# Instance-Adaptive and Geometric-Aware Keypoint Learning for Category-Level 6D Object Pose Estimation

Xiao Lin[1] Wenfei Yang[1,2] Yuan Gao[1] Tianzhu Zhang[1,*]
[1]University of Science and Technology of China
[2]Jianghuai Advance Technology Center, Hefei, 230000, China
{llinxiao,wazs98}@mail.ustc.edu.cn {yangwf,tzzhang}@ustc.edu.cn

## Abstract

*Category-level 6D object pose estimation aims to estimate the rotation, translation and size of unseen instances within specific categories. In this area, dense correspondence-based methods have achieved leading performance. However, they do not explicitly consider the local and global geometric information of different instances, resulting in poor generalization ability to unseen instances with significant shape variations. To deal with this problem, we propose a novel Instance-**A**daptive and **G**eometric-Aware Keypoint Learning method for category-level 6D object pose estimation (AG-Pose), which includes two key designs: (1) The first design is an Instance-Adaptive Keypoint Detection module, which can adaptively detect a set of sparse keypoints for various instances to represent their geometric structures. (2) The second design is a Geometric-Aware Feature Aggregation module, which can efficiently integrate the local and global geometric information into keypoint features. These two modules can work together to establish robust keypoint-level correspondences for unseen instances, thus enhancing the generalization ability of the model.Experimental results on CAMERA25 and REAL275 datasets show that the proposed AG-Pose outperforms state-of-the-art methods by a large margin without category-specific shape priors.*

## 1. Introduction

The 6D object pose estimation task aims to predict the rotation, translation and size of objects with 2D/3D observations. Due to its great potential in many real-world applications such as robotic manipulation [25, 38, 39], augmented reality [1, 23] and autonomous driving [3, 7, 22], this task has been attracting increasing attention in the research community. While many previous methods [10, 11, 19, 27, 34, 41] have achieved significant performance on instance-level 6D object pose estimation, their reliance on instance-level CAD models hinders their generalization to novel instances. To deal with this problem, category-level 6D object pose estimation has been introduced in [35], which aims to estimate the poses of unseen instances within specific categories without using their CAD models.

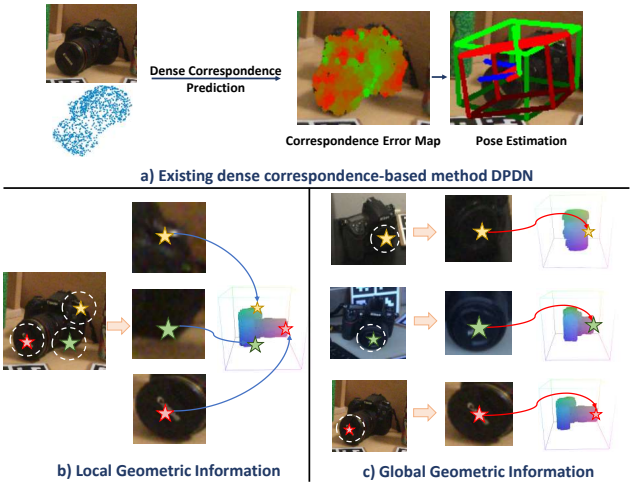


Figure 1. a) The visualization for the correspondence error map and final pose estimation of the dense correspondence-based method, DPDN [17]. Green/red indicates small/large errors and GT/predicted bounding box. b) Points belonging to different parts of the same instance may exhibit similar visual features. Thus, the local geometric information is essential to distinguish them from each other. c) Points belonging to different instances may exhibit similar local geometric structures. Therefore, the global geometric information is crucial for correctly mapping them to the corresponding NOCS coordinates.

In the category-level 6D object pose estimation task, networks are trained with numerous CAD models of different instances and required to estimate the 6D poses of unseen instances which belong to the same category but have different 3D shapes. Consequently, the main challenge of this task lies in the significant intra-category shape variation. Most of the existing category-level methods aim to establish dense correspondences between observed image points (RGB or RGB-D) and the Normalized Object Coor-

*Corresponding author.

dinate Space (NOCS) [35]. They can be categorized into two groups, i.e., the direct dense prediction methods and the two-stage deform-and-match methods. The direct dense prediction methods train their networks to directly map observed image points into the NOCS to obtain dense correspondences [13, 14, 16, 18, 35]. On the other hand, the two-stage methods predict deformation fields on a categorical shape prior [30] to reconstruct instance models first, and then establish dense correspondences between observed points and reconstructed models. With point-level dense correspondences, the final pose and size can be retrieved via Umeyama algorithm [31] or regression networks. Although dense correspondence-based methods have made significant progress, they tend to generate numerous incorrect correspondences for instances with large shape variations (Figure 1 (a)). We owe this problem to that these methods do not explicitly consider the geometric information of different instances, which is of significant importance for generalization to unseen instances. As shown in Figure 1 (b), the three points corresponding to different parts of the camera have similar visual features but distinct local structures. Thus, incorporating local geometric information is crucial to effectively distinguish them from each other. On the other hand, as illustrated in Figure 1 (c), the three points on different cameras exhibit similar local structures but correspond to different NOCS coordinates due to the large shape variations on the camera lenses. These points can be effectively discriminated by taking the global geometric information into account.

To facilitate each point with geometric information in dense correspondence-based methods, an intuitive way is to employ vanilla attention mechanism [32] to aggregate features from all other points. However, it will greatly increase the computational overhead due to the vast number of points. Different from existing methods, we aim to utilize a set of sparse keypoints to represent the shapes of different instances and extract geometric-aware keypoint features to establish robust keypoint-level correspondences for 6D pose estimation. Essentially, the global geometric information can be represented by the relative positions among keypoints, and the local geometric information can be represented by the relative positions between keypoints and their neighboring points. Nevertheless, it's non-trivial because of following challenges: 1) The shapes of different instances vary significantly. Thus, the model needs to adaptively detect representive keypoints for different instances to comprehensively represent their shapes. 2) The local and global geometric information are indispensable. How to efficiently encode such geometric information into keypoint features needs to be carefully considered to achieve better generalization on unseen instances.

Motivated by the above discussions, we propose a novel Instance-**A**daptive and **G**eometric-Aware Keypoint Learning method for category-level 6D object pose estimation (AG-Pose). The framework is shown in Figure 2 (a). The proposed method has two key designs: the Instance-Adaptive Keypoint Detection (IAKD) module and the Geometric-Aware Feature Aggregation (GAFA) module. The IAKD is designed to adaptively detect keypoints for various instances. Specifically, we initialize a set of category-shared learnable queries as keypoint detectors. We first convert them to instance-adaptive detectors by aggregating the object features into learnable queries via attention mechanism [32]. Subsequently, the heatmap of keypoints is obtained by calculating similarities between instance-adaptive detectors and object features. However, the IAKD alone can not guarantee the detected keypoints to focus on different local parts to fully represent the shapes of different instances. Thus, we further design a diversity loss and an object-aware chamfer distance loss to constrain the distribution of keypoints, yielding dispersed and object-focused keypoints. The GAFA module is designed to efficiently extract the local and global geometric information for detected keypoints. In particular, to incorporate local geometric information, GAFA selects the spatially nearest K points for each keypoint and incorporate their relative positions to aggregate features from its neighbors. To incorporate global geometric information, we further integrate global feature and relative positional embeddings into keypoint features. By combining these two modules together, the proposed method can efficiently learn instance-adaptive and geometric-aware keypoints to establish robust keypoint-level correspondences for pose estimation.

In summary, our contributions are as follows:

- We propose a novel instance-adaptive and geometric-aware keypoint learning method for category-level 6D object pose estimation, which can better generalize to unseen instances with large shape variations. To the best of our knowledge, this is the first adaptive keypoint-based method for category-level 6D object pose estimation.
- We evaluate our framework on widely adopted CAMERA25 and REAL275 datasets, and results demonstrate that the proposed method sets a new state-of-the-art performance without using categorical shape priors.

## 2. Related Works

### 2.1. Instance-level 6D object pose estimation

Instance-level 6D object pose estimation aims to estimate the pose of a known object given its CAD model. In recent years, numerous methods based on direct regression [4, 24, 33, 42], fixed keypoint detection [6, 27, 40, 41], and dense 2D-3D correspondences [26, 43] have been proposed. Direct regression-based methods take the observation image (RGB or RGBD) as input and directly predict the 3D rotation and translation via regression networks. The ad-

vantage of these methods lies in end-to-end pose estimation without extra post-processing. On the other hand, methods based on fixed keypoint detection usually predefine a set of keypoints on the object CAD model and detect their corresponding positions in the input image. Subsequently, pose estimation is achieved through Perspectiven-Point (PnP) algorithms. For example, [10, 11, 27, 41] utilize a dense voting strategy to detect keypoints and achieve state-of-the-art performance. Dense 2D-3D correspondence-based methods aim to predict the corresponding position on the object CAD model for every point [19, 34], which are more robust to occlusion. Among them, the approaches based on fixed keypoints are most related to our method. However, they adopt fixed keypoints on a given CAD model while our method aims to adaptively detect keypoints for different instances because the CAD models are not accessible during inference in category-level task.

### 2.2. Category-level 6D object pose estimation

To improve the generalization ablity on unseen instances, the category-level 6D object pose estimation is introduced, which aims to estimate the poses of all instances belonging a specific category without using their CAD models. NOCS [35] proposes to use a shared canonical representation called Normalized Object Coordinate Space (NOCS) to represent the shapes of all instances. They first predict the NOCS coordinates of observed points and then apply Umeyama [31] algorithm to recover the pose and size. To handle the intra-category shape variation, SPD [30] proposes a deformation and matching strategy. They first reconstruct instance models by deforming a categorical shape prior and then match observations to the reconstructed models. Inspired by SPD, multiple works [2, 17, 36, 44] have been proposed to improve the processes of shape prior deformation, correspondence matching and et al., continuously improving the pose estimation performance. However, the above methods have not explicitly taken the local and global geometric information of different instances into consideration, which results in poor generalization ability to unseen instances with significant shape variations.

## 3. Methodology

### 3.1. Overview

Given a RGB-D image, we first employ a segmentation model such as [20, 21] to obtain the segmentation mask and category label for each object. Following previous works [17, 46], here we employ an off-the-shelf MaskRCNN [9]. For each segmented object, we use the segmentation mask to get the cropped RGB image $\mathbf{I}_{obj} \in \mathbb{R}^{H\times W\times 3}$ and the point cloud $\mathbf{P}_{obj} \in \mathbb{R}^{N\times 3}$, where $N$ is the number of points and $\mathbf{P}_{obj}$ is acquired by back-projecting the cropped depth image using camera intrinsics followed by a downsampling process. By taking $\mathbf{I}_{obj}$ and $\mathbf{P}_{obj}$ as input, the proposed method aims to estimate 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 the target object.

The framework of proposed AG-Pose is shown in Figure 2 (a). Our method consists of four main components: Feature Extractor (Sec. 3.2), Instance-Adaptive Keypoint Detector (Sec. 3.3), Geometric-Aware Feature Aggregator (Sec. 3.4) and Pose&Size Estimator (Sec. 3.5). Details about each component are as follows.

### 3.2. Feature Extractor

For the input point cloud $\mathbf{P}_{obj}$, we utilize the PointNet++ [28] to extract point features $\mathbf{F}_P \in \mathbb{R}^{N\times C_1}$. For the RGB image $\mathbf{I}_{obj}$, following [33], a PSP network [45] with ResNet-18 [8] is applied to extract pixel-wise appearance features from $\mathbf{I}_{obj}$. We then choose those pixel features corresponding to $\mathbf{P}_{obj}$ to obtain the RGB features $\mathbf{F}_I \in \mathbb{R}^{N\times C_2}$. Lastly, we concatenate $\mathbf{F}_I$ and $\mathbf{F}_P$ to form $\mathbf{F}_{obj} \in \mathbb{R}^{N\times C}$ as the input for the subsequent networks.

### 3.3. Instance-Adaptive Keypoint Detector

As introduced in Section 1, the geometric information is indispensable to establish robust correspondences. An intuitive idea to facilitate each point with geometric information is to employ the vanilla attention mechanism [32] to aggregate features from all other points. But it is computationally expensive due to the large number of points. Instead, we aim to utilize a set of sparse keypoints to represent the shapes of different instances for pose estimation. However, since the shapes vary across different instances and the instance models are not accessible during inference, fixed keypoint detection methods like [10, 11, 27, 41] are not applicable. Besides, methods like Farest Point Sampling are not end-to-end trainable, which may detect keypoints on outlier points caused by the noisy segmentation mask, the incorrect depth values and et al. Consequently, we design an Instance-Adaptive Keypoint Detection module to adaptively detect keypoints for instances with different shapes which can efficiently represent the object and avoid focusing on outlier points.

The pipeline of IAKD is illustrated in Figure 2 (b). Specifically, we initialize a set of category-shared learnable queries $\mathbf{Q}_{cat} \in \mathbb{R}^{N_{kpt}\times C}$ and each of them represents a keypoint detector. We first inject the object features $\mathbf{F}_{obj}$ into learnable queries via *Transformer Layers* [32]. This process converts the category-shared detectors $\mathbf{Q}_{cat}$ into instance-adaptive detectors $\mathbf{Q}_{ins} \in \mathbb{R}^{N_{kpt}\times C}$ conditioned on object features $\mathbf{F}_{obj}$. In detail,

$$\mathbf{Q} = \mathbf{Q}_{cat}\mathbf{W}^q, \mathbf{K} = \mathbf{F}_{obj}\mathbf{W}^k, \mathbf{V} = \mathbf{F}_{obj}\mathbf{W}^v, \quad (1)$$

$$\mathbf{Q}_{ins} = \mathit{Norm}(\mathbf{Q}_{cat} + \mathit{softmax}(\mathbf{Q}\mathbf{K}^T)\mathbf{V}). \quad (2)$$

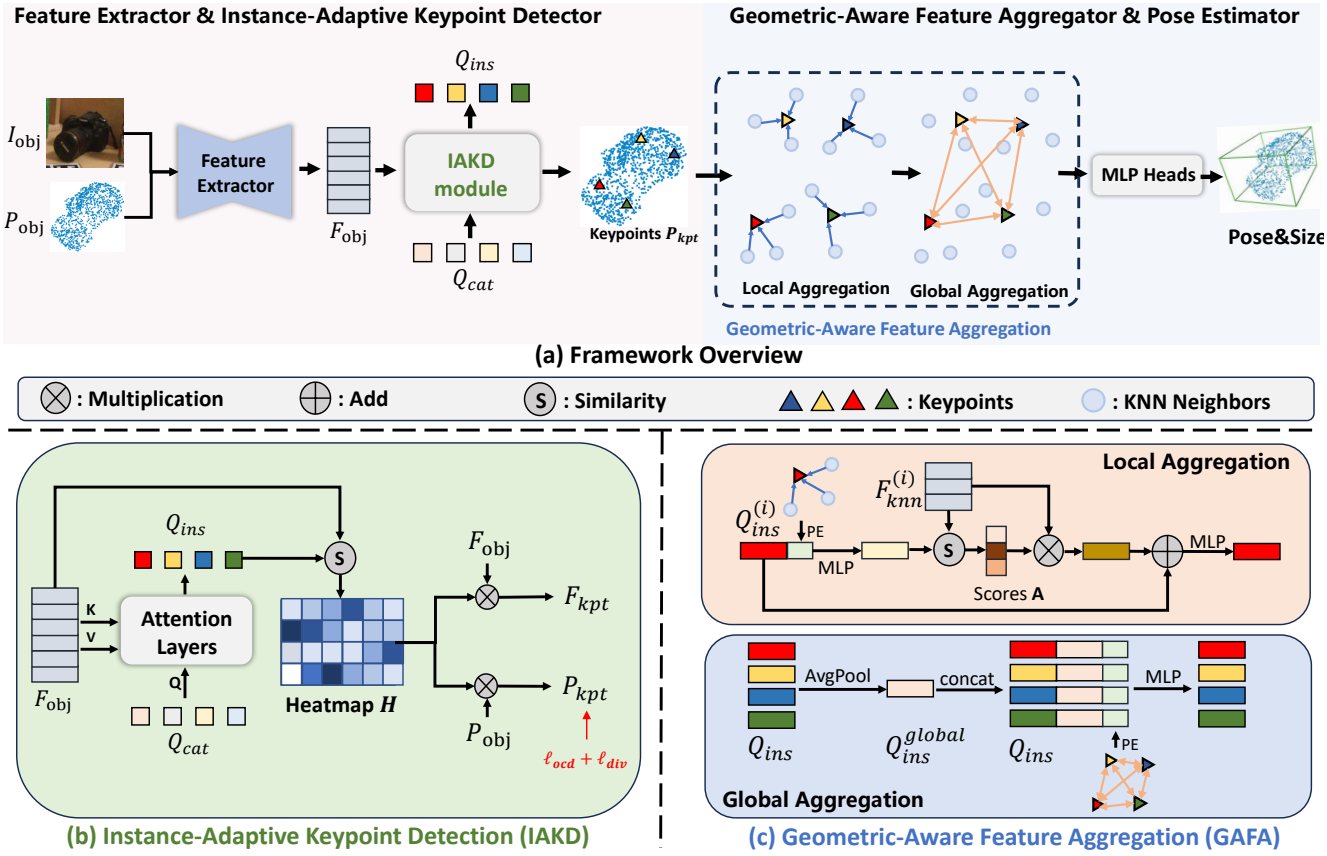


Figure 2. a) Overview of the proposed AG-Pose. b) Illustration of the IAKD module. We initialize a set of category-shared learnable queries and convert them into instance-adaptive detectors by integrating the object features. The instance-adaptive detectors are then used to detect keypoints for the object. To guide the learning of the IAKD module, we futher design the $L_{div}$ and $L_{ocd}$ to constrain the distribution of keypoints. c) Illustration of the GAFA module. Our GAFA can efficiently integrate the geometric information into keypoint features through a two-stage feature aggregation process.

Then we calculate the cosine similarities between the instance-adaptive detectors $\mathbf{Q}_{ins}$ and the object features $\mathbf{F}_{obj}$ to generate keypoint heatmap $\mathbf{H} \in \mathbb{R}^{N_{kpt} \times N}$. Subsequently, the 3D coordinates of keypoints $\mathbf{P}_{kpt} \in \mathbb{R}^{N_{kpt} \times 3}$ and their corresponding features $\mathbf{F}_{kpt} \in \mathbb{R}^{N_{kpt} \times C}$ under camera space are obtained by weighted sum as:

$$\mathbf{P}_{kpt} = softmax(\mathbf{H}) \times \mathbf{P}_{obj}, \tag{3}$$

$$\mathbf{F}_{kpt} = softmax(\mathbf{H}) \times \mathbf{F}_{obj}. \tag{4}$$

Since our target is to utilize a set of sparse keypoints to represent the geometric information of the object, the detected keypoints should be well distributed on the surface of the object. However, we find that the keypoints tend to cluster in small regions and often focus on non-surface or outlier points without explicit supervision. To encourage the keypoints to focus on different parts, we futher design a diversity loss $L_{div}$ to force the detected keypoints to be away from each other. In detail,

$$L_{div} = \sum_{i=1}^{N_{kpt}} \sum_{j=1, j \neq i}^{N_{kpt}} \mathbf{d}(\mathbf{P}_{kpt}^{(i)}, \mathbf{P}_{kpt}^{(j)}), \tag{5}$$

$$\mathbf{d}(\mathbf{P}_{kpt}^{(i)}, \mathbf{P}_{kpt}^{(j)}) = \max\{th_1 - \|\mathbf{P}_{kpt}^{(i)} - \mathbf{P}_{kpt}^{(j)}\|_2, 0\}, \tag{6}$$

where $th_1$ is a hyperparameter and $\mathbf{P}_{kpt}^{(i)}$ stands for the $i$-th keypoint. To encourage the keypoints to locate on the surface of the object and exclude outliers simultaneously, we design an object-aware chamfer distance loss $L_{ocd}$ to constrain the distribution of $\mathbf{P}_{kpt}$. Formally, we first utilize the ground truth pose and size $\mathbf{R}_{gt}, \mathbf{t}_{gt}, \mathbf{s}_{gt}$ to transform $\mathbf{P}_{obj}$ to the NOCS and remove outlier points based on instance model $\mathbf{M}_{obj} \in \mathbb{R}^{M \times 3}$ to produce $\mathbf{P}_{obj}^{\star}$. In formula,

$$\mathbf{P}_{obj}^{\star} = \{x_i | x_i \in \mathbf{P}_{obj} \quad and$$
$$\min_{y_j \in \mathbf{M}_{obj}} \| \frac{1}{\|\mathbf{s}_{gt}\|_2} \mathbf{R}_{gt}(x_i - \mathbf{t}_{gt}) - y_j \|_2 < th_2 \}. \tag{7}$$

The outlier filter process is also illustrated in Figure 3. Then

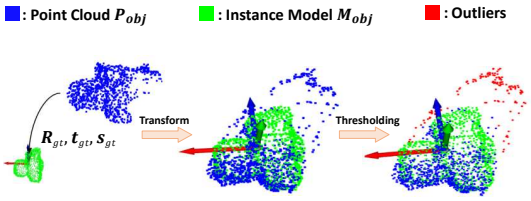


Figure 3. Illustration of the outlier filter process.

the $L_{ocd}$ is calculated as follows,

$$L_{ocd} = \frac{1}{|\mathbf{P}_{kpt}|} \sum_{x_i \in \mathbf{P}_{kpt}} \min_{y_j \in \mathbf{P}^{\star}_{obj}} \|x_i - y_j\|_2. \quad (8)$$

By constraining keypoints to be close to $\mathbf{P}^{\star}_{obj}$, the IAKD module can automatically learn to filter out outlier points during inference.

### 3.4. Geometric-Aware Feature Aggregator

With the detected keypoints for each object, a straightforward way is to predict the NOCS coordinates for these keypoints to establish keypoint-level correspondences. However, the keypoint features are lack of geometric information, which can result in numerous incorrect correspondences on unseen instances. Thus, we propose a Geometric-Aware Feature Aggregation module to efficiently incorporate geometric information into keypoint features.

The GAFA adopts a two-stage pipeline for geometric feature aggregation, as illustrated in Figure 2 (c). We first select the nearest $K$ neighbors in $\mathbf{P}_{obj}$ and their corresponding features in $\mathbf{F}_{obj}$ for each keypoint to obtain $\mathbf{P}_{knn} \in \mathbb{R}^{N_{kpt} \times K \times 3}$ and $\mathbf{F}_{knn} \in \mathbb{R}^{N_{kpt} \times K \times C}$. Essentially, the global geometric information can be represented by the relative positions among keypoints, and the local geometric information can be represented by the relative positions between keypoints and their neighboring points. Thus, we utilize the relative positional embeddings $\alpha$ and $\beta$ between keypoints and their neighbors to represent the local geometric features $f_l$ and global geometric features $f_g$, respectively. In detail,

$$\alpha_{i,j} = MLP(\mathbf{P}^{(i)}_{kpt} - \mathbf{P}^{(i,j)}_{knn}), f^{(i)}_{l} = AvgPool(\alpha_{i,:}), \quad (9)$$

$$\beta_{i,j} = MLP(\mathbf{P}^{(i)}_{kpt} - \mathbf{P}^{(j)}_{kpt}), f^{(i)}_{g} = AvgPool(\beta_{i,:}), \quad (10)$$

where $f^{(i)}_{l}, f^{(i)}_{g} \in \mathbb{R}^{1 \times C}$ and $\mathbf{P}^{(i,j)}_{knn}$ is the $j$-th neighboring point of $\mathbf{P}^{(i)}_{kpt}$. To incorporate local geometric information into keypoints, we combine the keypoint features $\mathbf{Q}_{ins}$ with $f_l$ and calculate local correlation scores $\mathbf{A}$ between keypoints and neighboring points, which is then used to aggregate features from neighbors. The local feature aggregation process for $i$-th keypoint feature $\mathbf{Q}^{(i)}_{ins}$ is as follows,

$$\mathbf{A} = sim(MLP(cat\left[\mathbf{Q}^{(i)}_{ins}, f^{(i)}_{l}\right]), \mathbf{F}^{(i)}_{knn}), \quad (11)$$

$$\mathbf{Q}^{(i)}_{ins} = MLP(softmax(\mathbf{A}) \times \mathbf{F}^{(i)}_{knn} + \mathbf{Q}^{(i)}_{ins}). \quad (12)$$

The above process is executed for all keypoints in parallel to extract representive local geometric features. However, as shown in Figure 1 (c), the local geometric information alone is insufficient to handle significant shape variation across different instances. Therefore, we futher inject the global geometric feature into $\mathbf{Q}_{ins}$ by leveraging all keypoint features and the global geometric features $f_g$. In formula,

$$\mathbf{Q}^{global}_{ins} = AvgPool(\mathbf{Q}_{ins}), \quad (13)$$

$$\mathbf{Q}^{(i)}_{ins} = MLP(concat\left[\mathbf{Q}^{(i)}_{ins}, \mathbf{Q}^{global}_{ins}, f^{(i)}_{g}\right]), \quad (14)$$

where $\mathbf{Q}^{global}_{ins} \in \mathbb{R}^{1 \times C}$ is the global feature of $\mathbf{Q}_{ins}$.

The above two-stage aggregation allows keypoints to adaptively aggregate local geometric features from neighbors and global geometric information from other keypoints. It's worth noting that since the number of keypoints is small, the above process is computationally efficient.

### 3.5. Pose&Size Estimator

After obtaining geometric-aware keypoint features, following previous work [17], we use MLP to predict the NOCS coordinates of keypoints $\mathbf{P}^{nocs}_{kpt} \in \mathbb{R}^{N_{kpt} \times 3}$ from $\mathbf{Q}_{ins}$ and regress the final pose and size $\mathbf{R}, \mathbf{t}, \mathbf{s}$ via this set of keypoint-level correspondences. In formula,

$$\mathbf{P}^{nocs}_{kpt} = MLP(\mathbf{Q}_{ins}), \quad (15)$$

$$\mathbf{f}_{pose} = concat\left[\mathbf{P}_{kpt}, \mathbf{F}_{kpt}, \mathbf{P}^{nocs}_{kpt}, \mathbf{Q}_{ins}\right], \quad (16)$$

$$\mathbf{R}, \mathbf{t}, \mathbf{s} = MLP_R(\mathbf{f}_{pose}), MLP_t(\mathbf{f}_{pose}), MLP_s(\mathbf{f}_{pose}). \quad (17)$$

For more details please refer to [17]. We use the 6D representation [47] for the rotation $\mathbf{R}$. And for the translation $\mathbf{t}$ we follow [46] to predict the residual translation between the ground truth and the mean value of the point cloud.

### 3.6. Overall Loss Function

The overall loss function is as follows:

$$L_{all} = \lambda_1 L_{ocd} + \lambda_2 L_{div} + \lambda_3 L_{nocs} + \lambda_4 L_{pose}, \quad (18)$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters to balance the contribution of each term. For $L_{pose}$, we simply use $L_1$ loss as follows,

$$L_{pose} = \|\mathbf{R}_{gt} - \mathbf{R}\|_2 + \|\mathbf{t}_{gt} - \mathbf{t}\|_2 + \|\mathbf{s}_{gt} - \mathbf{s}\|_2. \quad (19)$$

We generate ground truth NOCS coordinates of keypoints $\mathbf{P}^{gt}_{kpt}$ by projecting their coordinates under camera space $\mathbf{P}_{kpt}$ into NOCS using the ground-truth $\mathbf{R}_{gt}, \mathbf{t}_{gt}, \mathbf{s}_{gt}$. For $L_{nocs}$, we use the $SmoothL_1$ loss [17]. In formula,

$$\mathbf{P}^{gt}_{kpt} = \frac{1}{\|\mathbf{s}_{gt}\|_2} \mathbf{R}_{gt}(\mathbf{P}_{kpt} - \mathbf{t}_{gt}), \quad (20)$$

$$L_{nocs} = SmoothL_1(\mathbf{P}^{gt}_{kpt}, \mathbf{P}^{nocs}_{kpt}). \quad (21)$$

Table 1. **Quantitative comparisons with state-of-the-art methods on the REAL275 dataset.**

| Method | Use of Shape Priors | $IoU_{50}$ | $IOU_{75}$ | 5° 2 **cm** | 5° 5 **cm** | 10° 2 **cm** | 10° 5 **cm** |
|---|---|---|---|---|---|---|---|
| NOCS [35] | ✗ | 78 | 30.1 | 7.2 | 10 | 13.8 | 25.2 |
| DualPoseNet [16] | ✗ | 79.8 | 62.2 | 29.3 | 35.9 | 50 | 66.8 |
| GPV-Pose [5] | ✗ | — | 64.4 | 32 | 42.9 | — | 73.3 |
| IST-Net [18] | ✗ | 82.5 | 76.6 | 47.5 | 53.4 | 72.1 | 80.5 |
| Query6DoF [37] | ✗ | 82.5 | 76.1 | 49 | 58.9 | 68.7 | 83 |
| SPD [30] | ✓ | 77.3 | 53.2 | 19.3 | 21.4 | 43.2 | 54.1 |
| SGPA [2] | ✓ | 80.1 | 61.9 | 35.9 | 39.6 | 61.3 | 70.7 |
| SAR-Net [15] | ✓ | 79.3 | 62.4 | 31.6 | 42.3 | 50.3 | 68.3 |
| RBP-Pose [44] | ✓ | — | 67.8 | 38.2 | 48.1 | 63.1 | 79.2 |
| DPDN [17] | ✓ | 83.4 | 76 | 46 | 50.7 | 70.4 | 78.4 |
| AG-Pose | ✗ | **83.7** | **79.5** | **54.7** | **61.7** | **74.7** | **83.1** |

# 4. Experiments

**Datasets.** Following previous works [17, 18, 37], we conduct experiments on CAMERA25 and REAL275 [35], the most widely adopted datasets for category-level object pose estimation. CAMERA25 is a synthetic RGB-D dataset that contains 300K synthetic images of 1,085 instances from 6 different categories. In this dataset, 25,000 images of 184 instances are used for evaluation and the others are used for training. REAL275 dataset is a more challenging real-world dataset which shares the same categories with CAMERA25. It comprises 7K images from 13 different scenes. 2,750 images of 6 scenes are left for validation, including 3 unseen instances per category.

**Implementation details.** For a fair comparison, we use the same segmentation masks as SPD [30] and DPDN [17] from MaskRCNN [9]. For the data preprocessing, images are cropped and resized to $192 \times 192$ before feature extraction, and the number of points $N$ in point cloud is set as 1024. For model parameters, the number of keypoints is set as $N_{kpt} = 96$, and the local range for each keypoint in GAFA is set as $K = 16$. The feature dimensions are set as $C_1 = 128$, $C_2 = 128$ and $C = 256$, respectively. For the hyper-parameters in the loss functions, the two thresholds in LQKD are set as $th_1 = 0.01$ and $th_2 = 0.1$, and $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are 1.0,5.0,1.0,0.3, respectively. For model optimizing, we train our model on both synthetic and real datasets for evaluation on REAL275 while only on synthetic dataset for evaluation on CAMERA25. Following previous work [17], we use random translation $\Delta\mathbf{t} \sim U(-0.02, 0.02)$ and scaling $\Delta\mathbf{s} \sim U(-0.8, 1.2)$ for data augmentation and apply random rotation for each axis with rotation degree sampled from $U(0, 20)$. We train our network using the ADAM [12] optimizer and employ the triangular2 cyclical learning rate schedule [29] ranging from 2e-5 to 5e-4. All experiments are conducted on a single RTX3090Ti GPU with a batch size of 24.

**Evaluation metrics.** Following previous works [30, 35], we evaluate the model performance with two metrics.

- 3D IoU. We report the mean precision of Intersection-over-Union (IoU) for 3D bounding boxes with thresholds of 50% and 75%. This metric incorporates both the pose and size of the object.
- $n°$ $m$ **cm** is used for direct evaluation of the rotation and translation errors. Only predictions with rotation error less than $n°$ and translation error less than $m$ **cm** are considered correct.

## 4.1. Comparison with State-of-the-Art Methods

**Results on REAL275 dataset.** The comparisons between our AG-Pose and existing state-of-the-art methods on the challenging REAL275 dataset are shown in Table 1. As demonstrated by the results, our AG-Pose outperforms all previous methods in all metrics on REAL275 dataset. It should be noted that our method does not require the use of shape priors. In detail, on the most rigorous metric of 5° 2 **cm**, AG-Pose achieves the precision of 54.7%, surpassing the current state-of-the-art shape prior-based method DPDN [17] with a large margin by 8.7%. As for prior-free methods, our method surpasses Query6DoF [37], IST-Net [18] and GPV-Pose [5] by 5.7%, 7.2%, and 22.7%, respectively. It should be noted that Query6DoF aims to learn a set of sparse queries as implicit shape priors, and the query features are used to update the point features to establish better dense correspondences. Different from it, the proposed AG-Pose aims to use a sparse set of keypoints to explicitly model the geometric information of objects to establish robust keypoint-level correspondences for pose estimation. The superior performance of our method indicates the importance of geometric information in category-level 6D object pose estimation.

**Results on CAMERA25 dataset.** Table 2 shows the quantitative results of the proposed AG-Pose on CAMERA25 dataset. Our method achieves the best performance under most of metrics. In detail, the proposed AG-Pose outperforms the state-of-the-art method Query6DoF [37]

Table 2. **Quantitative comparisons with state-of-the-art methods on the CAMERA25 dataset.**

| Method | Use of Shape Prior | $IoU_{50}$ | $IoU_{75}$ | 5° 2 **cm** | 5° 5 **cm** | 10° 2 **cm** | 10° 5 **cm** |
|---|---|---|---|---|---|---|---|
| NOCS [35] | ✗ | 83.9 | 69.5 | 32.3 | 40.9 | 48.2 | 64.4 |
| DualPoseNet [16] | ✗ | 92.4 | 86.4 | 64.7 | 70.7 | 77.2 | 84.7 |
| GPV-Pose [5] | ✗ | 93.4 | 88.3 | 72.1 | 79.1 | — | 89 |
| Query6DoF [37] | ✗ | 91.9 | 88.1 | **78** | **83.1** | 83.9 | 90 |
| SPD [30] | ✓ | 93.2 | 83.1 | 54.3 | 59 | 73.3 | 81.5 |
| SGPA [2] | ✓ | 93.2 | 88.1 | 70.7 | 74.5 | 82.7 | 88.4 |
| SAR-Net [15] | ✓ | 86.8 | 79 | 66.7 | 70.9 | 75.3 | 80.3 |
| RBP-Pose [44] | ✓ | 93.1 | 89 | 73.5 | 79.6 | 82.1 | 89.5 |
| AG-Pose | ✗ | **93.8** | **91.3** | 77.8 | 82.8 | **85.5** | **91.6** |

Table 3. **Comparisons between the IAKD and FPS.**

| Setting | 5° 2 **cm** | 5° 5 **cm** | 10° 2 **cm** | 10° 5 **cm** |
|---|---|---|---|---|
| FPS | 46.2 | 55.5 | 67.0 | 80.2 |
| IAKD | **54.7** | **61.7** | **74.7** | **83.1** |

Table 4. **Ablation studies on the number of keypoints.**

| $N_{kpt}$ | 5° 2 **cm** | 5° 5 **cm** | 10° 2 **cm** | 10° 5 **cm** |
|---|---|---|---|---|
| 16 | 47.9 | 55.1 | 68.8 | 79.8 |
| 32 | 48.8 | 55.7 | 73.1 | 82.9 |
| 64 | 51 | 57.2 | 72.8 | 82 |
| 96 | **54.7** | **61.7** | **74.7** | 83.1 |
| 128 | 52.8 | 59.9 | 74.3 | **83.7** |

by 1.9% on $IoU_{50}$, 3.2% on $IoU_{75}$, 1.6% on 10° 2 **cm** and 1.6% on 10° 5 **cm**, respectively. Our method achieves comparable performance with Query6DoF on 5° 2 **cm** and 5° 5 **cm** (lower by 0.2% and 0.3%, respectively).

**Results of correspondence errors.** To validate that the keypoint-level correspondences of AG-Pose are more accurate than point-level correspondences, we calculate the NOCS error distributions of DPDN [17], ISTNet [18] and our method on the REAL275 validation set. As shown in Figure 4, the NOCS error of our keypoint-level correspondences is concentrated more on the interval from 0 to 0.1, while errors of dense correspondence-based methods are concentrated more on the interval from 0.15 to 0.5. It proves that the keypoint-level correspondences produced by AG-Pose exhibit better generalizability on unseen instances.

## 4.2. Ablation Studies

In this section, we conduct ablation experiments to demonstrate the effectiveness of the proposed method on the REAL275 dataset.

**Effects of the IAKD module.** To evaluate the effectiveness of the proposed Instance-Adaptive Keypoint Detection module, we first replace the IAKD with the most widely adopted Farest Point Sampling (FPS) strategy. Specifically,

Table 5. **Ablation studies on the proposed loss functions.**

| Loss | 5° 2 **cm** | 5° 5 **cm** | 10° 2 **cm** | 10° 5 **cm** |
|---|---|---|---|---|
| $L_{div}$+$L_{ocd}$ | **54.7** | **61.7** | **74.7** | **83.1** |
| $L_{div}$+$L_{ucd}$ | 49.8 | 57.3 | 74.4 | 82.0 |
| $L_{div}$ | 46.4 | 53 | 71 | 81.3 |
| $L_{ocd}$ | 30 | 36.1 | 55.0 | 68.6 |
| None | 29.3 | 35.6 | 56.4 | 69.6 |

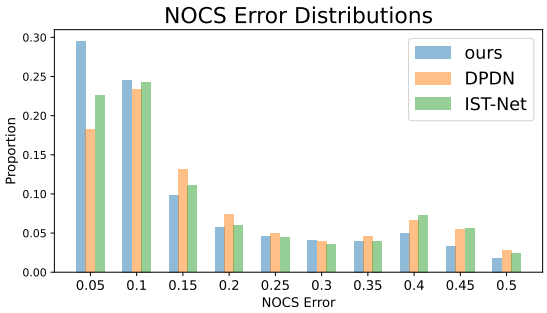


Figure 4. **Comparisons of NOCS error distributions.**

we use FPS to sample the same number of keypoints and use their corresponding features in $\mathbf{F}_{obj}$ as $\mathbf{Q}_{ins}$. For a fair comparison, we add extra MLPs to ensure that the number of parameters remain unchanged. As shown in Table 3, the proposed IAKD outperforms FPS on all metrics. We owe the advantage to the end-to-end trainable property of IAKD and the optimized distribution of keypoints brought by the proposed losses.

**Effects of different keypoint numbers.** In Table 4, we show the impact of different keypoint numbers $N_{kpt}$. Surprisingly, the result shows that our method can achieve comparable performance (47.9% on 5° 2 **cm**) with the SOTA methods by using only a very small number ($N_{kpt} = 16$) of keypoints, which demonstrates the superiority of proposed AG-Pose. And as the number of keypoints ($N_{kpt}$) increases, the performance of our method continues to improve. We choose $N_{kpt} = 96$ in our method for the balance between efficiency and accuracy.

**Effects of the $L_{div}$ and the $L_{ocd}$.** In our method, we

Table 6. **Ablation study on two-stage feature aggregation.**

| Setting | 5° 2 **cm** | 5° 5 **cm** | 10◦ 2 **cm** | 10° 5 **cm** |
|---|---|---|---|---|
| Full | **54.7** | **61.7** | **74.7** | **83.1** |
| w/o GAFA | 47.1 | 55.3 | 70.2 | 80.9 |
| w/o Local | 49 | 57.8 | 71.2 | 82.2 |
| w/o Global | 50.1 | 55.8 | 74.5 | 82.7 |
| w/ vanilla attn | 53 | 61 | 72.2 | 82.1 |

Table 7. **Ablation study on proposed GAFA.**

| | 5° 2 **cm** | 5° 5 **cm** | 10◦ 2 **cm** | 10° 5 **cm** |
|---|---|---|---|---|
| K=8 | 49.9 | 57.6 | 73.1 | 82.5 |
| K=16 | **54.7** | **61.7** | **74.7** | 83.1 |
| K=24 | 54.1 | 61.1 | 73.7 | **83.2** |
| K=32 | 52.7 | 59.9 | 73.6 | 82.8 |

utilize the proposed $L_{div}$ and $L_{ocd}$ to encourage the keypoints to be well distributed on the surface of the object. To verify the effectiveess of them, we conduct ablation experiments on these losses and the results are shown in Table 5. By replacing the proposed $L_{ocd}$ with the normal chamfer distance loss $L_{ucd}$, the performance of our model drops by $4.5\%$ on 5° 2 **cm**. The reason is that our object-aware chamfer distance loss can prevent the model from being affected by outliers, thereby improving the performance. The accuracy of the model further declines by $8.3\%$ after we remove the $L_{ocd}$. It is because $L_{ocd}$ encourages keypoints to distribute on the surface of objects, which can better represent the shapes of objects. Without $L_{div}$, the model performance declines significantly by $24.7\%$, since $L_{div}$ is indispensable for keypoints to focus on distinct regions of objects. The excessive clustering of keypoints can result in a degradation in model performance.

**Effects of the GAFA module.** Here we conduct ablation studies on the Geometric-Aware Feature Aggregation module. For a fair comparison, we replace the GAFA with MLPs that have the same number of parameters. The results are shown in Table 6. In particular, we observe that performance drops by $7.6\%, 4.6\%$ and $5.7\%$ on 5° 2 **cm** when removing the whole GAFA, the global feature aggregation and the local feature aggregation, respectively. As discussed in section 1, both local and global geometric information play crucial roles in establishing accurate correspondences on unseen objects. Our GAFA can effectively encode such geometric information into keypoint features, thus achieving better accuracy. Futhermore, we replace the geometric-aware feature aggregation operation with vanilla attention mechanism (w/ vanilla attn) in the GAFA. The experimental result shows that the proposed geometric-aware feature aggregation achieve better performance because the geometric information is indispensable in category-level 6D object pose estimation. Last, we explore the influence of the local aggregation range $K$ on the accuracy of our model in Table 7. The results demonstrate that $K = 16$ yields the most significant performance gains.

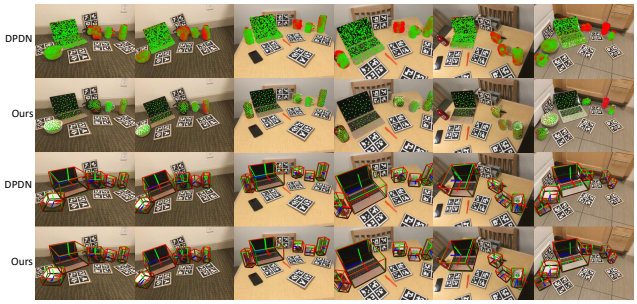


Figure 5. **Qualitative comparisons between our method and DPDN [17] on REAL275 dataset.** We visualize the correspondence error maps and pose estimation results of our AG-Pose and DPDN. Red/green indicates large/small errors and predicted/gt bounding boxes.

### 4.3. Visualization

**Qualitative Results.** The qualitative results of DPDN [17] and proposed AG-Pose are shown in Figure 5. Specifically, we visualize the NOCS prediction errors and the final pose predictions for both methods, in which green/red indicate small/large errors and gt/predicted results. The visualization results show that dense correspondence-based methods such as DPDN generate larger number of incorrect correspondences on novel instances with significant shape variations (e.g., cameras) as well as on outlier points (e.g., edge of objects), which severely degrade the performance of pose estimation. In contrast, our AG-Pose can perform instance-adaptive keypoint detection and extract geometric-aware features to establish accurate keypoint-level correspondences, leading to better pose estimation performance.

## 5. Conclusion

In summary, we present a novel Instance-**A**daptive and **G**eometric-Aware Keypoint Learning method for category-level 6D object pose estimation (AG-Pose). Specifically, we propose a Instance-Adaptive Keypoint Detection module represent the geometric information of different instances through a set of sparse keypoints. Futhermore, we propose a Geometric-Aware Feature Aggregation module to effectively incorporate local and global geometric information into keypoints to establish robust keypoint-level correspondences. We conduct comprehensive experiments and the experimental results verify the effectiveness of our method.

## 6. Acknowledgements

This work was partially supported by the National Nature Science Foundation of China (Grant 62306294), Dreams Foundation of Jianghuai Advance Technology Center (NO.Q/JH-063-2023-T02A/0) and Anhui Provincial Natural Science Foundation (Grant 2308085QF222).

## References


[1] Ronald T Azuma. A survey of augmented reality. Presence: teleoperators & virtual environments, 6(4):355–385, 1997. 1

[2] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2773–2782, 2021. 3, 6, 7

[3] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pages 1907–1915, 2017. 1

[4] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pages 12396–12405, 2021. 2

[5] 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. 6, 7

[6] Fabian Duffhauss, Tobias Demmler, and Gerhard Neumann. Mv6d: Multi-view 6d pose estimation on rgb-d frames using a deep point-wise voting network. In 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 3568–3575, 2022. 2

[7] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In 2012 IEEE conference on computer vision and pattern recognition, pages 3354–3361. IEEE, 2012. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 3

[9] 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

[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), 2020. 1, 3

[11] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021. 1, 3

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 6

[13] 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. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14891–14900, 2022. 2

[14] Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon. Tta-cope: Test-time adaptation for category-level object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 21285–21295, 2023. 2

[15] 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, 7

[16] 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, 2021. 2, 6, 7

[17] 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, 3, 5, 6, 7, 8

[18] Jianhui Liu, Yukang Chen, Xiaoqing Ye, and Xiaojuan Qi. Prior-free category-level pose estimation with implicit space transformation. arXiv preprint arXiv:2303.13479, 2023. 2, 6, 7

[19] Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Bowen Fu, Jiwen Tang, Xiquan Liang, Jingyi Tang, Xiaotian Cheng, Yukang Zhang, Gu Wang, and Xiangyang Ji. Gdrnpp. https://github.com/shanice-l/gdrnpp_bop2022, 2022. 1, 3

[20] Jiahao Lu, Jiacheng Deng, Chuxin Wang, Jianfeng He, and Tianzhu Zhang. Query refinement transformer for 3d instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 18516–18526, 2023. 3

[21] Jiahao Lu, Jiacheng Deng, and Tianzhu Zhang. Bsnet: Box-supervised simulation-assisted mean teacher for 3d instance segmentation. arXiv preprint arXiv:2403.15019, 2024. 3

[22] Fabian Manhardt, Wadim Kehl, and Adrien Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2069–2078, 2019. 1

[23] 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

[24] 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, pages 6718–6727, 2022. 2

[25] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 2901–2910, 2019. 1

[26] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In The IEEE International Conference on Computer Vision (ICCV), 2019. 2

[27] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. 1, 2, 3

[28] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In Advances in Neural Information Processing Systems (NeurIPS), pages 5099–5108, 2017. 3

[29] Leslie N Smith. Cyclical learning rates for training neural networks. In 2017 IEEE winter conference on applications of computer vision (WACV), pages 464–472. IEEE, 2017. 6

[30] 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), 2020. 2, 3, 6, 7

[31] 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

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 2, 3

[33] 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. 2, 3

[34] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. GDR-Net: Geometry-guided direct regression network for monocular 6d object pose estimation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 16611–16621, 2021. 1, 3

[35] 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), 2019. 1, 2, 3, 6, 7

[36] 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. 3

[37] 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. 6, 7

[38] 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

[39] Chaozheng Wu, Jian Chen, Qiaoyu Cao, Jianchi Zhang, Yunxin Tai, Lin Sun, and Kui Jia. Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps. Advances in Neural Information Processing Systems, 33: 13174–13184, 2020. 1

[40] Yangzheng Wu, Alireza Javaheri, Mohsen Zand, and Michael Greenspan. Keypoint cascade voting for point cloud based 6dof pose estimation. In 2022 International Conference on 3D Vision (3DV). IEEE, 2022. 2

[41] Yangzheng Wu, Mohsen Zand, Ali Etemad, and Michael Greenspan. Vote from the center: 6 dof pose estimation in rgb-d images by radial keypoint voting. In European Conference on Computer Vision (ECCV). Springer, 2022. 1, 2, 3

[42] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199, 2017. 2

[43] 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, 2022. 2

[44] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. In European Conference on Computer Vision, pages 655–672. Springer, 2022. 3, 6, 7

[45] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2881–2890, 2017. 3

[46] 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 (CVPR), pages 17163–17173, 2023. 3, 5

[47] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5745–5753, 2019. 5