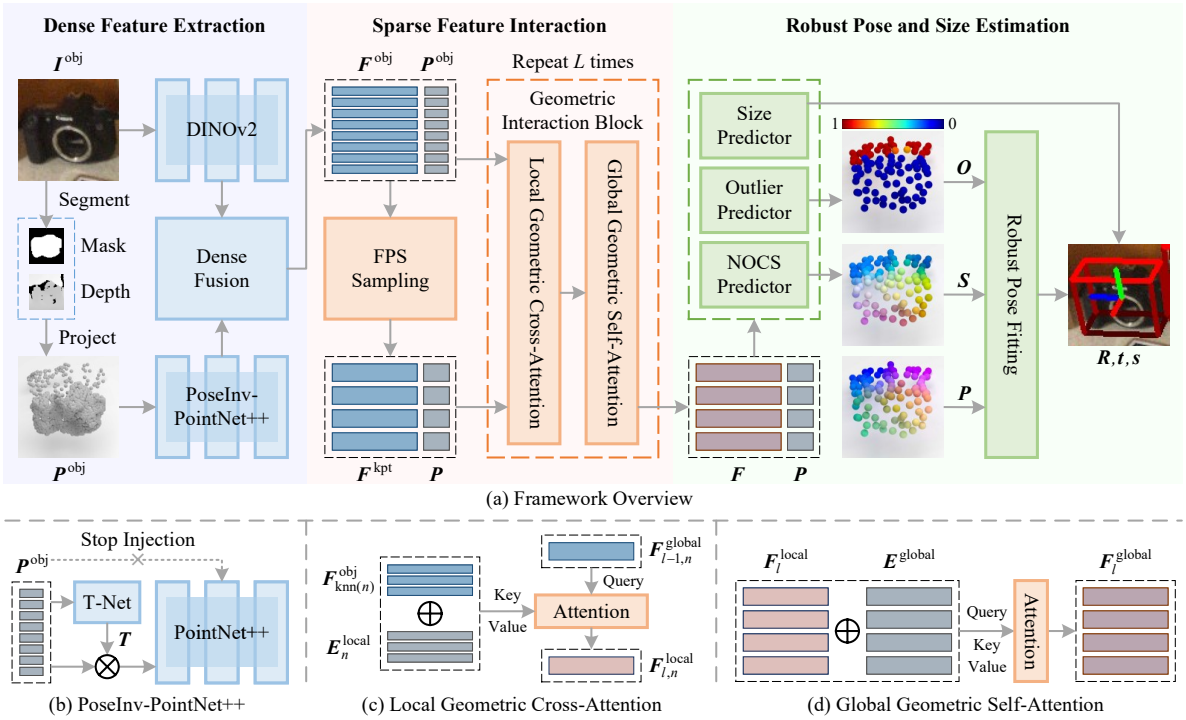

Figure 2. (a) Overview of the proposed SpotPose. Given the observation $\boldsymbol{I}^{\text{obj}}$ and $\boldsymbol{P}^{\text{obj}}$, dense point-wise features $\boldsymbol{F}^{\text{obj}}$ are first extracted with pose-invariance. Subsequently, the object shape is represented by a set of sparse keypoints $\boldsymbol{P}$, whose features $\boldsymbol{F}$ are enhanced with shape-sensitivity through L geometric interaction blocks. The final pose $\{\boldsymbol{R}, \boldsymbol{t}\}$ is derived from keypoint-wise correspondences after outlier removal, while the size $\boldsymbol{s}$ is directly regressed. (b) The PoseInv-PointNet++ is adapted from PointNet++ to extract pose-invariant point cloud features. (c) The local geometric cross-attention layer aims to enrich sparse keypoint-wise features with local context information from dense points. (d) The global geometric self-attention layer aims to holistically enhance keypoint-wise features with shape-sensitivity.

are retrieved from $\boldsymbol{F}^{\text{obj}}$. Then, the local geometric descriptor $\boldsymbol{E}_n^{\text{local}} \in \mathbb{R}^{K^{\text{local}} \times D}$ is computed using Equation (1), with the attended points $\boldsymbol{P}^{\text{obj}}_{\text{knn}(n)}$ and K^{local}-wise angle. In the l-th block, the n-th locally enhanced keypoint feature $\boldsymbol{F}_{l,n}^{\text{local}} \in \mathbb{R}^D$ is derived with the attention mechanism [29] that incorporates geometric information, formulated as:

$$\begin{aligned} \boldsymbol{F}_{l,n}^{\text{local}} &= \text{GCA}(\boldsymbol{F}_{l-1,n}^{\text{global}}, \boldsymbol{F}^{\text{obj}}_{\text{knn}(n)}, \boldsymbol{E}_n^{\text{local}}) + \boldsymbol{F}_{l-1,n}^{\text{global}}, \\ l &\in \{1, \dots, L\}, \quad n \in \{1, \dots, N\}, \end{aligned} \tag{2}$$

where the initial input keypoint feature $\boldsymbol{F}_{0,n}^{\text{global}} = \boldsymbol{F}_n^{\text{kpt}}$ and GCA is the geometric cross-attention operation, defined as:

$$\text{GCA}(\boldsymbol{q}, \boldsymbol{C}, \boldsymbol{E}) = \text{Attention}(\boldsymbol{q}, \boldsymbol{C} + \boldsymbol{E}, \boldsymbol{C} + \boldsymbol{E}), \tag{3}$$

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \boldsymbol{A} \times (\boldsymbol{V}\boldsymbol{W}^V), \tag{4}$$

$$\boldsymbol{A} = \text{Softmax}\left(\frac{(\boldsymbol{Q}\boldsymbol{W}^Q)\,(\boldsymbol{K}\boldsymbol{W}^K)^\top}{\sqrt{D}} \right), \tag{5}$$

where $\boldsymbol{W}^Q, \boldsymbol{W}^K, \boldsymbol{W}^V \in \mathbb{R}^{D \times D}$ are the projection matrices for query, key and value, respectively. Through the local geometric interaction, the keypoint features $\boldsymbol{F}_l^{\text{local}} \in \mathbb{R}^{N \times D}$ are enhanced with the ability to represent local parts.

Global Geometric Self-Attention. After the local context integration, we apply a global geometric self-attention layer to facilitate shape-sensitive holistic interaction among the keypoint features. Concretely, the n-th keypoint $\boldsymbol{P}_n$ attends to all keypoints $\{\boldsymbol{P}_m \mid m = 1, \dots, N\}$ and the corresponding global geometric structure embedding $\boldsymbol{E}_{n,m}^{\text{global}} \in \mathbb{R}^D$ is computed using Equation (1) with the K^{global}-wise angle. To yield the geometric descriptor $\boldsymbol{E}_n^{\text{global}} \in \mathbb{R}^D$ for the n-th keypoint, we average $\boldsymbol{E}_{n,m}^{\text{global}}$ over m. Within the l-th block, the globally enhanced keypoint features $\boldsymbol{F}_l^{\text{global}} \in \mathbb{R}^{N \times D}$ are then derived with the geometric self-attention (GSA) operation, and this process is formulated as follows:

$$\begin{aligned} \boldsymbol{F}_l^{\text{global}} &= \text{GSA}(\boldsymbol{F}_l^{\text{local}}, \boldsymbol{E}^{\text{global}}) + \boldsymbol{F}_l^{\text{local}}, \\ l &\in \{1, \dots, L\}, \end{aligned} \tag{6}$$

$$\text{GSA}(\boldsymbol{C}, \boldsymbol{E}) = \text{Attention}(\boldsymbol{C} + \boldsymbol{E}, \boldsymbol{C} + \boldsymbol{E}, \boldsymbol{C} + \boldsymbol{E}), \tag{7}$$

where the Attention operation is defined in Equation (4). After L geometric interaction blocks, the final keypoint features $\boldsymbol{F} = \boldsymbol{F}_L^{\text{global}} \in \mathbb{R}^{N \times D}$ are acquired. Through the local and global geometric interaction, keypoint features are enhanced with shape-sensitivity, while retaining pose-

invariance owing to the pose-invariant geometric descriptor, which facilitates subsequent transformation learning.

3.3. Robust Pose and Size Estimation

Outlier-aware Correspondence Prediction. Once shape-sensitive and pose-invariant keypoint features are extracted, we follow previous work [16] to predict the corresponding NOCS coordinates $\boldsymbol{S} \in \mathbb{R}^{N\times 3}$ from $\boldsymbol{F}$ with an MLP-based NOCS predictor. The ground truth NOCS coordinates $\boldsymbol{S}^{\mathrm{gt}}$ are derived by applying the ground truth $\boldsymbol{R}^{\mathrm{gt}}, \boldsymbol{t}^{\mathrm{gt}}, \boldsymbol{s}^{\mathrm{gt}}$ to the keypoint coordinates $\boldsymbol{P}$, which is formulated as below:

$$\boldsymbol{S}_n^{\mathrm{gt}} = \frac{1}{\|\boldsymbol{s}^{\mathrm{gt}}\|_2}(\boldsymbol{R}^{\mathrm{gt}})^{\top}(\boldsymbol{P}_n - \boldsymbol{t}^{\mathrm{gt}}), \quad n \in \{1, \dots, N\}, \tag{8}$$

where $\top$ denotes the transpose operation. The NOCS loss function is then defined as the keypoint-wise L^2 distance:

$$\mathcal{L}_n^{\mathrm{nocs}} = \|\boldsymbol{S}_n - \boldsymbol{S}_n^{\mathrm{gt}}\|_2, \quad n \in \{1, \dots, N\}. \tag{9}$$

However, due to noise in the object segmentation mask and depth camera sensor, the observed point cloud may contain outlier points that deviate from the object surface. Moreover, inaccurate NOCS predictions can result in outlier correspondences that are detrimental to the pose fitting process. To this end, we further predict the outlier scores $\boldsymbol{O} \in \mathbb{R}^N$ from $\boldsymbol{F}$ and $\boldsymbol{P}$ with an intuitive outlier predictor as follows:

$$\boldsymbol{O} = \mathrm{Sigmoid}(\mathrm{MLP}([\boldsymbol{F}, \mathrm{MLP}(\boldsymbol{P})])), \tag{10}$$

where $[:,:]$ indicates the feature concatenation operation. To identify outlier points, we follow AG-Pose [14] to produce keypoint-wise $\boldsymbol{O}_n^{\mathrm{gt}}$ based on the instance CAD model $\boldsymbol{P}^{\mathrm{cad}} \in \mathbb{R}^{N^{\mathrm{cad}}\times 3}$, which contains N^{cad} points. In formula,

$$\boldsymbol{O}_n^{\mathrm{gt}} = \begin{cases} 0, & \text{if } \min\limits_{m\in\{1,\dots,N^{\mathrm{cad}}\}} \|\boldsymbol{S}_n^{\mathrm{gt}} - \boldsymbol{P}_m^{\mathrm{cad}}\|_2 < \eta \\ 1, & \text{otherwise} \end{cases}, \tag{11}$$

where η is the outlier threshold. The final outlier-aware correspondence loss function $\mathcal{L}^{\mathrm{corr}}$ is then formed by modulating the keypoint-wise NOCS loss $\mathcal{L}_n^{\mathrm{nocs}}$, defined as:

$$\mathcal{L}_n^{\mathrm{corr}} = \begin{cases} \boldsymbol{I}_n \mathcal{L}_n^{\mathrm{nocs}} - \lambda^{\mathrm{reg}} \log \boldsymbol{I}_n, & \boldsymbol{O}_n^{\mathrm{gt}} = 0 \\ \boldsymbol{I}_n, & \boldsymbol{O}_n^{\mathrm{gt}} = 1 \end{cases}, \tag{12}$$

$$\mathcal{L}^{\mathrm{corr}} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_n^{\mathrm{corr}}, \tag{13}$$

where $\boldsymbol{I}_n = 1 - \boldsymbol{O}_n$ denotes the keypoint-wise inlier score, and λ^{reg} is a hyper-parameter to balance the regularization. Note that when $\boldsymbol{O}_n^{\mathrm{gt}} = 1$, $\boldsymbol{O}_n$ is directly constrained to be 1 to identify outlier keypoints. And when $\boldsymbol{O}_n^{\mathrm{gt}} = 0$, the first term suppresses the inlier score $\boldsymbol{I}_n$ if $\mathcal{L}_n^{\mathrm{nocs}}$ is large, which in turn reduces the weight of $\mathcal{L}_n^{\mathrm{nocs}}$. The second term acts as a regularizer to prevent the NOCS predictor from being lazy and not learning at all. This loss function encourages the network to prioritize correspondences that are easier to predict, while deferring constraints on more challenging correspondences, which are then identified as outliers.

Pose and Size Estimation. Given the keypoint-wise camera coordinates $\boldsymbol{P}$, NOCS coordinates $\boldsymbol{S}$, and outlier scores $\boldsymbol{O}$, we first apply a threshold 0.5 on $\boldsymbol{O}$ to filter out outliers. The remaining inlier correspondences between $\boldsymbol{P}$ and $\boldsymbol{S}$ are then utilized to solve the rotation $\boldsymbol{R}$ and translation $\boldsymbol{t}$ via the Umeyama algorithm [28]. Through this intuitive outlier removal process, robust pose fitting can be achieved without relying on the time-consuming RANSAC algorithm [6]. As for the estimation of object size $\boldsymbol{s}$, we simply employ an MLP-based size predictor for regression from the globally averaged keypoint features $\mathrm{AvgPool}(\boldsymbol{F})$ and utilize L2 loss for supervision, which is written as:

$$\mathcal{L}^{\mathrm{size}} = \|\boldsymbol{s} - \boldsymbol{s}^{\mathrm{gt}}\|_2. \tag{14}$$

In summary, the overall loss function is as follows:

$$\mathcal{L}^{\mathrm{all}} = \mathcal{L}^{\mathrm{corr}} + \lambda^{\mathrm{size}}\mathcal{L}^{\mathrm{size}}, \tag{15}$$

where λ^{size} is a balancing hyper-parameter.

3.4. Discussion

In this section, we elaborate on the distinctions between our SpotPose and AG-Pose [14], which both utilize a set of sparse keypoints to represent object shapes and incorporate geometric information to establish keypoint-wise correspondences. The primary differences lie in four aspects. (1) In terms of keypoint selection, AG-Pose adaptively detects sparse keypoints for different instances, but it fails to provide a suitable shape representation at the early stage of training. In contrast, we simply adopt the training-free farthest point sampling, reducing the risk of network collapse. (2) With regard to holistic feature interaction, keypoint-wise features in AG-Pose are simply concatenated with globally averaged keypoint features, leaving limited feature interaction with each other. By contrast, we leverage the attention mechanism to enable thorough interaction among sparse keypoint-wise features, enhancing their shape-sensitivity. (3) As for geometric feature integration, AG-Pose employs the relative position embedding as its geometric descriptor, which is inherently pose-sensitive. We instead resort to the pose-invariant distance- and angle-based geometric descriptor, ensuring the pose-invariance of the aggregated features. (4) With respect to outlier removal, AG-Pose constrains the detected keypoints to lie on inlier object surfaces, and then uses a deep estimator to regress the pose from all keypoints-wise correspondences, which still suffers from outliers due to noise in the predictions. In contrast, we explicitly predict keypoint-wise outlier scores to directly identify and remove outliers, leading to a more robust object pose estimation.

Table 1. Performance comparison with state-of-the-art methods on the REAL275 and CAMERA25 datasets.

Method		REAL275						CAMERA25					
		IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm	IoU_{50}	IoU_{75}	5°2cm	5°5cm	10°2cm	10°5cm
Direct Regression	DualPoseNet [11]	79.8	62.2	29.3	35.9	50.0	66.8	92.4	86.4	64.7	70.7	77.2	84.7
	GPV-Pose [4]	-	64.4	32.0	42.9	-	73.3	93.4	88.3	72.1	79.1	-	89.0
	HS-Pose [37]	82.1	74.7	46.5	55.2	68.6	82.7	93.3	89.4	73.3	80.5	80.4	89.4
	GenPose [36]	-	-	52.1	60.9	72.4	84.0	-	-	79.9	**84.4**	84.6	89.6
	VI-Net [13]	-	-	50.0	57.6	70.8	82.1	-	-	74.1	81.4	79.3	87.3
	SecondPose [3]	-	-	56.2	63.6	74.7	86.0	-	-	-	-	-	-
Correspondence	NOCS [32]	78.0	30.1	7.2	10.0	13.8	25.2	83.9	69.5	32.3	40.9	48.2	64.4
	SPD [27]	77.3	53.2	19.3	21.4	43.2	54.1	93.2	83.1	54.3	59.0	73.3	81.5
	SGPA [1]	80.1	61.9	35.9	39.6	61.3	70.7	93.2	88.1	70.7	74.5	82.7	88.4
	SAR-Net [10]	79.3	62.4	31.6	42.3	50.3	68.3	86.8	79.0	66.7	70.9	75.3	80.3
	DPDN [12]	83.4	76.0	46.0	50.7	70.4	78.4	-	-	-	-	-	-
	IST-Net [16]	82.5	76.6	47.5	53.4	72.1	80.5	93.7	90.8	71.3	79.9	79.4	89.9
	Query6DoF [34]	82.5	76.1	49.0	58.9	68.7	83.0	91.9	88.1	78.0	83.1	83.9	90.0
	AG-Pose [14]	83.7	79.5	54.7	61.7	74.7	83.1	93.8	91.3	77.8	82.8	85.5	91.6
	SpotPose	**84.1**	**81.2**	**59.7**	**64.8**	**81.5**	**88.2**	**94.3**	**92.5**	**80.4**	83.8	**87.7**	**92.2**

4. Experiments

4.1. Experimental Setup

Datasets. We conduct experiments on three benchmarks including CAMERA25, REAL275 [32] and HouseCat6D [8]. CAMERA25 is a synthetic dataset that comprises 275K training and 25K testing images across 6 object categories. These images are generated using a mixed-reality approach, where foreground objects are rendered against real-world backgrounds. REAL275 is a real-world dataset consisting of 4.3K training images from 7 scenes and 2.75K testing images from 6 scenes, which shares the same object categories with CAMERA25. HouseCat6D is an emerging real-world dataset containing 20K training frames from 34 scenes, 3K testing frames from 5 scenes, and 1.4K validation frames from 2 scenes, spanning 10 household object categories. This collection encompasses photometrically challenging objects with comprehensive viewpoint and occlusion coverage, posing substantial challenges for pose estimation.

Evaluation Metrics. Following previous works [14, 32], we report the mean Average Precision (mAP) of $n°m$ cm for 6D pose evaluation, which indicates the percentage of prediction with rotation error less than $n°$ and translation error less than m cm. We also report the mAP of 3D Intersection over Union (IoU_x) at a threshold of $x\%$ for joint 6D pose and 3D size evaluation.

Implementation Details. For a fair comparison, we employ the same instance masks as previous works [14, 27] from Mask R-CNN [7]. For data preprocessing, images are first cropped and then resized to 224×224, and the number of sampled points is $N^{\text{obj}} = 1024$. For dense image feature extraction, images are fed into the frozen DINOv2 [19], followed by a bilinear interpolation upsampling to the original resolution for point-wise selection. For dense point cloud feature extraction, T-Net [21] is composed of a shared MLP(64,128,256) on each point, a MaxPool across points and a final MLP(128,64,9) to predict $\boldsymbol{T}$, while PointNet++ [22] is kept consistent with previous works [12, 14] except for the incorporation of absolute coordinates. The number of keypoints is $N = 96$ and the feature dimension is $D = 256$. For geometric feature interaction, we adopt $L = 6$ blocks by default, with geometric hyper-parameters $\sigma^{\text{dis}} = 0.2$, $\sigma^{\text{ang}} = 15.0$ and $K^{\text{local}} = 16$, $K^{\text{global}} = 3$. For outlier identification, the threshold η is set to 0.1. In the loss function, balancing hyper-parameters are $\lambda^{\text{reg}} = 0.1$ and $\lambda^{\text{size}} = 0.5$. We train our network using the Adam [9] optimizer with a initial learning rate of 0.001 and a cosine annealing schedule. All experiments are conducted on a single RTX3090Ti GPU with a batch size of 24.

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

Results on REAL275 and CAMERA25 datasets. Table 1 shows the comparison of our method with existing direct regression-based and correspondence-based methods on the REAL275 and CAMERA25 datasets. From the results we can see that on the REAL275 dataset, our SpotPose significantly outperforms all previous methods by a large margin in all evaluation metrics. In particular, when compared with the direct regression-based approaches, SpotPose surpasses SecondPose [3] by 3.5% in 5°2cm and 6.8% in 10°2cm, which also employs DINOv2 [19] as the image backbone. When compared with other correspondence-based methods, SpotPose outperforms the previous leading AG-Pose [14] by 5.0% in 5°2cm, 6.8% in 10°2cm and 1.7% in IoU_{75}. In line with the discussions in Section 3.4, the results demonstrate that our approach achieves a more accurate pose estimation by extracting shape-sensitive and pose-invariant features, while explicitly removing outlier correspondences. Similar results can be found on the CAMERA25 dataset. Although SpotPose falls 0.6% below GenPose [36] in the

Table 2. Performance comparison with state-of-the-art methods on the HouseCat6D dataset.

Method	IoU_{25}	IoU_{50}	5°2cm	5°5cm	10°2cm	10°5cm
FS-Net [2]	74.9	48.0	3.3	4.2	17.1	21.6
GPV-Pose [4]	74.9	50.7	3.5	4.6	17.8	22.7
VI-Net [13]	80.7	56.4	8.4	10.3	20.5	29.1
SecondPose [3]	83.7	66.1	11.0	13.4	25.3	35.7
NOCS [32]	50.0	21.2	-	-	-	-
AG-Pose [14]	81.8	62.5	11.5	12.0	32.7	35.8
SpotPose	**89.1**	**77.0**	**23.8**	**24.5**	**52.3**	**54.8**

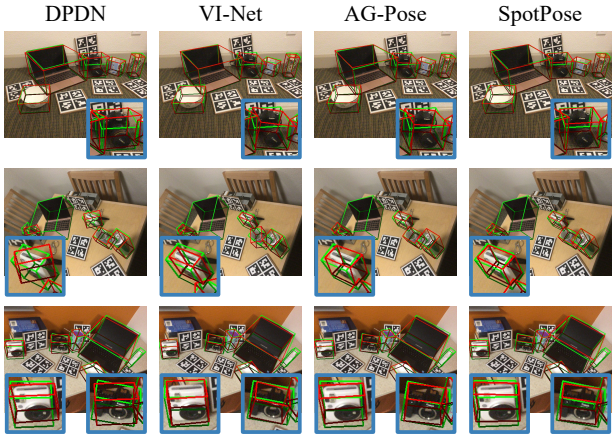

Figure 3. Qualitative comparison of DPDN [12], VI-Net [13], AG-Pose [14] and our SpotPose on the REAL275 dataset. Red/Green indicates the predicted/GT results.

5°5cm metric, it outperforms GenPose in all other metrics, *e.g.*, 3.1% higher in 10°2cm. Additionally, we are free from the costly sampling and filtering process of pose candidates.

Results on HouseCat6D dataset. Table 2 provides the performance comparison on the more challenging HouseCat6D dataset. Once again, our SpotPose achieves the best performance over state-of-the-art approaches by a large margin. Concretely, SpotPose exceeds SecondPose [3] by 12.8% in 5°2cm and 10.9% in IoU_{50}, and surpasses AG-Pose [14] by 12.3% in 5°2cm and 14.5% in IoU_{50}. As for the 10°5cm metric, SpotPose outperforms them by almost 20% (19.1% and 19.0%, respectively). The superior performance on this more comprehensive and challenging real-world dataset further demonstrates the effectiveness of our approach.

Qualitative Comparison. Figure 3 shows the qualitative comparison of SpotPose with the direct regression-based method, VI-Net [13] and correspondence-based methods, DPDN [12] and AG-Pose [14] on the REAL275 dataset. As highlighted in the blue boxes, our SpotPose yields a more accurate pose estimation across diverse shapes and poses.

Table 3. Ablation studies on the dense point cloud backbone.

Dense Point Cloud Backbone	5°2cm	5°5cm	10°2cm	10°5cm
PointNet++ [22]	51.4	56.2	78.9	86.6
PoseInv-PointNet++	**59.7**	**64.8**	**81.5**	**88.2**

Table 4. Ablation studies on the geometric descriptor. Dis-Ang denotes the distance- and angle-based geometric descriptor.

Geometric Descriptor		5°2cm	5°5cm	10°2cm	10°5cm
Local	Global				
None	None	55.8	62.0	77.0	85.0
Dis-Ang	Dis-Ang	**59.7**	64.8	**81.5**	**88.2**
None	Dis-Ang	58.5	64.0	79.0	85.9
PPF [5]	Dis-Ang	58.0	62.6	80.9	87.4
Relative [14]	Dis-Ang	56.4	61.7	80.2	86.7
Dis-Ang	None	59.2	**65.0**	79.6	86.9
Dis-Ang	HP-PPF [3]	59.3	64.6	80.5	87.4
Dis-Ang	Relative [14]	58.0	63.8	78.1	85.4

4.3. Ablation Studies

In this section, we conduct comprehensive ablation studies to shed more light on the superiority of our method on the REAL275 dataset.

Efficacy of Pose-invariant Feature Extraction. In order to validate the necessity of pose-invariant feature extraction, we first conduct ablation studies on the dense point cloud feature extractor in Table 3. The results reveal a significant performance drop of 8.3% in 5°2cm when replacing the proposed pose-invariant PoseInv-PointNet++ with the original pose-sensitive PointNet++ [22], highlighting the effectiveness of extracting inherently pose-invariant features. We further conduct comprehensive ablation studies on local and global geometric descriptors in Table 4. As seen from the table, when the local geometric descriptor is replaced from our pose-invariant Dis-Ang to the pose-sensitive relative position embedding adopted by AG-Pose [14], performance drops by 3.3% in 5°2cm. And when using other pose-invariant descriptors, such as PPF [5], or even omitting geometric descriptors to ensure pose-invariance, performance in 5°2cm is still higher than the relative position embedding. A similar trend can be observed in the ablation studies of the global geometric descriptor, where the pose-sensitive relative position embedding consistently performs worse than the pose-invariant HP-PPF [3] and Dis-Ang in all metrics. These results strongly demonstrate the essentiality of pose-invariant features, which can facilitate correspondence prediction and lead to a more precise object pose estimation performance.

Efficacy of Shape-sensitive Feature Interaction. Table 5 provides the performance comparison of different local and global feature interaction strategies. It can be observed that without feature interaction, performance drops by 12.0% in 5°2cm, indicating the importance of feature interaction

Table 5. Ablation studies on the strategies of feature interaction.

Feature Interaction		5°2cm	5°5cm	10°2cm	10°5cm
Local	Global				
None	None	47.7	55.1	70.6	80.5
Attention	Attention	**59.7**	**64.8**	**81.5**	**88.2**
None	Attention	56.3	62.0	77.6	85.2
AvgPool	Attention	57.7	62.6	79.0	85.8
Attention	None	51.4	58.7	72.6	82.6
Attention	AvgPool [14]	54.4	59.7	77.5	84.6

Table 6. Impact of the number of geometric interaction blocks.

Block Number L	5°2cm	5°5cm	10°2cm	10°5cm
0	47.7	55.1	70.6	80.5
2	54.7	60.9	74.7	82.8
4	56.0	62.2	79.9	88.1
6	**59.7**	**64.8**	**81.5**	**88.2**
8	57.9	63.9	79.9	87.9

from a holistic perspective. For local feature interaction, we experiment with the average pooling (AvgPool) operation, which performs better than no interaction but still lags behind the attention mechanism, which we attribute to the fact that simple AvgPool cannot adequately capture the detailed structure of local parts. For global feature interaction, when keypoint-wise features are simply concatenated with the globally averaged features yielded via an AvgPool, as in AG-Pose [14], performance drops by 5.3% in 5°2cm. The results indicate that this approach fails to capture the holistic structure information as thoroughly as the attention mechanism, limiting the shape-sensitivity of features. Additionally, as shown in Table 4, geometric structure embedding can facilitate the shape-sensitive interaction among features, leading to a 3.9% performance improvement in 5°2cm. We further evaluate the impact of varying number L of geometric interaction blocks in Table 6. As the number of blocks increases, the performance of pose estimation improves, reaching a saturation at six layers. Further increasing the block number does not yield additional performance gains, but instead increases the computational overhead. Therefore, we choose $L = 6$ by default.

Efficacy of Outlier Identification and Removal. We first conduct ablation studies on outlier-aware correspondence prediction in Table 7. When outlier scores are not predicted, all correspondences are indiscriminately learned, which can cause the network to be distracted by outlier points and challenging correspondences. As shown in the table, without outlier-awareness, performance drops by 9.2% in 5°2cm. Even with the RANSAC [6] algorithm for outlier removal, the pose estimation precision remains suboptimal. These results highlight the necessity of explicit identification and removal of outliers. Table 8 further present the performance and time consumption comparison of different pose fitting algorithms. When the Umeyama [28] algorithm

Table 7. Ablation studies on outlier-aware correspondence prediction. Without outlier-awareness, all correspondences are supervised by $\mathcal{L}_n^{\text{nocs}}$ and used for pose fitting. * denotes results with the RANSAC algorithm.

Correspondence Prediction	5°2cm	5°5cm	10°2cm	10°5cm
w/o Outlier-Aware	50.5	56.5	76.4	85.0
w/o Outlier-Aware*	53.1	59.4	77.2	85.8
w/ Outlier-Aware	**59.7**	**64.8**	**81.5**	**88.2**

Table 8. Ablation studies on the pose fitting algorithm. OR indicates outlier removal by applying a threshold on the outlier scores. We also report the average pose fitting time per image.

Pose Fitting	5°2cm	5°5cm	10°2cm	10°5cm	Time (ms)
Umeyama	46.4	51.4	68.0	78.6	1.55
w/ RANSAC [6]	58.0	62.4	81.3	88.0	6.50
Umeyama + OR	**59.7**	**64.8**	**81.5**	**88.2**	1.97
w/ RANSAC [6]	**59.7**	**64.8**	**81.5**	**88.2**	4.94
Deep Estimator [12]	57.1	64.0	78.5	87.6	6.41

is applied on all correspondences, performance is severely affected by outliers, with only 46.4% in 5°2cm. RANSAC improves performance to 58.0% in 5°2cm, but the pose fitting time increases from 1.55ms to 6.50ms due to the need for numerous iterations. In contrast, we identify and remove outliers with a lightweight outlier predictor, achieving 59.7% in 5°2cm with a faster pose fitting speed of 1.97ms per image. Adding RANSAC does not lead to further performance improvements, but rather slows down the pose fitting process, indicating the effectiveness and efficiency of the proposed outlier predictor. We also experiment with a deep estimator [12] for pose fitting, which leads to a 2.6% drop in 5°2cm, exhibiting a weaker robustness to outliers.

5. Conclusion

In this work, we perform a comprehensive analysis of the existing two-stage correspondence-based category-level pose estimation paradigm and introduce two core essentials: shape-sensitive and pose-invariant feature extraction to facilitate transformation learning during the correspondence prediction stage, and outlier correspondence removal to facilitate robust pose estimation during the pose fitting stage. Based on these insights, we propose a simple yet effective approach called SpotPose, which achieves superior performance with no bells and whistles. Extensive experimental results on three challenging benchmarks demonstrate the effectiveness of the proposed method.

6. Acknowledgement

This work was supported by the Open Fund of National Key Laboratory of Deep Space Exploration (Grant NKLDSE2023A009).

References

[1] 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. 1, 2, 3, 6

[2] 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. 2, 7

[3] Yamei Chen, Yan Di, Guangyao Zhai, Fabian Manhardt, Chenyangguang Zhang, Ruida Zhang, Federico Tombari, Nassir Navab, and Benjamin Busam. Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9959–9969, 2024. 2, 3, 6, 7

[4] 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. 2, 6, 7

[5] Bertram Drost, Markus Ulrich, Nassir Navab, and Slobodan Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 998–1005. Ieee, 2010. 7

[6] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2, 3, 5, 8

[7] 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. 3, 6

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

[9] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, San Diega, CA, USA, 2015. 6

[10] 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6707–6717, 2022. 6

[11] 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. 6

[12] 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, 2, 3, 6, 7, 8

[13] 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, 6, 7

[14] Xiao Lin, Wenfei Yang, Yuan Gao, and Tianzhu Zhang. Instance-adaptive and geometric-aware keypoint learning for category-level 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21040–21049, 2024. 1, 2, 3, 5, 6, 7, 8

[15] Zhifeng Lin, Changxing Ding, Huan Yao, Zengsheng Kuang, and Shaoli Huang. Harmonious feature learning for interactive hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12989–12998, 2023. 1

[16] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Ist-net: Prior-free category-level pose estimation with implicit space transformation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13978–13988, 2023. 2, 3, 5, 6

[17] Jian Liu, Wei Sun, Chongpei Liu, Xing Zhang, and Qiang Fu. Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation. *IEEE Transactions on Industrial Informatics*, 19 (11):11171–11181, 2023. 1

[18] Eric Marchand, Hideaki Uchiyama, and Fabien Spindler. Pose estimation for augmented reality: A hands-on survey. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 22(12):2633–2651, 2015. 1

[19] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 3, 6

[20] 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*, pages 4561–4570, 2019. 1

[21] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 2, 3, 6

[22] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 3, 6, 7

[23] Zheng Qin, Hao Yu, Changjian Wang, Yulan Guo, Yuxing Peng, and Kai Xu. Geometric transformer for fast and robust point cloud registration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11143–11152, 2022. 2, 3

[24] Huan Ren, Wenfei Yang, Xiang Liu, Shifeng Zhang, and Tianzhu Zhang. Learning shape-independent transformation via spherical representations for category-level object pose estimation. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[25] Alireza Rezazadeh, Snehal Dikhale, Soshi Iba, and Nawid Jamali. Hierarchical graph neural networks for proprioceptive 6d pose estimation of in-hand objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2884–2890. IEEE, 2023. 1

[26] Yongzhi Su, Jason Rambach, Nareg Minaskan, Paul Lesur, Alain Pagani, and Didier Stricker. Deep multi-state object pose estimation for augmented reality assembly. In *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pages 222–227. IEEE, 2019. 1

[27] Meng Tian, Marcelo H Ang, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 530–546. Springer, 2020. 1, 2, 3, 6

[28] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04): 376–380, 1991. 2, 3, 5, 8

[29] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3, 4

[30] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 1, 3

[31] 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. 1

[32] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2642–2651, 2019. 1, 2, 3, 6, 7

[33] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021. 2

[34] Ruiqi Wang, Xinggang Wang, Te Li, Rong Yang, Minhong Wan, and Wenyu Liu. Query6dof: Learning sparse queries as implicit shape prior for category-level 6dof pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14055–14064, 2023. 2, 3, 6

[35] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You only demonstrate once: Category-level manipulation from single visual demonstration. *Robotics: Science and Systems (RSS)*, 2022. 1

[36] Jiyao Zhang, Mingdong Wu, and Hao Dong. Generative category-level object pose estimation via diffusion models. *Advances in Neural Information Processing Systems*, 36, 2023. 6

[37] Linfang Zheng, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17163–17173, 2023. 2, 6