HS-Pose: Hybrid Scope Feature Extraction for Category-level Object Pose Estimation

Linfang Zheng[1,4]   Chen Wang[1,2]   Yinghan Sun[1]   Esha Dasgupta[4]   Hua Chen[1]

Aleš Leonardis[4]   Wei Zhang[*1,3]   Hyung Jin Chang[4]

[1]Department of Mechanical and Energy Engineering, Southern University of Science and Technology

[2]Department of Computer Science, the University of Hong Kong

[3]Peng Cheng Laboratory, Shenzhen, China

[4]School of Computer Science, University of Birmingham

{lxz948,exd949}@student.bham.ac.uk, cwang5@cs.hku.hk, sunyh2021@mail.sustech.edu.cn

{chenh6,zhangw3}@sustech.edu.cn,{a.leonadis,h.j.chang}@bham.ac.uk

Abstract

In this paper, we focus on the problem of category-level object pose estimation, which is challenging due to the large intra-category shape variation. 3D graph convolution (3D-GC) based methods have been widely used to extract local geometric features, but they have limitations for complex shaped objects and are sensitive to noise. Moreover, the scale and translation invariant properties of 3D-GC restrict the perception of an object's size and translation information. In this paper, we propose a simple network structure, the HS-layer, which extends 3D-GC to extract hybrid scope latent features from point cloud data for category-level object pose estimation tasks. The proposed HS-layer: 1) is able to perceive local-global geometric structure and global information, 2) is robust to noise, and 3) can encode size and translation information. Our experiments show that the simple replacement of the 3D-GC layer with the proposed HS-layer on the baseline method (GPV-Pose) achieves a significant improvement, with the performance increased by 14.5% on $5°2cm$ metric and 10.3% on IoU_{75}. Our method outperforms the state-of-the-art methods by a large margin (8.3% on $5°2cm$, 6.9% on IoU_{75}) on REAL275 dataset and runs in real-time (50 FPS)[1].

1. Introduction

Accurate and efficient estimation of an object's pose and size is crucial for many real-world applications [48], including robotic manipulation [15], augmented reality [36], and autonomous driving, among others. In these appli-
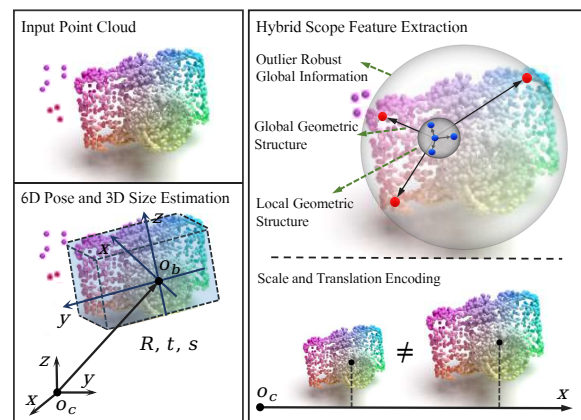

Figure 1. **Illustration of the hybrid scope feature extraction of the HS-layer**. As shown in the right figure, the proposed HS-layer possesses various advantages, including the capability of capturing both local and global geometric information, robustness to outliers, and the encoding of scale and translation information. Building upon the GPV-pose, the HS-layer is employed to develop a category-level pose estimation framework, namely **HS-Pose**. Upon receiving an input point cloud, HS-Pose outputs the estimated 6D pose and 3D size of the object, as shown in the left figure. Given the strengths of the HS-layer, HS-Pose is capable of handling complex object shapes, exhibits robustness to outliers, and achieves better performance compared with existing methods.

cations, it is essential that pose estimation algorithms can handle the diverse range of objects encountered in daily life. While many existing works [3,13,29,50] have demonstrated impressive performance in estimating an object's pose, they typically focus on only a limited set of objects with known shapes and textures, aided by CAD models. In contrast, category-level object pose estimation algorithms [7,22,34,45,49] address all objects within a given category and enable pose estimation of unseen objects during inference without the target objects' CAD models,

which is more suitable for daily-life applications. However, developing such algorithms is more challenging due to the shape and texture diversity within each category.

In recent years, category-level object pose estimation research [5, 55, 56] has advanced rapidly by adopting state-of-the-art deep learning methods. [2, 46] gain the ability to generalize by mapping the input shape to normalized or metric-scale canonical spaces and then recovering the objects' poses via correspondence matching. Better handling of intra-category shape variation is also achieved by leveraging shape priors [4, 42, 56], symmetry priors [20], or domain adaptation [17, 21]. Additionally, [5] enhances the perceptiveness of local geometry, and [7, 55] exploit geometric consistency terms to improve the performance further.

Despite the remarkable progress of existing methods, there is still room for improvement in the performance of the category-level object pose estimation. Reconstruction and matching-based methods [17, 42, 46] are usually limited in speed due to the time-consuming correspondence-matching procedure. Recently, various methods [5, 7, 20, 55, 56] built on 3D graph convolution (3D-GC) [23] have achieved impressive performance and run in real-time. They show outstanding local geometric sensitivity and the ability to generalize to unseen objects. However, only looking at small local regions impedes their ability to leverage the global geometric relationships that are essential for handling complex geometric shapes and makes them vulnerable to outliers. In addition, the scale and translation invariant nature of 3D-GC restrict the perception of object size and translation information.

To overcome the limitations of 3D-GC in category-level object pose estimation, we propose the hybrid scope latent feature extraction layer (HS-layer), which can perceive both local and global geometric relationships and has a better awareness of translation and scale. Moreover, the proposed HS-layer is highly robust to outliers. To demonstrate the effectiveness of the HS-layer, we replace the 3D-GC layers in GPV-Pose [7] to construct a new category-level object pose estimation framework, HS-pose. This framework significantly outperforms the state-of-the-art method and runs in real time. Our approach extends the perception of 3D-GC to incorporate other essential information by using two parallel paths for information extraction. The first path encodes size and translation information (STE), which is missing in 3D-GC due to its invariance property. The second path extracts outlier-robust geometric features using the receptive field with the feature distance metric (RF-F) and the outlier-robust feature extraction layer (ORL).

The main contribution of this paper is as follows:

- We propose a network architecture, the hybrid scope latent feature extraction layer (HS-layer), that can simultaneously perceive local and global geometric structure, encode translation and scale information,

and extract outlier-robust feature information. Our proposed HS-layer balances all these critical aspects necessary for category-level pose estimation.
- We use the HS-layer to develop a category-level pose estimation framework, HS-Pose, based on GPV-Pose. The HS-Pose, when compared to its parent framework, has an advantage in handling complex geometric shapes, capturing object size and translation while being robust to noise.
- We conduct extensive experiments and show that the proposed method can handle complex shapes and outperforms the state-of-the-art methods by a large margin while running in real-time (50FPS).

2. Related Works

Instance-level object pose estimation Instance-level object pose estimation estimates the pose of known objects with the 3D CAD model provided. Existing methods usually achieve the pose using end-to-end regression [14, 16, 18], template matching [1, 30, 35], or 2D-3D correspondence-matching [3, 8, 10, 28, 38, 43]. End-to-end regression-based methods estimate object pose directly from the visual observations and have a high inference speed. Template matching methods recover the object pose by comparing the visual observation and usually exhibit robustness to textureless objects. [11, 44] use the 3D models as templates, which achieve high accuracy but suffer from low matching speed. In recent years, latent feature-based template matching methods [6, 24, 39, 40] have achieved real-time performance and have gained popularity. 2D-3D correspondence matching-based methods [37, 52] first estimate the 2D-3D correspondences and then retrieve the objects' pose by PnP methods. They show outstanding results for textured objects. The correspondences can be sparse bounding box corners [33, 41], or distinguishable points on the object's surface [19, 31, 32]. While the aforementioned methods have shown impressive capabilities in estimating object pose, their applicability is limited to a few objects and usually needs the corresponding CAD models.

Category-level object pose estimation Category-level methods estimate the pose of unseen objects within specific categories [12, 22, 27, 34]. NOCS [46] suggests mapping the input shape to a normalized canonical space (NOCS) and retrieving the pose by point matching. [2, 12, 17] enhance NOCS using a shape prior [42], mapping the shape to a metric scale space [2], or domain adaptation [17]. [4, 21] leverage structural similarity between the shape prior and the observed object. TransNet [53] extends the targets to transparent objects. However, they show limited speed and are unsuitable for real-time applications. CATRE [26] explored real-time pose refinement for pose estimation. FS-Net [5] explored local geometric relationships using 3D-GC [23], which shows robustness to rotation estimation and

runs in real-time. [7, 20, 55, 56] inherit the utilization of 3D-GC and enhance the pose estimation performance in different ways. SAR-Net [20] proposes shape alignment and symmetry-aware shape reconstruction. GPV-Pose [7] presents geometric-pose consistency terms and point-wise bounding box (Bbox) voting. [55, 56] further enhance [7] by shape deformation [56] and residual Bbox voting [55]. Nonetheless, they only look at local geometric relationships and are limited in handling more complex shapes.

3. Methodology

This paper considers the category-level pose estimation problem of estimating the 6D pose and 3D size of an arbitrary instance in the same category based on visual observation. In particular, our approach estimates the 3D rotation $\mathbf{R} \in SO(3)$, the 3D translation $\boldsymbol{t} \in \mathbb{R}^3$, and the size $\boldsymbol{s} \in \mathbb{R}^3$ of object instances based on a depth image, the objects' categories, and segmentation masks. The segmentation mask and category information can be generated by object detectors (*e.g.* MaskRCNN [9]). We use point cloud data $\mathcal{P} \in \mathbb{R}^{N \times 3}$ as the direct input of our network, which is achieved by back-projecting the segmented depth data and downsampling.

Due to the fact that geometric features are essential for determining an object's pose across different shapes, the 3D graph convolution (3D-GC) [23] is widely adopted in recent category-level object pose estimation methods [5,7,20,55,56]. In particular, GPV-Pose [7] uses a 3D-GCN encoder, formed by 3D-GC layers, together with geometric consistency terms for category-level object pose estimation and achieves state-of-the-art performance. However, 3D-GC cannot perceive global geometric features, limiting its capability to handle complex geometric shapes and being sensitive to noise. Also, it is invariant to scale and translation, which contradicts category-level pose estimation tasks (*i.e.*, size and translation estimation).

In this paper, we propose the hybrid scope geometric feature extraction layer (HS-layer) which is based on 3D-GC and keeps its local geometric sensitivity while extending it to have the following characteristics: 1) perception of global geometric structural relationships, 2) robustness to noise, and 3) encoding of size and translation information, particularly for category-level object pose estimation tasks.

3.1. Background of 3D-GC

The core unit of 3D-GC is a deformable kernel that generalizes the convolution kernel used in 2D image processing to deal with unstructured point cloud data. In particular, a 3D-GC kernel K^S is defined as:

$$K^S = \{(\boldsymbol{k}_C, \boldsymbol{w}_C), (\boldsymbol{k}_1, \boldsymbol{w}_1), \dots, (\boldsymbol{k}_S, \boldsymbol{w}_S)\}, \quad (1)$$

where S is the total number of support vectors, $\boldsymbol{k}_C = [0,0,0]^T$ is the central kernel point, $\{\boldsymbol{k}_s \in \mathbb{R}^3\}_{s=1}^S$ are the

support kernel vectors and $\boldsymbol{w}$ is the weight associated with each kernel vector. The 3D-GC kernel performs a convolution on the receptive field $R^M(\boldsymbol{p}_i)$, which is the point along with its neighbors and their associated features $\boldsymbol{f}$:

$$R^M(\boldsymbol{p}_i) = \{(\boldsymbol{p}_i, \boldsymbol{f}_i), (\boldsymbol{p}_m, \boldsymbol{f}_m) | \boldsymbol{p}_m \in \mathcal{N}^M(\boldsymbol{p}_i)\}. \quad (2)$$

Here $\mathcal{N}^M(\boldsymbol{p}_i)$ is the set of the M nearest neighbor points of $\boldsymbol{p}_i$. In particular, in [23] the receptive field with point distance metric (RF-P) is used for finding which of the nearest neighbors is within the point distance metric:

$$\text{dist}_p(\boldsymbol{p}_i, \boldsymbol{p}_j) = \|\boldsymbol{p}_i - \boldsymbol{p}_j\|. \quad (3)$$

For more details, the readers can refer to the original work [23]. It should be noted that 3D-GC has size and translation invariance by design. Although this invariance may be benefit tasks like segmentation and classification, it harms the pose estimation task as the size and translation are the targets to estimate.

3.2. Overall Framework

The overview of the framework, HS-Pose, is shown in Figure 2. We use the proposed HS-layer to form an encoder (HS-encoder) to extract the hybrid scope latent features from the input point cloud data. Then, the extracted latent features are fed into the downstream branches for object pose estimation. To demonstrate the effectiveness of the proposed HS-layer, which can be inserted into any category-level object pose estimation method, we construct our hybrid scope pose estimation network (HS-Pose) based on the state-of-the-art 3D-GC based GPV-Pose with minimal modification. Specifically, we only replace the 3D-GC layers of the 3D-GCN encoder of GPV-Pose with the HS-layer and keep all the other settings the same as the original GPV-Pose, which include network layers, network connection structure, and the downstream branches. Therefore, the extracted features from the encoder along with the input point cloud are fed into three modules for object pose regression, symmetric-based point cloud reconstruction, and bounding box voting. During inference, only the encoder and the pose regression module are used.

Inside the HS-layer, we extract the hybrid scope latent features of the input using two parallel paths. The first path performs scale and translation encoding (STE), which provides essential information for size and translation estimation. The second path extracts outlier-robust geometric features by leveraging local and global geometric relationships, as well as global information in two phases. In the first phase, we form the receptive fields of points based on their feature distances (RF-F), then feed them to a graph convolution (GC) layer to extract high-level geometric features. The output of the GC layer is taken as the second phase's input and passes through an outlier-robust feature extraction layer (ORL), where each point feature is adjusted by
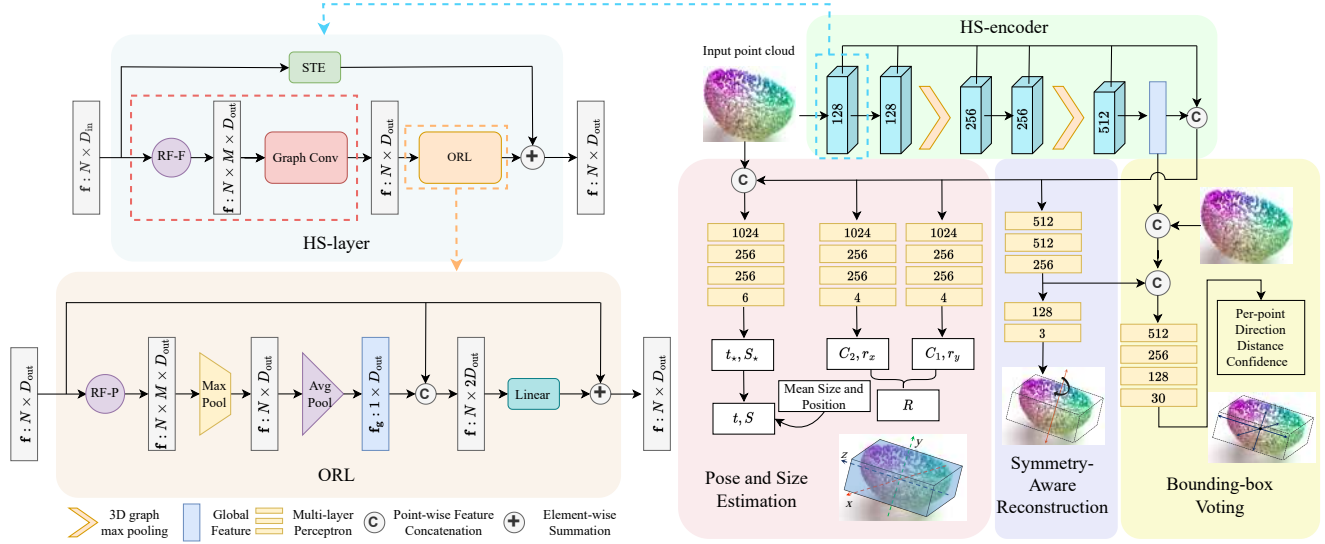
Figure 2. **Overview of the proposed HS-Pose.** The core unit of our framework is HS-layer, which extracts the hybrid scope features of the input data in two paths to gain scale and translation encoding and capture outlier robust geometric features. We stack HS-layers and 3D graph max pooling layers to form an HS-encoder, and then connect it to three sub-modules to form HS-Pose. The three sub-modules are used for pose regression, symmetry-aware point cloud reconstruction, and bounding box voting, respectively.

an outlier robust global information. The final output of the HS-layer is the element-wise summation of the features of both paths.

3.3. Scale and Translation Encoding (STE)

As mentioned earlier, even though 3D-GC provides geometric features crucial in rotation estimation, it loses the essential translation and scale information necessary for pose estimation. To address this problem, existing 3D-GC-based methods try to use another network for translation and size estimation [5] or concatenate the point cloud data with the extracted features for downstream estimation tasks with the assistance of other modules (*i.e.*, bounding box voting) [7, 55, 56]. While these methods are effective and all achieve improvements from the baseline, we emphasize the scale and translation information is beneficial during the latent feature extraction phase.

As shown in Figure 2, our suggestion is to connect in parallel a linear layer (see STE in HS-layer in the figure) to the geometric extraction path and then perform element-wise summation for their output features:

$$\boldsymbol{f}_n^{\text{out}} = \boldsymbol{g}(\boldsymbol{f}_n) + \boldsymbol{h}(\boldsymbol{f}_n), \tag{4}$$

where $\mathbf{h}$ and $\boldsymbol{g}$ apply linear transformation and geometric feature extraction on the features of the points, respectively, and $\boldsymbol{f}_n$ is the n-th point's feature. In particular, we use the points' positions for size and translation encoding in the first layer since there are no features in the original point cloud. Our ablation study in Table 1 shows that this design choice keeps the advantage of geometric feature extraction, and boosts the performance of translation and scale estimation.
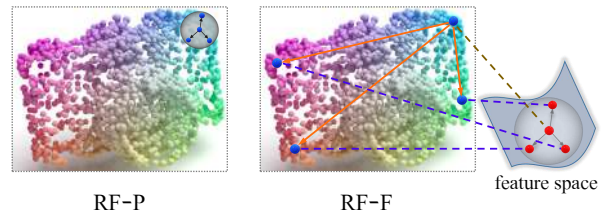

Figure 3. **The illustration and comparison of the receptive field between RF-P and RF-F.** RF-P could only capture geometric structures in a small local region, while RF-F could capture more complex global geometric relationships among the latent features for each point in a high-dimensional hyperspace.

3.4. Receptive field with feature distance (RF-F)

As introduced in Sec. 3.1, 3D-GC learns awareness of local geometric features by forming receptive fields with point Euclidean distance metric (RF-P) and then using the deformable kernel-based graph convolution to extract geometric features for the receptive fields. However, RF-P restricts the perception to small local regions. Even though the perceived regions can be enlarged when cooperating with 3D graph pooling, it can not perceive the global geometric relationships essential for complex geometric structures. This limitation is also exhibited in the performance of category-level object pose estimation tasks [7], where the methods show impressive capability in handling simple geometric shapes (*e.g.* bowl) while encountering difficulty with more complex shapes (*e.g.* mug and camera). However, this limitation has not been well addressed. To this end, we extend the 3D-GC and propose a simple manner to leverage global geometric structural relationships.

We suggest forming the receptive field with the feature distance metric (RF-F). Specifically, we find $\boldsymbol{p}_i$'s neighbors
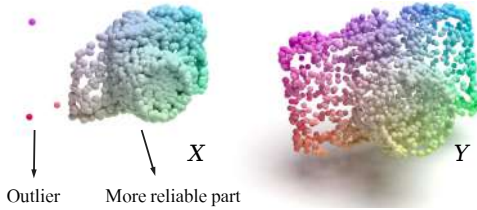
Figure 4. **The design intuition of the outlier robust feature extraction layer (ORL).** X is the input point cloud of a camera with outliers, and Y is the complete shape. Having a perception of global information, especially the more reliable part, helps the network gain resistance to noise.

using the feature distance metric:

$$\text{dist}_f(\boldsymbol{p}_i, \boldsymbol{p}_m) = \|\boldsymbol{f}_i - \boldsymbol{f}_m\|. \tag{5}$$

In other words, with the feature distance metric, the distance between two points is the Euclidean distance between their associated features. We denote the corresponding receptive fields as $R_f^M(\boldsymbol{p}_i)$. This receptive field has the advantage that it is not restricted to local regions; distant points with similar features can also be included.

Figure 3 shows the difference between RF-P and RF-F. RF-F can capture a larger receptive field and, therefore, can capture geometric relationships in a larger area, while the RF-P always formed with local regions. For initialization, in the first layer, we use RF-P and set all the features $\boldsymbol{f}$ to 1. The RF-F is used in the following layers for extracting higher-level geometric relationships.

3.5. Outlier robust feature extraction layer (ORL)

3D-GC's sensitivity to noise influences the category-level methods [5, 7, 55, 56] that are based on it. To address this problem, we introduce an outlier robust feature extraction layer (ORL) on top of the 3D-GC layer, which enhances the method's robustness to noise. The ORL is constructed as follows. Denote the input to this layer as $\{(\boldsymbol{p}_1, \boldsymbol{f}_1), \ldots, (\boldsymbol{p}_N, \boldsymbol{f}_N)\}$, where $\boldsymbol{f}_n \in \mathbb{R}^D$ is the feature of point $\boldsymbol{p}_n$. As illustrated in Figure 4, outliers are distractive, and their features $\boldsymbol{f}$ should not be trusted. To focus on the global information of the more reliable part, we need a mechanism to alleviate the deviation caused by the outliers. Using the global average or maximum pooling directly is limited in addressing this, as all points are taken equally in the pooling procedure.

To lower outliers' influence, we propose using the local region as a guide to extract the global feature. As shown in Figure 2 (see ORL), we first use RF-P to find the M nearest neighbors of each point $\mathcal{N}_p^M(\boldsymbol{p}_n)$. Then, we extract the channel-wise max features of $\mathcal{N}_p^M(\boldsymbol{p}_n)$ using a maximum pooling layer. It should be noted that the points in the reliable parts are more likely to be presented in other points' receptive fields and thus contribute more to the results of the max pooling. The output of the max pooling layer is then passed to a global average pooling layer to get the global feature $\boldsymbol{f}^{\text{global}}$. We then generate an adjusting feature using the $\boldsymbol{f}^{\text{global}}$ and the original input per-point feature $\boldsymbol{f}_n$ by first concatenating them and then feeding them to a linear layer. The final output of ORL is the result of the summation of the adjusting feature and the input features $\boldsymbol{f}_n$ of this layer.

4. Experiments

Implementation details: To rigorously verify the effectiveness of the proposed HS-layer and ensure a fair comparison with the baseline GPV-Pose, we construct the HS-Pose by replacing GPV-Pose's 3D-GC layer with the HS-layer while keeping the overall network structure and network parameters identical to the GPV-Pose, as shown in Figure 2. For a fair comparison, we choose 10 neighbors for the RF-F, consistent with the RF-P in GPV-Pose. The neighbor number of ORL is the same as the RF-F. No other parameters need to be set for the HS-layer as they only depend on the input and output. We also keep the settings, data augmentation strategy, loss terms, and their parameters, the same as those in GPV-Pose's official code[2]. Following GPV-Pose, the off-the-shelf object detector MaskRCNN [9] is employed to generate instance segmentation masks, and 1028 points are randomly sampled as the input to the network. The code is developed using `PyTorch`. We run all experiments on a computer equipped with an Intel(R) Core(TM) i9-10900K CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3090 GPU. All categories are trained together with a batch size of 32, and the training epochs are set to 150 and 300 for REAL275 and CAMERA25 datasets, respectively. The Ranger optimizer [25, 51, 54] is used with the learning rate starting at $1e^{-4}$ and then decreasing based on a cosine schedule for the last 28% training phase.

Baseline methods: We use GPV-Pose [7] as the baseline for the ablation study. Since GPV-Pose did not provide the performance of $10°2\text{cm}$, 2cm, and $5°$, we generate them using their official code[2]. To ensure a fair comparison of their relative speeds, we report GPV-Pose's speed on our machine using the same evaluation code as ours. The results of the other methods are taken directly from the corresponding papers.

Datasets: We evaluate our method on REAL275 [46] and CAMERA25 [46], the two most popular benchmark datasets for category-level object pose estimation. REAL275 is a real-world dataset that provides 7k RGB-D images in 13 scenes. It contains 6 categories of objects (can, laptop, mug, bowl, camera, and bottle), and every category contains 6 instances. The training data comprises 4.3k images from 7 scenes, with 3 objects from each category shown in different scenes. The testing data includes 2.7k images from 6 scenes and 3 objects from each category. CAMERA25 is a synthetic RGB-D dataset that contains the

---

[2]https://github.com/lolrudy/GPV_Pose

Table 1. **Ablation studies on REAL275.**

Higher score means better performance. Overall best results are in bold, and the second-best results are underlined.

Row	Method	IoU_{25}	IoU_{50}	IoU_{75}	$5°2cm$	$5°5cm$	$10°2cm$	$10°5cm$	$2cm$	$5°$	Speed(FPS)
A0	GPV-Pose [7] (baseline)	84.2	**83.0**	64.4	32.0	42.9	55.0	73.3	69.7	44.7	**69**
B0	A0 + STE	84.2	82.2	73.1	36.4	45.1	62.2	76.7	75.6	47.4	<u>66</u>
B1	A0 + RF-F	84.2	<u>82.8</u>	67.7	38.9	52.3	62.1	81.8	71.7	56.1	65
B2	A0 + STE + RF-F	84.1	82.0	72.0	42.7	53.7	63.4	79.2	75.7	57.0	64
C0	A0 + STE + RF-F + Average Pool	84.1	81.7	73.4	43.7	54.8	65.7	81.6	75.7	<u>58.5</u>	62
C1	A0 + STE + RF-F + Max Pool	84.2	81.7	<u>74.8</u>	44.3	54.5	66.9	81.8	77.3	58.1	62
D0	A0 + STL + RF-F + ORL (**Full**)	84.2	82.1	74.7	**46.5**	<u>55.2</u>	<u>68.6</u>	<u>82.7</u>	**78.2**	58.2	50
E0	D0: Neighbor number: $10 \rightarrow 20$	**84.3**	<u>82.8</u>	**75.3**	<u>46.2</u>	**56.1**	**68.9**	**84.1**	<u>77.8</u>	**59.1**	38

same categories as REAL275. It provides 1085 objects for training and 184 for testing. The training set contains 275K images, and the testing set contains 25K.

Evaluation metrics: Following [7,55], we use the mean average precision (mAP) of the *3D Intersection over Union (IoU)* with thresholds of 25%, 50%, and 75% to evaluate the object's size and pose together. We evaluate the rotation and translation estimation performance using the metrics of $5°$, $10°$, $2cm$ and $5cm$, which means an estimation is considered correct if its corresponding error is lower than the threshold. The pose estimation performance is also evaluated using the combination of rotation and translation thresholds: $5°2cm$, $5°5cm$, $10°2cm$, and $10°5cm$.

4.1. Ablation Study

To validate the proposed architecture, we conduct intensive ablation studies using the REAL275 [46] dataset. We incrementally add the proposed strategies (STE, RF-F, and ORL) on the baseline (GPV-Pose) to study their influences. The full ablation study results are shown in Table 1.

[AS-1] Scale and translation encoding (STE). To demonstrate the effectiveness of STE and highlight the significance of scale and translation awareness when extracting latent features, we parallelly connected a single linear layer to each 3D-GC layer in the encoder of the GPV-Pose. The results in Table 1, specifically the [B0] row, indicate that the inclusion of STE has a significant positive impact on scale and translation estimation (**8.7%** improvement on IoU_{75} and **5.9%** improvement on $2cm$) while also slightly improving rotation estimation (2.7% improvement on $5°$). As shown in Table 2, such a simple addition even outperforms the SSP-Pose in several strict metrics (IoU_{75}, $5°2cm$, and $5°5cm$) and shows a notable improvement of 6.8% on the IoU_{75} metric, despite that the SSP-Pose extends the GPV-Pose using a much more complex shape deformation module. The experiment results demonstrate the effectiveness of STE.

[AS-2] Receptive field with feature distance (RF-F). To show the usefulness of the proposed RF-F strategy and to demonstrate the importance of the global geometric re-

lationships, we apply RF-F on GPV-Pose. From the results in Table 1 ([B1]), we see that RF-F has a substantial impact on rotation estimation and brings a performance leap by **11.4%** on $5°$ metric. In addition, it improves the performance on IoU_{75} and $2cm$ by 3.3% and 2.0%, respectively, thanks to the fact that having a sense of the global geometric relationships is helpful in finding the object's center and shape boundary. When comparing the experimental results with the state-of-the-art methods in Table 2, our simple RF-F strategy achieves comparable performance with the state-of-the-art methods and outperforms them on the stricter metrics (*e.g.* $5°2cm$ and $5°5cm$).

[AS-3] The combination of RF-F and STE. To exhibit the benefit of leveraging global geometric relationships and size-translation awareness, we conduct an experiment that combines RF-F and STE. As shown in [B2], the cooperation of RF-F and STE enhances each other and contributes to a better performance than their individual results. When compared with the baseline method, GPV-Pose, the combination of RF-F and STE improves $5°5cm$ by **10.8%**, $5°$ by **12.3%** and IoU_{75} by **7.6%**.

[AS-4] Outlier robust feature extraction layer (ORL). To demonstrate the effectiveness of the ORL, we add the ORL on top of [AS-3]. The results shown in the [D0] row of Table 1 demonstrate that using global features to adjust per-point feature extraction is helpful for both pose and size estimation with an improvement of 5.2% ($10°2cm$) and 2.7% (IoU_{75}), respectively. To check the effectiveness of the outlier robust global feature, we further conduct two experiments by replacing the outlier robust global feature with two popular global pooling methods: average pooling [C0] and max pooling [C1]. The results of [D0], [C0], and [C1] all show the contribution of global information to pose estimation. The comparison between [D0] and [C0, C1] shows that the outlier robust global feature plays a positive role and enhances the overall performance.

[AS-5] Capability of handling complex shapes. To exhibit the proposed method's capability in handling complex geometric shapes, we compare the rotation estimation results of the three proposed strategies (STE, RF-F,
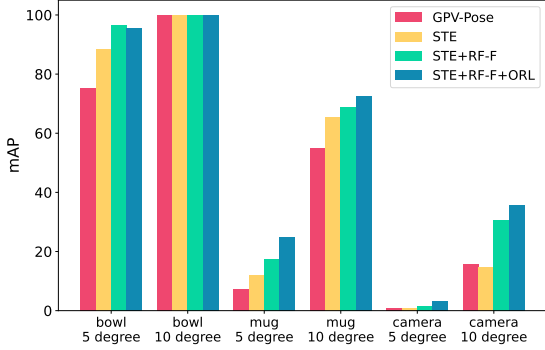
Figure 5. **The rotation estimation of the proposed three strategies and GPV-Pose on categories with different geometric complexity.** The figure shows the rotation estimation mAP ($5°$ and $10°$) on objects with different geometric complexities (*i.e.* bottle is the simplest and the camera is the most complex one). Our method boosted the rotation estimation of the simple shape (bowl) to almost 100% and increased the rotation mAP on more complex objects (mug and camera) by a large margin.

Figure 6. **The comparison of noise resistance between GPV-Pose and the proposed HS-Pose under different outlier ratios (from 0.0% to 40.0%).** Our method outperforms GPV-Pose by a large margin across all outlier ratio levels and is steadier when the outlier ratio increases.

and ORL) and GPV-Pose on categories with different shape complexity in Figure 5. As shown in the figure, the proposed method increases the mAP of categories with complex shapes (*i.e.* mug and camera) and handles simple shapes (*i.e.* bowl) with ease. The figure also demonstrates the effectiveness of leveraging global geometric relationships (STE+RF-F *vs*. STE) and shows the usefulness of outlier robust global information guided feature extraction in ORL (STE+RF-F+ORL *vs*. STE+RF-F).

[AS-6] Noise resistance. To demonstrate the outlier robustness of the proposed method, we tested GPV-Pose and our method under different outlier ratios. As shown in Figure 6, our method outperforms GPV-Pose by a large margin across a range of outlier ratios and is steadier when the outlier ratio increases. More details are in the Supplementary.

[AS-7] Neighbor numbers. We investigate the influence of neighbor numbers used in ORL and RF-F on the performance. The details are presented in the supplementary. The results show that the performance is best when the neighbor numbers are in a certain range. We also observed that using the same neighbor numbers in ORL and RF-F enhances the performance: the precision results are best when the neighbor numbers for both ORL and RF-F are 20 or 30. The results for 20 neighbor numbers are shown in row [E0] of Table 1, which outperforms the results with 10 neighbors. It should be noted that, for a fair comparison with GPV-Pose and focusing on the HS-layer's structural design, we use the results with 10 neighbors (as GPV-Pose) in all tables and figures if not specified.

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

Results on REAL275 dataset: We compare the performance of the proposed HS-Pose with the state-of-the-art methods in Table 2, which shows the mAP scores in different metrics. We choose methods that use depth only for
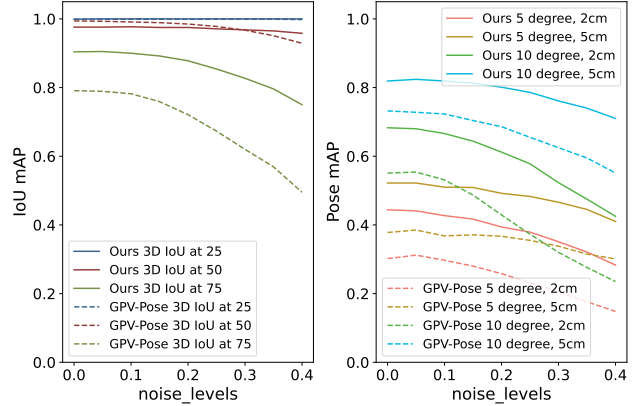
pose estimation for a fair comparison. As shown in the table, our method outperforms the state-of-the-art methods in all metrics except the IoU_{50} in which our method also have comparable performance. Besides, our method can run in real-time. It is worth noting that our method outperforms the second rank on strict metrics by a large margin, with **8.3%** improvement on $5°2cm$, and **7.1%** on $5°5cm$, and **6.9%** on IoU_{75}. We also provide the comparison with methods [2,4,21,22,42,46,47] and that by using other data modalities (*e.g.* RGB and RGB-D) in the supplementary, we outperform the state-of-the-art on 5 metrics out of 9 and achieved the second rank on 3 metrics. Notably, most of them leverage synthetic data, whose datasets contain many more images and objects for training purposes, and also exhibit a limited inference speed. Our method is trained using REAL275 with only 1.6k images and 18 objects while achieving real-time performance. A qualitative comparison between GPV-Pose and our method is shown in Figure 7. Our method achieves a better size and pose estimation (*e.g.* the first three columns), shows robustness to occlusion (*e.g.* the laptop in the last column), and handles complex shapes better (*e.g.* the cameras and mugs in each column).

Results on CAMERA25 Dataset: The performance comparison of the proposed method and the state-of-the-art is shown in Table 3. Our method ranks top and second on all the metrics without prior information. Of the four scores ranked second, three are close to the tops with negligible differences (0.1% on $10°5cm$ and IoU_{75} metrics, and 0.2% on $5°2cm$ metric). It is also worth noting that CAMERA25 is a synthetic dataset that contains no noise, so one main contribution of the proposed method, noise robustness, is not reflected in this dataset. However, this contribution can be identified by comparing the proposed and the state-of-the-art methods' performance on the CAMERA25

---

[3] We use the result provided by GPV-Net, which is higher than the reported result in the FS-Net paper.

Table 2. **Comparison with the state-of-the-art methods (depth only) on REAL275 dataset.**
Higher score means better performance. Overall best results are in bold, and the second-best results are underlined.

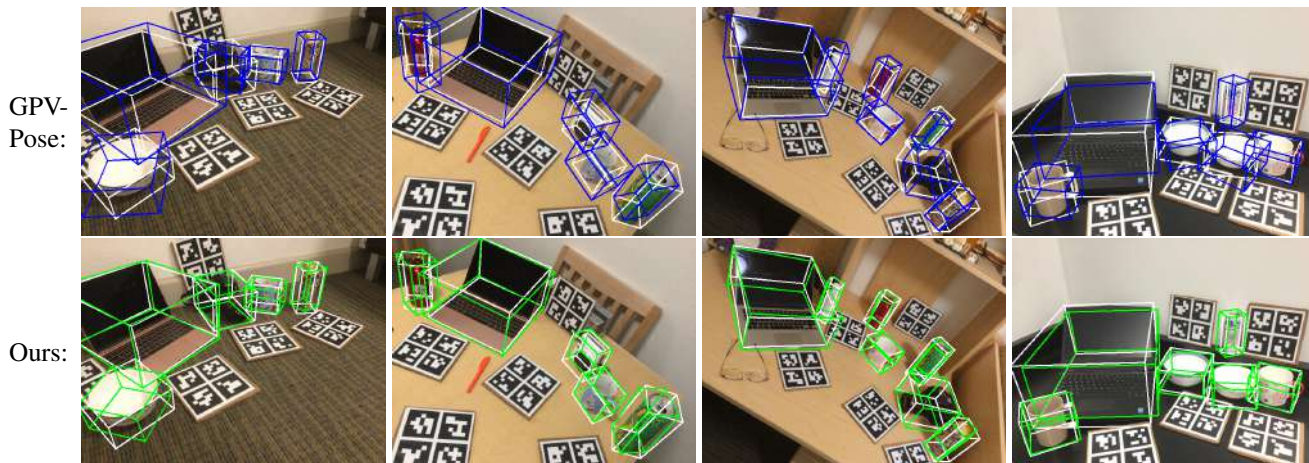| Method	IoU$_{25}$	IoU$_{50}$	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm	Speed(FPS)
SAR-Net [20]	-	79.3	62.4	31.6	42.3	50.3	68.3	-	10
FS-Net[3] [5]	84.0	81.1	63.5	19.9	33.9	-	69.1	71.0	20
UDA-COPE [17]	-	79.6	57.8	21.2	29.1	48.7	65.9	-	-
SSP-Pose [56]	84.0	_82.3_	66.3	34.7	44.6	-	77.8	_79.7_	25
RBP-Pose [55]	-	-	_67.8_	_38.2_	_48.1_	_63.1_	_79.2_	-	25
GPV-Pose [7]	84.1	**83.0**	64.4	32.0	42.9	55.0	73.3	74.6	**69**
Ours	**84.2**	82.1	**74.7**	**46.5**	**55.2**	68.6	**82.7**	**83.7**	_50_

Figure 7. **Qualitative results of our method (green line) and the GPV-Pose (blue line).** The ground truth results are shown with white lines. The estimated rotations of symmetric objects (*e.g.* bowl, bottle, and can) are considered correct if the symmetry axis is aligned.

Table 3. **Comparison with state-of-the-art methods (depth-only) on CAMERA25 dataset.** Overall best results are in bold, and the second-best results are underlined. *Prior* denotes whether the method uses shape priors.

Method	Prior	IoU$_{50}$	IoU$_{75}$	5°2cm	5°5cm	10°2cm	10°5cm
SAR-Net [20]	✓	86.8	79.0	66.7	70.9	75.3	80.3
SSP-Pose [56]	✓	-	86.8	64.7	75.5	-	87.4
RBP-Pose [55]	✓	93.1	_89.0_	**73.5**	_79.6_	**82.1**	**89.5**
GPV-Pose [7]		**93.4**	88.3	72.1	79.1	-	89.0
Ours		_93.3_	**89.4**	_73.3_	**80.5**	_80.4_	_89.4_

and the REAL275 dataset. The REAL275 dataset contains the same object categories as the CAMERA25 but is real-world collected and contains complex noise. It can be observed that the performance drop of our method is much less than other methods when encountering real-world noises in the REAL275. This demonstrates that our method is more noise-robust compared with other methods. A more comprehensive comparison with methods using RGB and RGB-D data is included in the supplementary, in which our method still shows competitive results despite using depth-only data.

5. Conclusion

In this paper, we proposed a hybrid scope latent feature extraction layer, the HS-layer, and used it to construct a category-level object pose estimation framework HS-Pose. Based on the advantages of the HS-layer, HS-Pose can handle complex shapes, capture an object's size and translation, and is robust to noise. The capability of the overall framework is demonstrated in the experiments. The comparisons with the existing methods show that our HS-Pose achieves state-of-the-art performance. In future work, we plan to apply our proposed HS-layer to other problems where unstructured data needs to be processed, and the combination between the local and the global information becomes critical.

Acknowledgements

References

[1] Dingding Cai, Janne Heikkilä, and Esa Rahtu. Ove6d: Object viewpoint encoding for depth-based 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6803–6813, June 2022. 2

[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 (CVPR)*, June 2020. 2, 7

[3] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation, 2022. 1, 2

[4] 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 (ICCV)*, pages 2773–2782, October 2021. 2, 7

[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 (CVPR)*, pages 1581–1590, June 2021. 2, 3, 4, 5, 8

[6] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. PoseRBPF: A Rao-Blackwellized Particle Filter for6D Object Pose Estimation. In *Proceedings of Robotics: Science and Systems*, FreiburgimBreisgau, Germany, June 2019. 2

[7] 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, 2022. 1, 2, 3, 4, 5, 6, 8

[8] Rasmus Laurvig Haugaard and Anders Glent Buch. Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. *CoRR*, abs/2111.13489, 2021. 2

[9] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017. 3, 5

[10] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[11] Stefan Hinterstoisser, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige. Going further with point pair features. In *Computer Vision – ECCV 2016*, pages 834–848. Springer International Publishing, 2016. 2

[12] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi object shape appearance and pose optimization. 2022. 2

[13] Xiaoke Jiang, Donghai Li, Hao Chen, Ye Zheng, Rui Zhao, and Liwei Wu. Uni6d: A unified cnn framework without projection breakdown for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11174–11184, 2022. 1

[14] 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 *IEEE International Conference on Computer Vision (ICCV)*, pages 1530–1538, 10 2017. 2

[15] Nikunj Kothari, Misha Gupta, Leena Vachhani, and Hemendra Arya. Pose estimation for an autonomous vehicle using monocular vision. pages 424–431, 01 2017. 1

[16] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. CosyPose: Consistent multi-view multi-object 6D pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 2

[17] Taeyeop Lee, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon. UDA-COPE: unsupervised domain adaptation for category-level object pose estimation. *CoRR*, abs/2111.12580, 2021. 2, 8

[18] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. *International Journal of Computer Vision*, 128, 03 2020. 2

[19] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[20] 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 (CVPR)*, pages 6707–6717, June 2022. 2, 3, 8

[21] 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. 2, 7

[22] 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 (ICCV)*, pages 3560–3569, October 2021. 1, 2, 7

[23] Zhi-Hao Lin, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1797–1806, 2020. 2, 3

[24] Zheng Linfang, Leonardis Ales, Tse Tze Ho, Elden, Horanyi Nora, Chen Hua, Zhang Wei, and Chang Hyung Jin. Tp-ae: Temporally primed 6d object pose tracking with auto-encoders. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*, 2022. 2

[25] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond, 2019. 5

[26] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. CATRE: iterative point clouds alignment for category-level object pose

refinement. In *European Conference on Computer Vision (ECCV)*, October 2022. 2

[27] Fabian Manhardt, Manuel Nickel, Sven Meier, Luca Minciullo, and Nassir Navab. CPS: class-level 6d pose and shape estimation from monocular images. *CoRR*, abs/2003.05848, 2020. 2

[28] Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, and Guoquan Huang. Symmetry and uncertainty-aware object slam for 6dof object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14901–14910, June 2022. 2

[29] Ningkai Mo, Wanshui Gan, Naoto Yokoya, and Shifeng Chen. Es6d: A computation efficient and symmetry-aware 6d pose regression framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6718–6727, June 2022. 1

[30] Van Nguyen Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit. Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions, 2022. 2

[31] Kiru Park, Timothy Patten, and Markus Vincze. Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation. In *IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[32] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[33] Mahdi Rad and Vincent Lepetit. BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth. In *IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 2

[34] Caner Sahin and Tae-Kyun Kim. Category-level 6d object pose recovery in depth images, 08 2018. 1, 2

[35] Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic. Osop: A multi-stage one shot object pose estimation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6835–6844, June 2022. 2

[36] 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, 2019. 1

[37] 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, 2022. 2

[38] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6825–6834, 2022. 2

[39] Martin Sundermeyer, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras,

and Rudolph Triebel. Multi-path learning for object pose estimation across domains, 2019. 2

[40] Martin Sundermeyer, Zoltan Marton, Maximilian Durner, and Rudolph Triebel. Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection. *International Journal of Computer Vision (IJCV)*, 128, 10 2019. 2

[41] Bugra Tekin, Sudipta Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. pages 292–301, 06 2018. 2

[42] Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 2, 7

[43] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects, 2018. 2

[44] Joel Vidal, Chyi-Yeu Lin, and Robert Martí. 6d pose estimation using an improved method based on point pair features, 2018. 2

[45] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. 2020. 1

[46] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5, 6, 7

[47] 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. 7

[48] Bowen Wen and Kostas Bekris. Bundletrack: 6d pose tracking for novel objects without instance or category-level 3d models. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8067–8074, 2021. 1

[49] 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. *arXiv preprint arXiv:2104.03437*, 2021. 1

[50] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14880–14890, June 2022. 1

[51] Hongwei Yong, Jianqiang Huang, Xiansheng Hua, and Lei Zhang. Gradient centralization: A new optimization technique for deep neural networks, 2020. 5

[52] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. DPOD: 6D Pose Object Detector and Refiner. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[53] Huijie Zhang, Anthony Opipari, Xiaotong Chen, Jiyue Zhu, Zeren Yu, and Odest Chadwicke Jenkins. Transnet: Category-level transparent object pose estimation, 2022. 2

[54] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back, 2019. 5

[55] 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. 2, 3, 4, 5, 6, 8

[56] 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. 2, 3, 4, 5, 8