MMFEIR: Multi-attention Mutual Feature Enhance and Instance Reconstruction for Category-level 6D Object Pose Estimation

Haotian Lei[a], Xiangyu Liu[a], Yan Zhou[a], Guo Niu[a], Changan Yi[a], Yuexia Zhou[a], Xiaofeng Liang[a], Fuhe Liu[a]

[a]*SCHOOL OF ELECTRONIC AND INFORMATION ENGINEERING, Foshan University, Foshan, China*

Abstract

Category-level 6D object pose estimation is a fundamental problem in fields such as robotic manipulation and augmented reality. The goal of this task is to predict the rotation, translation, and size of the object. Current research typically extracts the deformation field from observed point cloud of the object for estimating 6D pose. However, they did not fully consider the interaction between the observed point cloud, prior shape, and image of the object, resulting in the loss of geometric and texture features of the object, thereby affecting the accuracy of pose estimation for objects with large intra class configuration differences. In this paper, we propose a Multi-attention Mutual Feature Enhance Module (MMFEM) to enhance the inherent linkages among different perception data of objects. MMFEM enhances the interaction between images, observed point cloud, and prior shape through multiple attention modules. This enables the network to gain a deeper understanding of the differences between distinct instances. In addition, to improve the feature expression of geometric details for objects, we propose the Instance Reconstruction Deformation Module (IRDM). IRDM reconstructed the three-dimensional instance point cloud for each object, enhancing the model's ability to identify differences in geometric configurations of objects. Extensive experiments on the CAMERA25 and REAL275 datasets show that the proposed methods have achieved 79.0% and 91.2% on the 3D75 metric, 52.6% and 75.9% on the 5°2cm metric, respectively, outperforming current mainstream methods.

Keywords: 6D object pose estimation, Feature enhancement, Instance

reconstruction, Category-level

1. INTRODUCTION

Object pose estimation plays an indispensable role in a variety of domains such as robotic manipulation [1, 2], augmented reality [3, 4], and 3D scene understanding [5, 6]. Traditional methods have largely focused on instance-level pose estimation [7, 8, 9, 10, 11, 12], which involves aligning observed point cloud or images with known 3D models to predict the 6D pose of objects. However, these methods are heavily dependent on the availability of specific 3D models of target objects. Their predictive accuracy can significantly decline when encountering unseen instances within the same category, highlighting the need for improved generalization. This leads to a notable impact on pose estimation performance in real-world scenarios. To address this issue, category-level methods have been introduced. Wang et al. [13] utilizes the concept of Normalized Object Coordinate Space (NOCS), predicting the 3D model point cloud of the object within NOCS (i.e., NOCS point cloud), and solving for the 6D pose of the object by leveraging the correspondences between the observed point cloud obtained from depth maps and the NOCS point cloud. This method effectively addresses the limitations of instance-level approaches, significantly enhancing the generalization capability of pose estimation. It maintains a certain level of predictive accuracy even when applied to unseen objects.

In the tasks of category-level 6D pose estimation, accurately predicting the NOCS point cloud for each instance becomes a challenge due to significant geometric and color differences among intra-class objects. To overcome this issue, some existing studies [14, 15, 16, 17] use prior shapes for specific categories to deform and generate more accurate NOCS point cloud. These prior shapes represent the common geometric shapes of the category, helping to improve prediction accuracy.

However, existing research typically uses point cloud features, or point cloud features concatenated with color information, for training their models [18, 19, 20]. We believe that merely learning the relationship between image and point cloud features is insufficient to address the issue of intra-class deformation. When the model encounters objects with significant shape variations (e.g., cameras), the lack of generalization results in inaccurate pose estimation. Since we already have a prior shape as the common geometric

form for objects within a class, we believe that the prior shapes should be used as a third input to guide the model in better learning the relationship between images and point cloud, thereby improving generalization ability. However, we found that directly concatenating the three types of features does not effectively improve the accuracy of 6D pose estimation. This is because direct concatenation lacks the guidance provided by the prior shapes and does not effectively integrate the relationships among the three input features.

Therefore, in this paper, we design an attention-based multi-feature mutual enhancement module that enables the prior features to more effectively guide and enhance the relationships between the color and point cloud features. At the same time, the color and point cloud features also enhance the prior features, ultimately achieving mutual complementary enhancement among all three types of features. This method improves the model's understanding of the relationships between different instances within the same class by leveraging multiple attention matrices. When the model encounters unseen images and point cloud, it can enhance the relevant features in the image and point cloud by leveraging the known prior features, significantly improving its generalization ability. With this approach, the network can accurately analyze the relationship among the prior shapes, the object's color, and its geometric information, achieving precise pose prediction for unseen objects.

To obtain more accurate NOCS point cloud, traditional methods [14, 16, 17, 21, 22, 23] often rely on prior shapes and apply deformation field to achieve better results, as shown in Fig.1(a). However, we find that this method based on prior shapes deformation cannot accurately capture the specific shape characteristics of each instance. For example, there are significant differences between the prior shapes of a camera model and the actual instance model. Additionally, we observed that the observed point cloud, derived from the depth map, is sparse and incomplete. Using the observed point cloud to replace the shape prior for deformation will inevitably introduce significant errors.

To address this issue, we design a simple and effective instance reconstruction deformation module (IRDM). This module aims to reconstruct the 3D instance point cloud of the observed object and then generate the NOCS point cloud by deforming this complete model. Due to the innovative methods in this article, our prediction accuracy has improved by 10% compared to traditional methods. The specific performance improvements will be de-

tailed in the ablation study section. Finally, we used the DPDN [15] pose estimator, which employs a regression method to determine the 6D pose of the observed object.

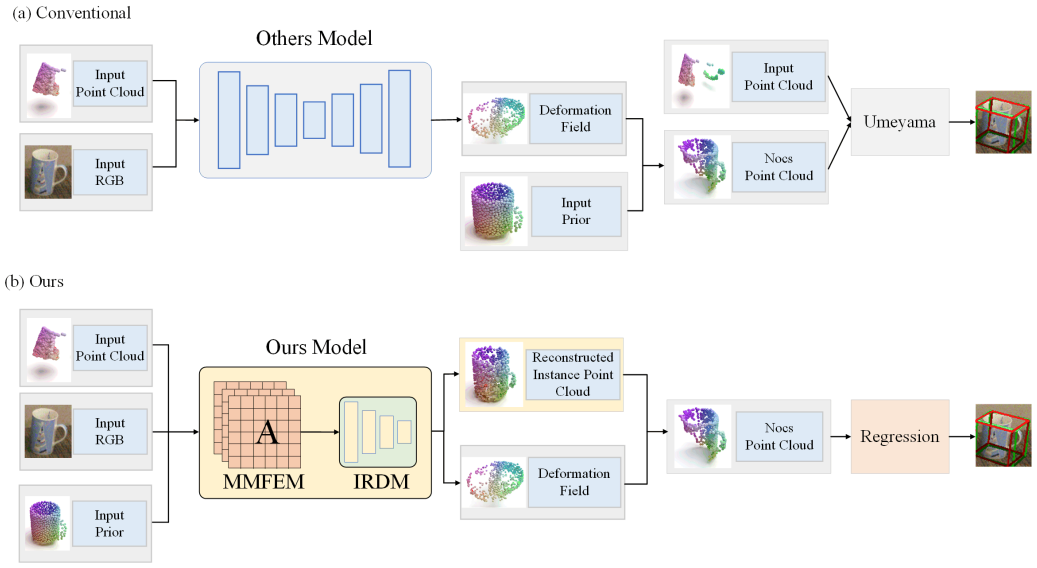

Figure 1: Compared with traditional methods: (a) The traditional methods predicted a deformation field, which is then applied to the prior shapes to generate the NOCS Point Cloud. The 6D pose of the object was estimated using the Umeyama algorithm, based on the input point cloud and the NOCS point cloud. (b) Our method utilizes the multi-attention mutual feature enhance module (MMFEM) to integrate three types of input data. The model then separately predicts the deformation field and the instance point cloud. The deformation field is applied to the instance point cloud to generate the NOCS point cloud. The 6D pose of the object is estimated through regression.

In summary, the main contributions of this paper are as follows:

1) We propose a multi-attention mutual feature enhance module (MMFEM), where each attention module learns the attention weights between two features to enhance the third feature. Through multiple attention matrices, the mutual enhancement and information exchange between the three input features (images, point cloud, and prior shapes) can be effectively achieved, enabling feature complementarity among the three types of input information.

2) We propose an instance reconstruction deformation module, reconstructs the instance point cloud and deformation field, thereby replacing the

use of prior shapes for deformation. Compared to the prior shapes, the instance point cloud provides a more detailed local geometric representation and can more accurately reflect the specific shape of the object.

The rest of this paper is organized as follows. Section 2 provides an overview of the related works. Section 3 describes the design of the method in this paper, including the overall framework and detailed modules. Section 4 presents and analyzes our experimental results. Section 5 summarizes the content of the entire paper.

2. Related Works

2.1. Prior-free Methods

Category-level methods can generally be distinguished based on whether they introduce prior shape, divided into Prior-free Methods and Prior-based Methods. Prior-free Methods do not rely on 3D prior templates to guide pose estimation but mainly focus on designing network structures to better fit the training data. Firstly, He Wang et al. [13] proposed a category-level method that maps different instances of the same category objects to a normalized object coordinate space (NOCS), and then reconstructs the input object point cloud in NOCS. The Umeyama algorithm [24] is applied to determine the object's pose and size by establishing the correspondence between the observed point cloud and the reconstructed point cloud in NOCS. Zheng et al. [19] proposed an HS layer that extends 3D graph convolution to extract mixed-range latent features from point cloud data, while being robust to noise and being able to perceive both local and global geometric structures and global information. Liu et al. [25] proposed a simple prior-free implicit spatial transformation network IST-Net, which transforms camera space features into world space features and implicitly establishes correspondences between them.

Prior-free category-level methods focus more on designing network frameworks. These frameworks are designed to learn the relationship between images and point cloud, enabling 6D pose estimation through regression. However, prior-free methods struggle to handle large intra-class variations due to the absence of prior shapes for the observed target objects. In contrast, prior-based methods can compensate for the prior shape of objects in each category, achieving more accurate precision.

2.2. Prior-based Methods

The main idea of prior-based methods is to utilize category-specific 3D prior shapes to guide pose estimation in order to address the issue of excessive intra-class variation. Tian et al. [14] first proposed SPD to learn a deformation field from pre-learned prior shapes to handle intra-class variations.Specifically, by designing an autoencoder trained on a set of object models to obtain prior shapes. Then, a deformation field is predicted through a deep network and applied to the prior shape to establish the NOCS point cloud. Finally, the 6D pose is obtained through a similarity transformation between the observed point cloud and the NOCS point cloud. Based on this idea, many researchers have proposed their own improved solutions.

Ren et al. [16] proposed Dual-COPE, designed an estimation network with dual prior deformation and transformer-based matching, and also proposed an efficient dual Sim2Real domain adaptation module to reduce the semantic and geometric feature distribution differences between synthetic and real data. Yu et al. [17] proposed a new Transformer-based category-level 6D pose estimation method called CatFormer. Its main feature is the introduction of an end-to-end iterative refinement module, which allows the previous point cloud to undergo multiple deformations based on real scene features. Zou et al. [23] proposed CD-Pose, a geometric consistency and geometric discrepancy learning framework based on depth map point clouds. The framework uses a MLP-based Pose-Consistent Module to extract consistent geometric features and a multi-scale region-guided transformer network to capture instance-level geometric discrepancies. Ultimately, the framework reconstructs the object's NOCS model and estimates the 6D pose via a similarity transformation. Jiang et al. [26] proposed HoPENet which collects high-order statistical information through a global high-order enhancement module and a global high-order pooling operations model. The method fuses global geometric features while capturing long-range statistical correlations and contextual information.

However, the methods mentioned above all deform the prior to obtain the NOCS point cloud, as shown in Fig. 1(a)). We find that using prior shapes for deformation fails to fully capture the uniqueness of individual instances, overlooking their local differences. Therefore, the MMFEM is proposed to fuse image, point cloud, and prior features. The IRDM is designed to reconstruct instances of the objects with large intra-class variations. Finally, by deforming the instance point cloud, the NOCS point cloud is obtained to predict the 6D pose, as shown in Fig.1(b).

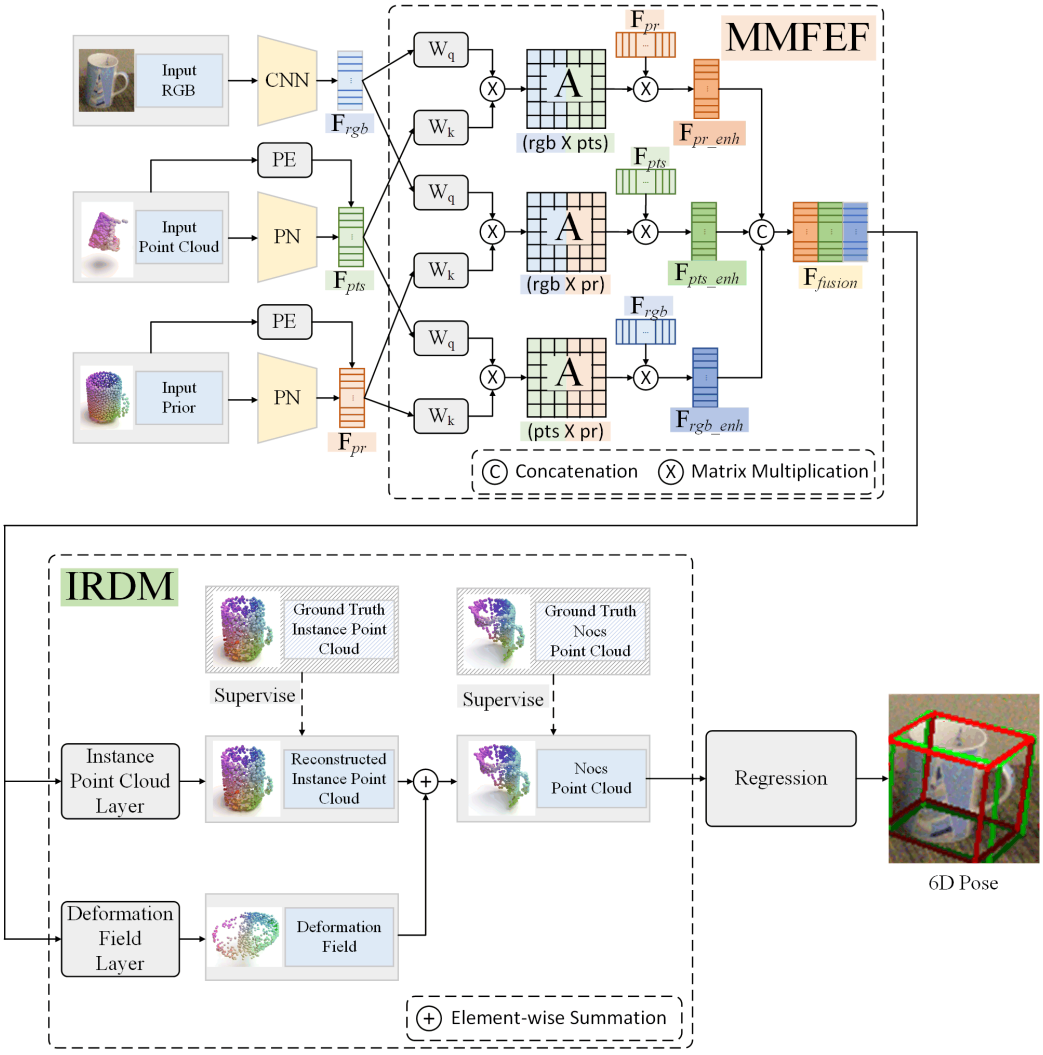

Figure 2: The general framework of the proposed approach. F_{rgb}, F_{pts}, and F_{pr} denote the features of RGB image, point cloud and prior, respectively. The features are mutually enhanced by the MMFEM. In IRDM, we reconstruct the object's instance point cloud and NOCS point cloud, both of which are supervised separately using the ground truth.

3. Method

3.1. Overview

In this section, we will introduce the detailed architecture of the proposed framework, as shown in Fig. 2. We first provide an RGB image $rgb \in \mathbb{R}^{h \times w \times 3}$

with a size of $h \times w$, a point cloud $pts \in \mathbb{R}^{N\times 3}$ with N points, and the prior point cloud of the corresponding category $pr \in \mathbb{R}^{N\times 3}$. We use the CNN backbone network ResNet to extract image features F_{rgb}, and PointNet++ to extract point cloud features F_{pts} and prior features F_{pr}. PointNet++ is insensitive to positional information, which is crucial for accurate translation and rotation estimation. To address this, we use MLP layers to encode the positional information of the point cloud and prior point cloud, integrating it with the features extracted by the network. The feature extraction modules mentioned above are represented by equations (1), (2) and (3):

$$F_{rgb} = \mathrm{CNN}\left(rgb\right), F_{rgb} \in \mathbb{R}^{N\times d} \tag{1}$$

$$F_{pts} = \left[\mathrm{PN}\left(pts\right), \mathrm{MLP}\left(pts\right)\right], F_{pts} \in \mathbb{R}^{N\times (d+d')} \tag{2}$$

$$F_{pr} = \left[\mathrm{PN}\left(pr\right), \mathrm{MLP}\left(pr\right)\right], F_{pr} \in \mathbb{R}^{N\times (d+d')}, \tag{3}$$

where PN represents the PointNet++ network, d denotes the feature dimension extracted by the network, and d' indicates the feature dimension extracted by the MLP layers. After feature extraction, the fusion feature F_{fusion} is obtained through the multi-attention mutual feature enhance module (MMFEM). This module strengthens the relationships among the three feature, as represented by Equation (4) and detailed in Section 3.2.

$$F_{fusion} = \mathrm{MMFEM}(F_{rgb}, F_{pts}, F_{pr}), F_{fusion} \in \mathbb{R}^{N\times M}, \tag{4}$$

where M represents the dimension of the features after fusion. Then we designed a concise and effective instance reconstruction deformation module (IRDM) to reconstruct the target object's 3D instance point cloud P_{model} and NOCS point cloud P_{nocs}. Finally, we utilize the pose estimator from DPDN [15] to regress the pose $\{R, t, s\}$, as represented by equations (5) and (6), respectively, which will be elaborated in Sections 3.3 and 3.4.

$$P_{model}, P_{nocs} = \mathrm{IRDM}(F_{fusion}) \tag{5}$$

$$R, t, s = \mathrm{PoseEstimator}\left(P_{model}, P_{nocs}, F_{rgb}, F_{pts}, F_{pr}\right). \tag{6}$$

3.2. Multi-attention Mutual Feature Enhance Module

Attention mechanisms are typically applied either within a single entity (self-attention) or between two entities (cross-attention). Their primary function is to amplify the important parts of the input features while suppressing

the less significant ones. Building on this idea, we designed a multi-attention mutual feature enhance module. This module more effectively captures the interrelationships between the three input features (F_{rgb}, F_{pts}, F_{pr}) and enables their mutual enhancement.

The attention matrix plays a crucial role in this process, as it can capture the weight relationship between each point of two input features. In the attention matrix, points with higher weight values indicate stronger similarity and greater importance between the two features. Conversely, points with lower weight values indicate a weaker correlation between these features. By applying the attention matrix to previously unprocessed features, we not only increase the weight of their corresponding points but also indirectly integrate the relationships among the three feature types.

However, a single attention matrix can only capture the relationship between two features. To achieve mutual attention and enhancement among the three input features, we adopted a strategy that utilizes multiple attention matrices. The first attention matrix, between image features and prior features, enhances point cloud features. The second attention matrix, between image features and point cloud features, enhances prior features. The third attention matrix, between point cloud features and prior features, enhances image features. The overall process can be represented by the following equations (7) and (8):

$$A_{a\times b} = W_q\left(F_a\right) \times W_k\left(F_b\right)^T, A \in \mathbb{R}^{N\times N} \tag{7}$$

$$F_{c\_enh} = \text{Softmax}\left(A_{a\times b}\right) \times F_c, \tag{8}$$

where W_q and W_k are weight matrices that need to be updated through learning. F_a, F_b and F_c can be replaced with different feature values respectively. $A_{a\times b}$ represents the attention matrix of F_a and F_b. F_{c_enh} represents the enhanced feature F_c. In this way, we obtain mutually enhanced image, point cloud, and prior features, which are also closely connected to each other. Finally, we concatenate all the enhanced features obtained to arrive at the fused feature F_{fusion}.

To better enhance the three types of features, we have established three attention matrices, namely $A_{rgb\times pts}$, $A_{rgb\times pr}$ and $A_{pts\times pr}$. These attention matrices are used to enhance F_{pr}, F_{pts} and F_{rgb} respectively, in order to achieve feature enhancement among them.

3.3. Instance Reconstruction Deformation Module

The core objective of the IRDM is to reconstruct the 3D instance point cloud of the target object and predict its corresponding NOCS point cloud. This module takes the fused feature F_{fusion} output by the multi-attention mutual feature enhance module as the input foundation. This framework utilizes two parallel prediction heads (Consisting of MLP layers). The first prediction head focuses on reconstructing the instance point cloud P_{model} of the object. The second prediction head is committed to predicting a deformation field D, which is then applied to the point cloud instance to generate the object's NOCS point cloud P_{nocs}. The deformation field essentially describes the point-to-point mapping relationship between points of the instance point cloud and the NOCS point cloud. Compared with traditional deformation prior point cloud, to deformed instance point cloud can more accurately capture the local geometric details of the observed object. This improvement not only enhances the model's expressive power but also significantly increases the accuracy of the NOCS point cloud prediction.

To further enhance the performance of the IRDM module, we propose the Selective Instance Reconstruction (SIR) method. This method is based on the symmetry and intra-class variability of objects. For objects with large intra-class differences and asymmetry, instance reconstruction is performed. In contrast, for objects with small intra-class differences and symmetry, the prior point cloud is directly used for deformation processing. The process of the IRDM can be mathematically described by the following equations (9), (10) and (11):

$$P_{\text{model}} = \begin{cases} \text{MLP}\left[F_{\text{fusion}}, \text{AvgPool}\left(F_{\text{fusion}}\right)\right], & \text{if cls} \notin \text{sym} \\ piror, & \text{if cls} \in \text{sym} \end{cases} \tag{9}$$

$$D = \text{MLP}\left(F_{fusion}\right) \tag{10}$$

$$P_{nocs} = P_{model} + D, \tag{11}$$

where cls represents the category of the current object, *prior* represents the prior model of the current object, and sym indicates the category of objects that have symmetry. The AvgPool$(\cdot)$ represents the global average pooling operation applied to the feature to extract global context information.

3.4. Pose Estimation

Our pose estimation module refers to the pose estimation method proposed in DPDN [15] to achieve a precise 6D pose estimation of the target

object. We first perform positional encoding on the NOCS point cloud and the instance point cloud, ensuring effective integration of spatial information. Subsequently, we combine these encoded features with image features, point cloud features, and prior features to form the fused feature F_{feat}. To further capture the global information of the features, we perform global pooling on the fused feature F_{feat}. Subsequently, it is input into a decoder composed of three parallel MLP layers, which are used to predict the rotation, translation, and scale of the target object. This process can be mathematically described by the following equations (12), (13) and (14):

$$F_{feat} = [\mathrm{MLP}(P_{model}), \mathrm{MLP}(P_{nocs}), F_{rgb}, F_{pts}, F_{pr}] \tag{12}$$

$$F_{pose} = [F_{feat}, \mathrm{AvgPool}\,(F_{feat})] \tag{13}$$

$$R = \mathrm{MLP_r}\,(F_{pose})\,, t = \mathrm{MLP_t}\,(F_{pose})\,, s = \mathrm{MLP_s}\,(F_{pose})\,. \tag{14}$$

3.5. Overall Loss Function

Our overall loss function is shown in the following equation (15):

$$L = \lambda_1 L_{pose} + \lambda_2 L_{model} + \lambda_3 L_{nocs}, \tag{15}$$

where λ_1, λ_2 and λ_3 are hyper parameters that balance the loss weights. L_{pose} is used to supervise the difference between the predicted $\{R_{pred}, t_{pred}, s_{pred}\}$ and the ground truth $\{R_{gt}, t_{gt}, s_{gt}\}$, hence the L2 loss is employed to supervise the regression task, as shown in equation (16). For the supervision loss of the instance point cloud L_{model} and the NOCS point cloud L_{nocs}, a smooth L1 loss with a threshold of 0.1 is used for both loss.

$$L_{pose} = \parallel R_{pred} - R_{gt} \parallel_2 + \parallel t_{pred} - t_{gt} \parallel_2 + \parallel s_{pred} - s_{gt} \parallel_2. \tag{16}$$

4. Experiments

4.1. Datasets

Our method was trained on the synthetic scene dataset CAMERA25 [13] and the real scene dataset REAL275 [13], and was evaluated on the test set of REAL275. The CAMERA25 dataset comprises synthetic images blended with real-world backgrounds using rendering techniques. It contains a total of 300,000 RGB-D images showcasing 1,085 distinct synthetic objects. The REAL275 dataset comprises 8,000 RGB-D images captured from 18 real-world scenarios. These images reflect the complexity and diversity of the real

world, offering the model more challenging testing conditions. Both datasets encompass six categories of objects: mug, camera, laptop, can, bowl, and bottle. The selection of these categories is intended to ensure that the model can generalize to a variety of common objects in daily life.

4.2. Implementation Details

In the implementation of this study, we drew on previous research schemes for category-level 6D pose estimation [15] and adopted the widely recognized Mask R-CNN [27] framework to generate instance masks consistent with SPD [14]. Using these masks, we precisely cropped the instances. Furthermore, using the camera's intrinsic parameters, we converted the depth images into point cloud and randomly sampled 1,024 points from each for subsequent processing. The RGB images were resized to 192×192 pixels and enhanced using standard data augmentation techniques. These techniques included random rotation, uniform noise addition, and pixel perturbation, as proposed by FS-Net [28]. The prior point cloud was provided by SPD [14], and we sampled 1,024 points from each.

In terms of feature extraction, for the image network CNN (Convolutional Neural Networks), we employed a PSP (Pyramid Scene Parsing) network [29] based on ResNet-18 [30] for semantic feature extraction. For the point cloud backbone network and the prior backbone network, we utilized PointNet++ [31] for point-level feature extraction, a method proven effective for handling point cloud data.

We set specific parameters: N=1024, d=128, and d'=64. The weights $\lambda_1, \lambda_2, \lambda_3$ in the loss function were set to 1.0, 10.0, 10.0, respectively, to balance the contributions of the different loss terms. All experiments were conducted on a single RTX4090 GPU.

4.3. Evaluation Metrics

In the category-level pose estimation task, we follow the guidance of the literature [13] and adopt the widely recognized evaluation metrics in the industry to quantify the prediction accuracy of rotation, translation, and size. Specifically, we used the following two metrics:

- **3D IoU (Intersection over Union)**: This metric assesses the degree of overlap between the predicted 3D bounding box and the actual 3D bounding box. The prediction is deemed accurate when the ratio of the overlapping volume to the union volume surpasses a certain threshold. The commonly used thresholds are 25%, 50%, and 75%.

- **n° m cm**: This metric is used to evaluate the difference between the predicted pose and the actual pose. It includes angular error (in degrees) and translational error (in centimeters). A prediction is considered accurate if the difference between the predicted and actual rotation angles is less than the threshold of n°, and the Euclidean distance between the predicted and actual translation vectors is less than the threshold of m cm. Common threshold combinations include 5°2cm, 5°5cm, 10°2cm, 10°5cm, and 10°10cm.

These evaluation metrics provide us with a standardized method to measure the performance of the model and allow us to compare with other methods in the existing literature. Through these quantitative measures, we can gain a comprehensive understanding of the model's performance under different conditions.

Table 1: Quantitative comparisons of different methods for category-level 6D object pose estimation on REAL275 dataset (%, Mean Average Precision(MAP)).

Method	3D50	3D75	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm
NOCS(2019)[13]	80.5	30.1	7.2	10	13.8	25.2	26.7
DPDN(2022)[15]	83.4	76.0	46.0	50.7	70.4	78.4	-
HS-Pose(2023)[19]	82.1	74.7	46.5	55.2	68.6	82.7	83.7
Query6Dof(2023)[20]	82.5	76.1	49.0	**58.9**	68.7	83.0	-
IST-Net(2023)[25]	82.5	76.6	47.5	53.4	72.1	80.5	82.6
VI-Net(2023)[32]	-	48.3	50.0	57.6	70.8	82.1	-
CatFormer(2024)[17]	83.1	73.8	47.7	53.7	69.0	79.5	-
Dual-COPE(2024)[16]	80.9	44.6	47.9	52.3	68.5	82.3	-
HoPENet(2024)[26]	82.7	76.2	50.5	57.9	70.6	82.5	-
DGPF6D(2024)[33]	70.2	47.5	40.8	49.6	60.5	73.9	-
CD-Pose(2024)[23]	81.0	68.6	39.8	44.9	61.8	71.6	-
RPG6D(2024)[34]	82.1	66.1	36.7	40.3	64.7	79.7	-
Diff9D(2025)[35]	76.5	41.7	35.3	43.9	54.8	70.0	-
UFGZS(2025)[36]	63.5	-	30.6	33.3	50.2	57.7	-
Ours	**83.8**	**79.0**	**52.6**	58.0	**75.7**	**83.9**	**86.0**

4.4. Comparison with Current Methods

Our method was experimentally compared with current methods in the REAL275 and CAMERA25 dataset, the results summarized in Table 1 and

Table 2: Quantitative comparisons of different methods for category-level 6D object pose estimation on CAMERA25 dataset (%, MAP).

Method	3D50	3D75	5°2cm	5°5cm	10°2cm	10°5cm
NOCS(2019)[13]	83.9	69.4	32.3	40.9	48.2	64.6
GPV-Pose(2022)[18]	93.4	88.3	72.1	79.1	-	89.0
HS-Pose(2023)[19]	93.3	88.3	73.3	80.5	80.4	89..4
VI-Net(2023)[32]	-	79.1	74.1	81.4	79.3	87.3
CD-Pose(2024)[23]	92.2	87.7	68.6	73.0	81.6	87.3
RPG6D(2024)[34]	93.4	89.2	70.2	72.6	**86.7**	90.4
CatFormer(2024)[17]	93.5	89.9	74.9	79.8	85.3	90.2
Diff9D(2025)[35]	79.8	55.8	50.5	57.1	72.1	81.5
Ours	**94.1**	**91.2**	**75.9**	**81.5**	83.7	**90.6**

Table 2. The object samples in the test set are not included in the training set. This means that all samples in the test set are previously unseen objects.

Table 1 presents the performance comparison between our method and the current methods on the REAL275 dataset. On the $3D_{75}$ metric, our method achieved an accuracy of 79.0%, an increase of 2.4% over the IST-Net [25] method. On the 5°2cm metric, we achieved a precision of 52.6%, which is 2.1% higher than the HoPENet [26] method. On the 10°2cm metric, our method surpassed IST-Net with an accuracy of 75.7%, an improvement of 3.6%. On the 10°5cm metric, our method achieved 83.9%, outperforming Query6Dof [20] by 0.9%. On the 10°10cm metric, our method slightly outperformed HS-Pose with an accuracy rate of 86.0%, an improvement of 2.3%. However, on the 5°5cm metric, our result is 0.9% lower than Query6Dof. After comparing the data, we found that Query6Dof achieves nearly 60% for the 5° metric, while our method reaches only 55%. For the 5cm metric, both methods achieve approximately 97%. However, for the 2cm metric, our method achieves nearly 86%, while Query6Dof reaches only 80%. This means that while our method is less accurate in finer rotation estimation, it outperforms other methods in more precise translation estimation.

Table 2 presents the performance comparison between our method and the current methods on the CAMERA25 dataset. As can be seen from the table 2, our method achieves the highest precision of 94.1%, 91.2%, 75.9%, 81.5%, and 90.6% on the 3D50, 3D75, 5°2cm, 5°5cm, and 10°5cm metrics, respectively, demonstrating a significant performance advantage. On the 10°2cm metric, our method achieves a precision of 83.7%, which is slightly

lower than the 86.7% of RPG6D [34], but still shows strong competitiveness overall. These results fully validate the effectiveness and robustness of the propose method in the 6D pose estimation task.

To further demonstrate the advantages of our method, we conducted reproducibility experiments using the publicly available weights of VI-Net and compared our results with theirs. As shown in Fig.3, our method outperformed VI-Net in terms of 3D IOU for the camera category (green line), where VI-Net's line begins to drop sharply at a threshold of 0.4, while our method shows a downward trend at a threshold of 0.7. In the rotation prediction for the camera category, our method also performed better than VI-Net, with our curve being steeper than VI-Net's. Under the same 10° metric, our accuracy reaches 0.4, while VI-Net's accuracy is 0.28. Under the 20° metric, our accuracy reaches 0.9, while VI-Net's accuracy is 0.79. Fig.4 visually presents the comparison between our method and VI-Net in actual performance. As highlighted in blue boxes, the laptop predicted by VI-Net exhibits significant discrepancies from the ground truth. Additionally, the camera's predicted rotations also show some deviations from the ground truth. These comparisons not only verify the effectiveness of our method but also highlight its potential advantages over existing technologies in specific categories and evaluation metrics.

4.5. Ablation Studies and Performance Analyses

Effects of Each Modules. To assess the specific impact of the proposed modules on system performance, we conducted a series of ablation studies. These studies verified the contribution of each module through different configuration combinations. The results are summarized in Table 3. First, we explored the effectiveness of the multi-attention mutual feature enhance module (MMFEM). In the experiment, we removed the MMFEM and retrained the model, with the results listed in line 1 of Table 3. By comparing the results across categories, we found that for the camera category, the absence of prior shape guidance in the input features leads to significant performance degradation across all evaluation metrics. It dropped by 9.5% on the 3D75 metric, 3% on the 5°2cm and 5°5cm metrics, 15% on the 10°2cm metric, and 23% on the 10°5cm metric. Similarly, in the mug category, both the rotation and translation metrics show significant degradation as well. It decreased by 6.2% on the 5°2cm metric, 6.9% on the 5°5cm metric, 10.8% on the 10°2cm metric, and 10.9% on the 10°5cm metric. The detailed results comparisons for these two categories can be found in Table 3. For other categories, the

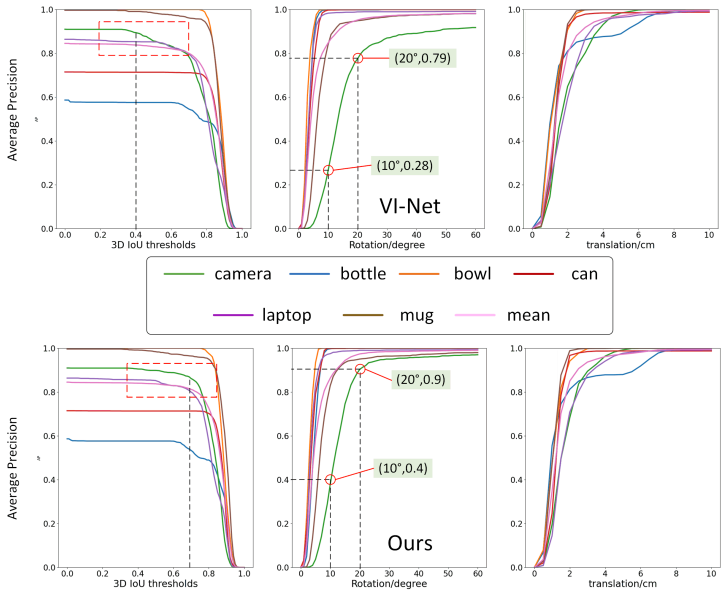

Figure 3: Comparisons with VI-Net on REAL275 in terms of average precision in 3D IoU, Rotation and Translation. In rotation estimation, our method shows significant advantages in camera categories.

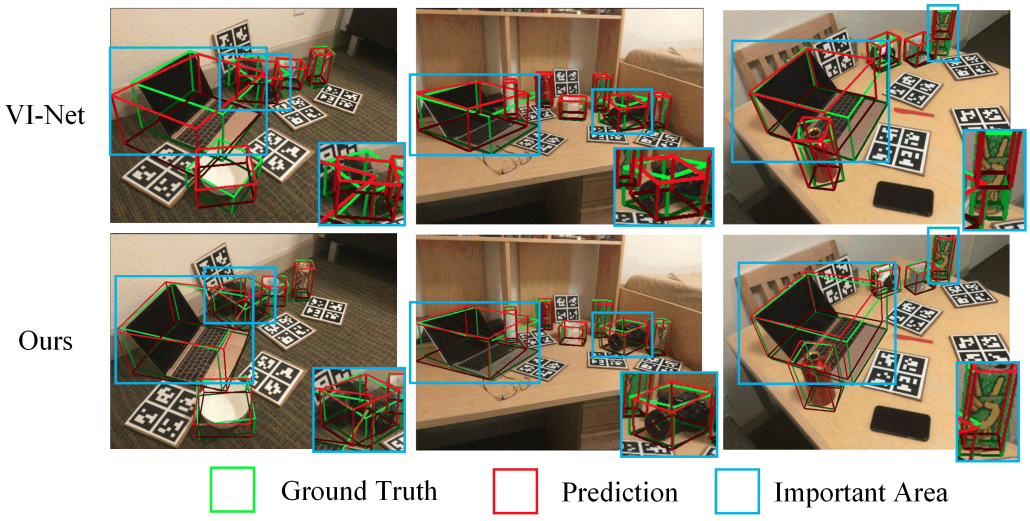

Figure 4: Qualitive comparison between Ours and VI-Net on REAL275 dataset. The red bounding boxes indicate the predicted results, while the green bounding boxes represent the ground truth.

results show little difference. This indicates that for categories with large intra-class variations (such as cameras and mugs), our proposed MMFEM method has a substantial impact.

Furthermore, in line 2 of Table 3, we completely removed the Instance Reconstruction Deformation Module (IRDM), meaning we did not predict the object instance point cloud and the NOCS point cloud. This change led to a significant decrease in overall accuracy, highlighting the key role of the IRDM module in the entire system. To verify the importance of predicting the instance point cloud, in line 3, we replaced the instance point cloud of each category with the prior point cloud while keeping the other modules unchanged. The results show that the overall accuracy suffered a significant loss. This phenomenon indicates that the prior point cloud, lacking a detailed description of each specific instance object, cannot effectively support the accurate prediction of the NOCS point cloud. In line 4, we performed instance reconstruction for all categories. In line 5, we adopted the Selective Instance Reconstruction (SIR) method, which performs instance reconstruction only for asymmetric objects, while using the prior as the instance model for symmetric objects. Compared to line 4, the SIR method exhibits 2.5% and 3% improvements in 5°2cm and 5°5cm metrics, respectively.

Through these ablation studies, we not only verified the necessity of each module but also demonstrated the synergistic effect they play in improving the accuracy of pose estimation.

Table 3: Ablation experimental results on different configurations of network architectures. MMFEM refers to Multi-attention mutual Feature Enhancement Module, IRDM refers to Instance Reconstruction Deformation Module and PM represents the instance model within the IRDM (%, MAP).

	MMFEM	IRDM	PM	3D50	3D75	5°2cm	5°5cm	10°2cm	10°5cm
1	-	✓	All	82.5	77.8	48.4	53.6	68.3	78.2
2	✓	-	-	80.0	68.1	34.2	41.4	58.8	72.5
3	✓	✓	Prior	79.0	68.6	43.5	48.7	67.9	76.0
4	✓	✓	All	83.8	79.0	50.1	55.0	75.7	83.9
5	✓	✓	SIR	**83.8**	**79.0**	**52.6**	**58.0**	**75.7**	**83.9**

Effects of Position Encoding Term. This part aims to evaluate the specific impact of positional encoding (PE) on the model's performance. We introduced positional encoding before MMFEM and recorded its effects on different evaluation metrics, with the results summarized in Table 5. On the

Table 4: The experimental results comparisons for the camera and mug categories. (%, MAP).

Categories	MMFEM	3D50	3D75	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm
CAMERA	-	82.6	72.5	0.6	0.6	13.5	14.9	14.9
CAMERA	✓	**89.9**	**82.0**	**4.6**	**4.7**	**29.0**	**37.7**	**38.0**
MUG	-	99.0	96.8	20.7	20.7	71.9	72.0	72.0
MUG	✓	98.4	95.9	**28.8**	**28.8**	**82.7**	**82.9**	**82.9**

5°2cm and 10°2cm metrics, the accuracy improved by 4.2% and 7.6%, respectively. This result indicates that on a fine scale of translation, positional encoding can effectively complement the positional information that may be lost during the feature extraction process, thus benefiting the precise regression of 6D pose estimation. The introduction of positional encoding enhances the model's perception of the spatial layout of objects, especially under strict standards for evaluating the accuracy of pose estimation. These findings emphasize the potential of positional encoding in improving the accuracy of pose estimation, especially in application scenarios where high translation accuracy is required.

Table 5: Ablation experimental results on position encoding term. PE refers to the position encoding module (%, MAP).

PE	3D50	3D75	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm
✓	**83.8**	**79.0**	**52.6**	**58.0**	**75.7**	**83.9**	**86.0**
×	83.4	76.9	45.9	55.0	68.1	81.6	83.7

Effects of input data. This part aims to evaluate the specific impact of three input data types (image, point cloud and prior shape) on the model's performance. We conducted two sets of experiments: without the image input(woRGB) and the other without the prior shape input(woPr). In the absence of the third input, the multi-attention feature enhancement module is modified accordingly. We designed two sets of cross-attention modules for mutual feature enhancement, while maintaining the rest of the structure unchanged. The results of our ablation experiments are shown in Table 6. The experimental results show that removing the image input leads to a significant drop in overall performance. This indicates that in our method, it is difficult to perform 6D pose estimation using only the point cloud and prior shape. When the prior shape input is removed, the experimental accuracy

shows a slight decline. We found that the pose estimation results for the camera category were still relatively low, which led to an overall decrease in pose estimation performance. This demonstrates that adding prior input can partially address the issue of intra-class variations, thereby enabling more accurate pose estimation.

Table 6: Ablation experimental results on input data. woRGB" refers to experiments without the image input, and "woPr" refers to experiments without the prior input. (%, MAP).

Method	3D50	3D75	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm
woRGB	77.1	56.6	21.2	23.7	46.1	54.9	56.9
woPr	83.1	75.0	48.7	54.7	68.2	77.0	79.1
Ours	**83.8**	**79.0**	**52.6**	**58.0**	**75.7**	**83.9**	**86.0**

5. DISCUSSION

This paper proposes a category-level 6D pose estimation method based on multi-attention feature enhancement and instance reconstruction. The model takes RGB images, depth images, and category priors as inputs. Through the multi-attention feature enhancement module, the three types of features are mutually enhanced. By concatenating the mutually enhanced features, the model can better represent the differences between instances of the same category, thereby reducing errors caused by intra-class variations. The instance reconstruction deformation module reconstructs instances and predicts deformation fields based on the enhanced fused features. The deformation fields are then applied to the instance point clouds to obtain NOCS (Normalized Object Coordinate Space) point clouds. Experimental results on the international standard datasets CAMERA25 and REAL275 demonstrate that the proposed methods outperforms existing methods.

The proposed method has achieved promising results in the task of category-level 6D pose estimation. In future work, we will study the pose estimation methods for complex scenarios with dynamics changes and multiobjective interferences. The complex scenarios pose higher requirements for the robustness and generalization ability of the model. Combining temporal information and active perception mechanisms is an effective way to address this challenge. By using temporal information, the motion trajectories of the target in consecutive video frames will be tracked to compensate for the

insufficiency of instantaneous image information and enhance the stability of pose estimation in dynamic scenes. The active perception mechanism will actively adjust the data acquisition strategy according to the current uncertainty and requirements to obtain more discriminative information.

6. CONCLUSIONS

In this work, we propose a novel method called MMFEIR for category-level 6D object pose estimation. Our work consists of two main components: the multi-attention mutual feature enhance module and the instance reconstruction deformation module. The former leverages attention mechanisms to enhance the interaction between multi-modal input feature values and improve feature expression capabilities. The latter utilizes the enhanced feature values to design an instance reconstruction deformation module, which can accurately reconstruct the instance point cloud. IRDM enhances the ability to understand the geometric configuration differences between different instance point cloud within the same category, thereby improving the accuracy of the 6D pose estimation. The results of extensive experiments on the challenging benchmark demonstrate that the proposed method outperforms several existing approaches. The results of this research contribute to advances in cutting-edge domains such as robotic grasping, intelligent robotic manipulation, and augmented reality. In future work, we will study the pose estimation algorithms for complex scenarios with dynamics changes and multiobjective interferences and further expand the application fields of pose estimation technology.

Acknowledgments

This research was supported by National Natural Science Foundation of China(No.61972091), Natural Science Foundation of Guang-dong Province of China(No.2022A1515010101), Research Projects of Ordinary Universities in Guangdong Province under Grant(No.2023KTSCX133), Guangdong Basic and Applied Basic Research Foundation under Grant (No.2022A1515140103), Featured Innovation Project of Foshan Education Bureau(2022DZXX06).

References

[1] K. Wada, E. Sucar, S. James, D. Lenton, A. J. Davison, Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion,

in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 14540–14549.

[2] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, K. Bousmalis, Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12627–12637.

[3] P. Cipresso, I. A. C. Giglioli, M. A. Raya, G. Riva, The past, present, and future of virtual and augmented reality research: a network and cluster analysis of the literature, Frontiers in psychology 9 (2018) 2086.

[4] M. Gattullo, G. W. Scurati, M. Fiorentino, A. E. Uva, F. Ferrise, M. Bordegoni, Towards augmented reality manuals for industry 4.0: A methodology, robotics and computer-integrated manufacturing 56 (2019) 276–286.

[5] Y. Nie, X. Han, S. Guo, Y. Zheng, J. Chang, J. J. Zhang, Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 55–64.

[6] C. Zhang, Z. Cui, Y. Zhang, B. Zeng, M. Pollefeys, S. Liu, Holistic 3d scene understanding from a single image with implicit representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 8833–8842.

[7] H. Sun, T. Wang, E. Yu, A dynamic keypoint selection network for 6dof pose estimation, Image and Vision Computing 118 (2022) 104372.

[8] A. Trabelsi, M. Chaabane, N. Blanchard, R. Beveridge, A pose proposal and refinement network for better 6d object pose estimation, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 2382–2391.

[9] G. Wang, F. Manhardt, F. Tombari, X. Ji, Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 16611–16621.

[10] Z. Jiang, X. Wang, X. Huang, H. Li, Triangulate geometric constraint combined with visual-flow fusion network for accurate 6dof pose estimation, Image and Vision Computing 108 (2021) 104127.

[11] Z. Chai, C. Liu, Z. Xiong, Multi-pyramid-based hierarchical template matching for 6d pose estimation in industrial grasping task, Industrial Robot: the international journal of robotics research and application 50 (2023) 659–672.

[12] Y. Wu, Y. Fu, S. Wang, Deep instance segmentation and 6d object pose estimation in cluttered scenes for robotic autonomous grasping, Industrial Robot: the international journal of robotics research and application 47 (2020) 593–606.

[13] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, L. J. Guibas, Normalized object coordinate space for category-level 6d object pose and size estimation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2642–2651.

[14] M. Tian, M. H. Ang, G. H. Lee, Shape prior deformation for categorical 6d object pose and size estimation, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16, Springer, 2020, pp. 530–546.

[15] J. Lin, Z. Wei, C. Ding, K. Jia, Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks, in: European Conference on Computer Vision, Springer, 2022, pp. 19–34.

[16] X. Ren, N. Guo, Z. Zhu, X. Jiang, Dual-cope: Category-level object pose estimation with dual sim2real unsupervised domain adaptation module, Available at SSRN 4772808 (2024).

[17] S. Yu, D.-H. Zhai, Y. Xia, Catformer: Category-level 6d object pose estimation with transformer, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, 2024, pp. 6808–6816.

[18] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, F. 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, 2022, pp. 6781–6791.

[19] L. Zheng, C. Wang, Y. Sun, E. Dasgupta, H. Chen, A. Leonardis, W. Zhang, H. J. 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, 2023, pp. 17163–17173.

[20] R. Wang, X. Wang, T. Li, R. Yang, M. Wan, W. 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, 2023, pp. 14055–14064.

[21] Z. Li, Y. Hu, M. Salzmann, X. Ji, Sd-pose: Semantic decomposition for cross-domain 6d object pose estimation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, 2021, pp. 2020–2028.

[22] J. Liu, W. Sun, C. Liu, X. Zhang, Q. Fu, Robotic continuous grasping system by shape transformer-guided multiobject category-level 6-d pose estimation, IEEE Transactions on Industrial Informatics 19 (2023) 11171–11181.

[23] L. Zou, Z. Huang, N. Gu, G. Wang, Learning geometric consistency and discrepancy for category-level 6d object pose estimation from point clouds, Pattern Recognition 145 (2024) 109896.

[24] S. Umeyama, Least-squares estimation of transformation parameters between two point patterns, IEEE Transactions on Pattern Analysis & Machine Intelligence 13 (1991) 376–380.

[25] J. Liu, Y. Chen, X. Ye, X. Qi, Ist-net: Prior-free category-level pose estimation with implicit space transformation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 13978–13988.

[26] C. Jiang, X. Mu, B. Zhang, M. Xie, C. Liang, Category level object pose estimation via global high-order pooling, Electronics 13 (2024) 1720.

[27] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2961–2969.

[28] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, A. 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, 2021, pp. 1581–1590.

[29] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2881–2890.

[30] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

[31] C. R. Qi, L. Yi, H. Su, L. J. Guibas, Pointnet++: Deep hierarchical feature learning on point sets in a metric space, Advances in neural information processing systems 30 (2017).

[32] J. Lin, Z. Wei, Y. Zhang, K. Jia, Vi-net: Boosting category-level 6d object pose estimation via learning decoupled rotations on the spherical representations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 14001–14011.

[33] J. Liu, W. Sun, H. Yang, C. Liu, X. Zhang, A. Mian, Domain-generalized robotic picking via contrastive learning-based 6-d pose estimation, IEEE Transactions on Industrial Informatics (2024).

[34] H. Sun, Y. Zhang, H. Sun, K. Hashimoto, Refined prior guided category-level 6d pose estimation and its application on robotic grasping, Applied Sciences 14 (2024) 8009.

[35] J. Liu, W. Sun, H. Yang, P. Deng, C. Liu, N. Sebe, H. Rahmani, A. Mian, Diff9d: Diffusion-based domain-generalized category-level 9-dof object pose estimation, arXiv preprint arXiv:2502.02525 (2025).

[36] W. Qu, C. Meng, H. Li, J. Cheng, C. Ma, H. Wang, X. Zhou, X. Deng, P. Tan, Universal features guided zero-shot category-level object pose estimation, arXiv preprint arXiv:2501.02831 (2025).