2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Implicit Coarse-to-Fine 3D Perception for Category-level Object Pose Estimation from Monocular RGB Image

Jia Li[1,2], Li Jin[1,2], Xibin Song,[3] Yeheng Chen[4], Nan Li[4,*], Xueying Qin[1,2,*]

***Abstract*— Category-level object pose estimation demonstrates robust generalization capabilities that benefit robotics applications. However, exclusive reliance on RGB images without leveraging any 3D information introduces ambiguity in the translation and size of objects, leading to suboptimal performance. In this paper, we propose a framework for category-level pose estimation from a single RGB image in an end-to-end manner, *i.e.,* Feature Auxiliary Perception Network (FAP-Net). To address inaccurate pose estimation caused by the inherent ambiguity of RGB images, we design a coarse-to-fine approach that first harnesses geometry supervision to facilitate coarse 3D feature perception and subsequently refines the features based on pose and size constraints. Experimental results on REAL275 and CAMERA25 demonstrate that FAP-Net achieves significant improvements (14.7% on 10°10cm and 11.4% on IoU50 on the real-scene REAL275 dataset) over the state-of-the-art and real-time inference (42 FPS).**

I. INTRODUCTION

Category-level object pose estimation predicts the translation, rotation and size of previously unseen objects [1] occupies a significant role in real-world applications such as robotics [2], [3], augmented reality (AR) [4] and 3D scene understanding [5]. The category-level methods, independent of CAD models, need to represent the NOCS canonical space of objects (explicit point cloud or implicit features in the object-coordinate system) based on observed data. This representation is used for aligning with the camera-coordinate system to achieve object pose estimation.

Nonetheless, the complexity lies in inferring the geometry of objects as a prior to representing the object-coordinate system when relying solely on a single RGB image from a partial observation. As illustrated in Fig.1. (a), when translation estimation is inaccurate, perfect fitting of the observation can still be achieved by adjusting the object's size. This situation is further ambiguous due to a lack of geometry information, and the challenge becomes particularly pronounced when the object is previously unseen.

To mitigate the ambiguity in translation and size when performing category-level pose estimation, approaches [6], [7] are proposed which incorporate depth captured by sensors to confine the searching scope of translation within a narrower range. However, the utilization of depth maps is restricted by equipment availability and application contexts, whereas RGB images are more prevalent and readily accessible.

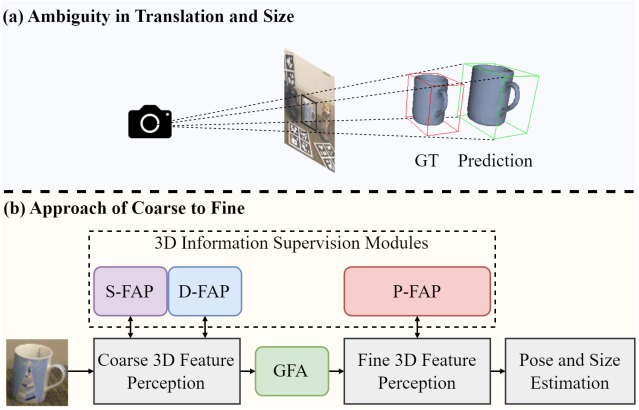

Fig. 1. Illustration of the ambiguity problem and our framework. The ground truth (GT) in the red box and the prediction in the green box with different poses and sizes can be projected onto the same observation. We design 3D information supervision modules to facilitate implicit coarse-to-fine perception.

When relying solely on RGB images, category-level pose estimation encounters a severe degradation issue [8] due to the projection of 3D spatial geometry onto 2D images. In the absence of complete geometry about unseen objects, accurate pose estimation based on a single RGB image becomes an ill-posed problem. RGB-based methods [9], [10] address this hurdle by estimating the absolute depth of objects. However, they can only approximate the relatively coarse depth maps. Besides, relying merely on the estimated depth maps for pose estimation is insufficient and imprecise.

To tackle these problems, we present a novel end-to-end framework, *i.e.,* Feature Auxiliary Perception Network (FAP-Net), explicitly designed for category-level object pose estimation based on RGB images. We are dedicated to progressively perceiving 3D implicit features from the monocular image, which can benefit from the rich supervision provided by the proposed modules. We introduce a coarse-to-fine 3D feature perception strategy (Fig. 1. (b)) to mitigate issues correlated with ambiguity in translation and size. In the coarse perception phase, we employ 3D information (point cloud and depth maps) as auxiliary supervision (S-FAP and D-FAP) to facilitate representing the geometry of objects, which is not evident in RGB images. In the refinement phase, we transform geometry features into high-level pose features encompassing camera-coordinate features and object-coordinate features, guided by the supervision of pose and size (P-FAP).

The main contributions of FAP-Net are summarized as

This work is supported by the National Key R&D Program of China (No. 2022YFB3303203), NSF of China (No. U21B6001, No. 62172260), and Shandong Provincial Natural Science Foundation (No. ZR2022LZH007).

[1] School of Software, Shandong University, P.R. China
[2] Engineering Research Center of Digital Media Technology, Ministry of Education, P.R. China
[3] XR Vision Labs, Tencent, P.R. China
[4] Intelligent Robot Research Center, Zhejiang Lab, P.R. China
[*] Corresponding Author.

follows:

- Our work introduces an innovative end-to-end network for RGB-based category-level object pose estimation, which employs a coarse-to-fine strategy to extract 3D features, effectively addressing the ambiguity issue in translation and size associated with pose estimation.
- Two essential modules are designed as supervision to extract low-level geometry features during the coarse stage: the Shape-based 3D Feature Auxiliary Perception (S-FAP) module and the Depth-based 3D Feature Auxiliary Perception (D-FAP) module.
- We introduce the Pose-based 3D Feature Auxiliary Perception (P-FAP) module, which employs a siamese architecture to supervise the extraction of high-level pose features during the refinement stage.
- To demonstrate the effectiveness of FAP-Net, we conduct extensive experiments on REAL275 and CAMERA75 datasets and achieve state-of-the-art performance and real-time inference (42 FPS).

II. RELATED WORK

A. Instance-Level Object Pose Estimation

Instance-level object pose estimation can utilize objects' accurate CAD models. CAD models provide ample information about the objects' shape and size, mitigating the ambiguity issue mentioned above.

For methods grounded in RGB-D modalities, DenseFusion [11] leverages depth information and dense feature fusion for pose estimation. For less computational complexity and more robustness, PVN3D [12] predicts 3D keypoints of the target object, aligns them with the object's CAD model, and ultimately achieves pose estimation through the iterative closest point (ICP) algorithm. Furthermore, OVE6D [13] is trained on synthetic data and demonstrates generalization to real-world objects utilizing depth and masks.

For RGB-based methods, PoseNet [14] employs convolutional neural networks to estimate the 6D camera pose end-to-end solely from a single RGB image. To effectively enhance pose accuracy, PoseCNN [15] estimates translation by predicting the object's center position depth in RGB images and then utilizes quaternion representation for rotation estimation. However, it suffers from limitations in handling occlusions. To this end, PVNet [16] employs a per-pixel voting strategy to predict 2D keypoints on RGB images and address the Perspective-n-Point (PnP) problem. Unlike the sparse keypoint correspondence in [16], DPOD [17] establishes dense 2D-3D correspondence for better robustness. Moreover, SMOC-Net [18] employs a knowledge distillation framework to achieve pose estimation for self-supervised pose estimation, exhibiting superior generalization. Despite the relative maturity of instance-level object pose estimation, it is hindered by the dependence on the availability of CAD models.

B. Category-Level Object Pose Estimation

CAD models of objects are unavailable in category-level object pose estimation. The method [1] introduces the concept of category-level and proposes Normalized Object Coordinate Space (NOCS) as the object-coordinate system representation. It defines a normalized canonical space for intra-category shapes, predicts pixel-level NOCS representations, and ultimately resolves the alignment with depth maps. Several other methods, such as SPD [19], SGPA [20] and DPDN [6], consider it challenging to predict NOCS representations directly. These methods fuse intra-category instance models into categorical shape priors using auto-encoder networks and derive NOCS representations based on shape priors through deformation. Following this, IST-Net [7] implicitly expresses the deformation process by building correspondence between features in different coordinate systems, thereby eliminating the necessity for shape priors. Based on 3DGC [21], other methods [22], [23], [24], [25], extract local geometry relationships from observed point cloud and directly solve pose estimation without using deformation priors. Nevertheless, depth information is indispensable for these methods.

Methods [9], [10] are dedicated to category-level object pose estimation based on RGB images. These methods acquire geometry information from RGB images by estimating depth maps in the camera-coordinate system. Furthermore, they predict the NOCS representation and achieve pose and size estimation through alignment between coordinate systems. To be more specific, [9] employs a two-stage approach, using the metric-scale object mesh obtained from Mesh-RCNN [26] as an intermediary to predict depth maps. However, it is worth noting that predicting intermediate representations introduces additional computational costs. In contrast, OLD-Net [10], based on SPD [19] framework, decouples rotation and translation to reconstruct object-level depth maps directly. Nonetheless, depth estimation approaches are limited to approximating coarse depth maps. It becomes apparent that more complex geometry constraints are needed to address the issue of ambiguity and achieve precision in pose estimation.

III. METHOD

We focus on RGB-based category-level object pose estimation, which is deeply affected by the ambiguity in translation and size. For clarity in subsequent sections, (H, W) represents the size of the RGB image, N represents the number of sampling object points corresponding to the RGB image, and the superscript " $\sim$ " denotes the variables used for supervision.

A. Overview

We present an end-to-end framework FAP-Net, as shown in Fig.2, which achieves accurate RGB-based category-level object pose estimation through a coarse-to-fine strategy. FAP-Net consists of a) Coarse Geometry Feature Perception; b) Fine Pose Feature Perception; and c) Object Pose and Size Estimation.

Geometry information embeds within the 3D observed data. Employing this information for supervision enables the network to infer object geometry, thereby mitigating

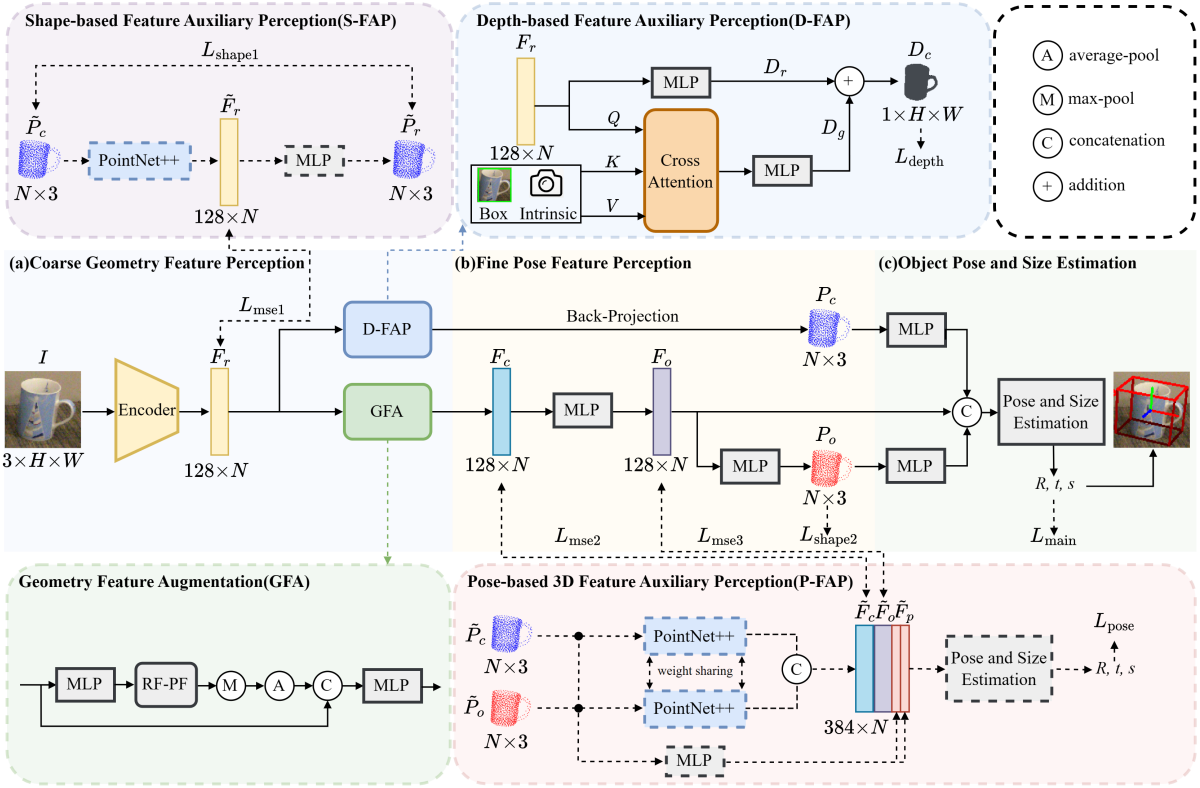

Fig. 2. Illustration of FAP-Net. The framework is divided into three stages: (a) coarse geometry feature perception, (b) fine pose feature perception, and (c) object pose and size estimation. The core modules include S-FAP, D-FAP, GFA, and P-FAP, where S-FAP and P-FAP depicted with dashed lines are used only during the training phase.

ambiguity. Simultaneously, pose estimation necessitates a further understanding of the NOCS canonical space. The coarse-to-fine strategy empowers us to deduce high-level pose features from geometry features efficiently. At each stage, we leverage 3D information to supervise intermediate features, providing insightful constraints.

During the training stage, we first extract coarse geometry features $F_r \in \mathbb{R}^{128\times N}$ from the RGB image $I \in \mathbb{R}^{3\times H\times W}$ with the supervision of S-FAP and D-FAP (Sec. III-B). Then, the coarse features F_r will be refined into the pose features encompassing camera-coordinate features $F_c \in \mathbb{R}^{128\times N}$ and object-coordinate features $F_o \in \mathbb{R}^{128\times N}$ with the supervision of P-FAP (Sec. III-C). Finally, predicted P_c and P_o are encoded and concatenated with F_c and F_o for pose estimation in the Object Pose and Size Estimation module (Sec. III-D).

During the inference stage, only D-FAP and GFA are utilized. FAP-Net can directly extract the geometry features from a single RGB image and transform them into the pose features, ultimately enabling the regression of the object's pose and size. The overall loss function is presented in Sec. III-E.

B. Coarse Geometry Feature Perception

Considering that a single RGB image provides only a 2D observation of a 3D object, inferring the comprehensive geometry of the object becomes challenging. Nevertheless, geometry information is inherently present in 3D observed data. Therefore, we aid the network in perceiving the geometry features through 3D supervision.

Given the RGB image I, we first use the CNN-based multi-scale feature extraction network [27] and select N pixels corresponding to object points to obtain semantic features F_r. Then, the features are supervised by the S-FAP module, which is based on auto-encoder architecture.

Shape-based 3D Feature Auxiliary Perception (S-FAP): The PointNet++ [28] based auto-encoder is applied to retrieve the shape features $\tilde{F}_r$ from the point cloud $\tilde{P}_c$ which is projected from the depth map corresponding to I. Subsequently, $\tilde{F}_r$ is decoded into the point cloud $\tilde{P}_r$ through a Multi-Layer Perception (MLP). $\tilde{P}_r$ should closely resemble $\tilde{P}_c$ for constraining $\tilde{F}_r$ to represent the shape properties well. We adopt smooth-L1 loss as the constraint:

$$L_{\text{shape1}} = L_{\text{SL1}}\left(\tilde{P}_c, \tilde{P}_r\right), \tag{1}$$

Then, shape features $\tilde{F}_r$ is used to directly supervise F_r as:

$$\begin{aligned} L_{\text{mse1}} &= L_{\text{mse}}\left(F_r, \tilde{F}_r\right) \\ &= \frac{1}{n\cdot d}\sum_{i=1}^{n}\sum_{j=1}^{d}\left\| f_r^{ij} - \tilde{f}_r^{ij}\right\|_2, \end{aligned} \tag{2}$$

Intermediate features $\tilde{F}_r$ in the S-FAP module serve as standard geometry shape representations, providing direct constraints on intermediate features.

Depth-based 3D Feature Auxiliary Perception: Given F_r as input, an MLP operation estimates the depth map. It's important to clarify that the input F_r refers to the image

features without being filtered by pixel-to-point correspondences. As the optimization objective focuses on features corresponding to pixel-to-point pairs, they are uniformly represented as F_r in Fig.2. Due to the crop and resize of the input image, global object localization information in the scene is lost, further intensifying the problem of ambiguity. To overcome the issue, we employ the attention mechanism within the D-FAP module to fuse the information from the detection box and camera intrinsic parameters with image features F_r, predicting the compensation for the depth map. Following [29], [10], we normalize the detection box and camera intrinsic parameters as Global Information (GI) of the image:

$$GI = \left[\frac{f_x}{r-l}, \frac{f_y}{b-t}, \frac{l-c_x}{f_x}, \frac{t-c_y}{f_y}, \frac{r-c_x}{f_x}, \frac{b-c_y}{f_y} \right], \quad (3)$$

where (l, t) and (r, b) denote the pixel coordinates of the top-left and bottom-right corners of the cropped image. f_x, f_y, c_x and c_y represent focal lengths and optical center of the camera. Subsequently, we utilize the cross-attention to fuse F_r with GI to assist depth compensation. The accompanying video illustrates the detailed structure of the Cross Attention module.

The fused features are fed into an MLP to predict one uniform depth value applied to each pixel for depth compensation. The final estimated depth map D_c is the addition of the rough depth map D_r and the depth compensation D_g. To generate more accurate depth maps, in addition to pixel depth values, we also utilize standard image features, gradients, and normals for supervision:

$$\begin{aligned} L_{\text{depth}} = l_{\text{depth}} \left(D_c, \tilde{D}_c \right) + l_{\text{gradient}} \left(D_c, \tilde{D}_c \right) \\ + l_{\text{normal}} \left(D_c, \tilde{D}_c \right), \end{aligned} \quad (4)$$

where $\tilde{D}_c$ is the ground truth. The details of the loss function is available in the accompanying video.

Geometry Feature Augmentation: The receptive field of 2D convolution is limited to the image plane. Expanding the receptive field capability is essential for perceiving the pose features. Given that we predict the depth map and back-project it into the point cloud in D-FAP. We can derive pixels corresponding to points with close distances and similar features. Referring to [21], [25], we jointly consider receptive field with point distance and feature distance, forming a more robust 3D-capable receptive field, denoted as RF-PF. It can simultaneously focus on local structural points with close distances and global points with similar features, and perform well in 2D-3D tasks.

C. Fine Pose Feature Perception

We preliminarily perceive coarse geometry features from RGB images. However, object pose estimation necessitates further inference on high-level pose features. Referring to IST-Net [7], P-FAP offers implicit feature representations in both camera and NOCS space, directly supervising the intermediate features. IST-Net focuses on the transformation between different coordinate features, while P-FAP is based on siamese architecture to jointly supervise different coordinate features, enhancing pose correspondence.

In detail, we take the output of the GFA module as the camera-coordinate features F_c and employ an MLP to transform F_c into the object-coordinate features F_o, which is then used to regress the point cloud P_o in the object coordinate by another MLP. We use smooth L1 loss to constrain F_o in representing the NOCS:

$$L_{\text{shape2}} = L_{\text{SL1}} \left(P_o, \tilde{P}_o \right), \quad (5)$$

Pose-based 3D Feature Auxiliary Perception: PointNet++ [28] based on siamese architecture is applied to extract the pose features $\tilde{F}_c$ and $\tilde{F}_o$ from point cloud $\tilde{P}_c$ and $\tilde{P}_o$ in camera and object coordinates, and an MLP is used for position encoding $\tilde{F}_p$. Among them, $\tilde{F}_c$ and $\tilde{F}_o$ are used to supervise the features F_c and F_o in the corresponding coordinate system of the main network by Mean Square Error (MSE) as:

$$L_{\text{mse2}} = L_{\text{mse}} \left(F_c, \tilde{F}_c \right), \quad (6)$$

$$L_{\text{mse3}} = L_{\text{mse}} \left(F_o, \tilde{F}_o \right), \quad (7)$$

$\tilde{F}_c$, $\tilde{F}_r$ and $\tilde{F}_p$ are concatenated and fed into Pose and Size Estimation module for the regression of pose and size.

It should be noted that siamese architecture offers a distinct advantage, as it learns a unified mapping that projects point clouds from different coordinate systems into one feature space related to pose and size. On the one hand, it simultaneously learns different coordinate systems of point clouds, enhancing its ability to perceive high-level information about pose and size. On the other hand, it strengthens the connection between the features of different coordinate systems, aiding in alignment.

D. Object Pose and Size Estimation

Following [6], [7], we directly regress the object's translation, rotation and size from the pose features and position encoding through the Pose and Size Estimation module architecture. Based on the predicted pose and size, we can visualize the 3D bounding box on the RGB image to measure the estimation result visually. It shares the same formulation as the loss employed in P-FAP:

$$L_{\text{main}} = \|R_{\text{pre}} - R_{\text{gt}}\|_2 + \|t_{\text{pre}} - t_{\text{gt}}\|_2 + \|s_{\text{pre}} - s_{\text{gt}}\|_2, \quad (8)$$

where R_{pre}, t_{pre}, and s_{pre} are defined as the predicted parameters for rotation, translation and size.

E. Loss Function

In summary, the loss function employed in our framework is as follows:

$$L = \begin{cases} L_{\text{shape1}} + L_{\text{depth}} + L_{\text{mse1}}, & n < 10 \\ \lambda(L_{\text{mse2}} + L_{\text{mse3}}) + L_{\text{shape2}} + L_{\text{pose}} + L_{\text{main}}, & n \geqslant 10 \end{cases} \quad (9)$$

where n defines the training epoch, λ is a hyper-parameter, and it is set to 10.

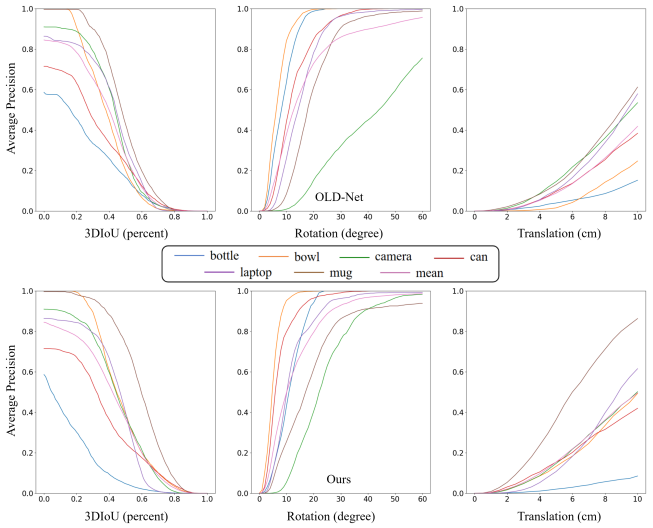

Fig. 3. Per-category quantitative comparison with OLD-Net [10] on REAL275.

IV. EXPERIMENTS

A. Datasets and Evaluation Metrics

Datasets: Our experiments are conducted on two widely used benchmark datasets REAL275 and CAMERA25. Both datasets include 6 categories of objects (bottle, bowl, camera, can, laptop, and mug). REAL275 captures 7k RGB-D images across 13 real-world scenes, 4.3k for training, and 2.75k for evaluation. CAMERA25 is a syntactic dataset containing 300k RGB-D images, and 2.75k images are for evaluation.

Evaluation Metrics: 3D intersection over union (3D IoU) and rotation/translation errors are standard metrics for category-level object pose estimation. Following [10], we report 3D IoU with thresholds of 25%, 50%, and 75%. Besides, the metrics of 10cm, 10°, and the combination metric 10°10cm are employed for measuring rotation and translation.

B. Implementation Details

Following previous work [1], [7], the pre-trained Mask-RCNN is employed to segment the objects from the images. The size of the RGB image is resized to 256 × 256, and the number of object points N is set to 1024. FAP-Net is trained on 2 NVIDIA RTX3090 GPUs with a batch size of 24, and the training epoch is set to 20. We use the training set of CAMERA25 for its evaluation, and for the evaluation of REAL275, both training sets are utilized. An integrated model is trained for 6 categories. For robot experiments, we utilize the Azure Kinect DK as the eye-out-of-hand visual sensor to capture RGB images and employ the UR10E robotic arm for object manipulation.

C. Comparison with State-of-the-Arts

In Table I, we present the performance of the FAP-Net in comparison to the state-of-the-art RGB-based methods. It should be noted that IST-Net [7] is the RGB-D based method, and we remove the depth input for a fair comparison. FAP-Net achieves the SOTA results across all metrics on both datasets. Surprisingly, we outperform the second-ranking method OLD-Net [10] by 11.4% and 14.7% on IoU50 and 10°10cm. As for CAMERA25, our method still demonstrates superiority over OLD-Net (7.1% on IoU25, 7.1% on IoU50, and 6.4% on 10°10cm). Importantly, when the depth input in IST-Net is removed without 3D information perception, it is severely affected by ambiguity issues in translation and size. The outstanding performance demonstrates the effectiveness of FAP-Net in utilizing the 3D information auxiliary module for coarse-to-fine supervision.

Furthermore, we present the quantitative performance of our method and OLD-Net on REAL275 in Fig. 3. Notably, we outperform on IoU and translation metrics in most categories. In addition, we perform better on the rotation metric, particularly for categories with large shape variations, such as camera. Qualitative visualization is shown in Fig. 4. Our method demonstrates improved pose and size estimation, even when the partial occlusion exists (*e.g.* the laptop in the last column, which is partially obscured by the mug).

To enhance the analysis of improvement sources, we compare chamfer distance errors of reconstructed point clouds back-projected from estimated depth maps with OLD-Net in Table II on REAL275. Our reconstruction accuracy is significantly better than the object-level reconstruction method like bowl and mug. When the accuracy is comparable, superior pose estimation is achieved for the camera, can, and laptop. However, poor reconstruction adversely impacts pose estimation, such as the bottle.

We also conduct robot grasping experiments to validate the robustness of FAP-Net in practical applications. Fig. 5 displays our results on object pose estimation in a real-world scene. We select appropriate grasp points related to categories for motion planning and grasping. Besides, our method runs in 42FPS on a single NVIDIA RTX 3090. In contrast, while OLD-Net has an inference time of 24ms, it requires an additional 22ms for ICP, underscoring the superiority of our end-to-end approach.

D. Ablation Study

We provide ablation studies on REAL275 and validate the effectiveness of the proposed modules.

Ablation on S-FAP and D-FAP modules: We first ablate S-FAP and D-FAP modules in Table III. With S-FAP, we can observe that the metric of 10°increases from 39.5% to 47.4%, and there is also an improvement in the relatively loose IoU metrics (5.8% on IoU25). The reason is that S-FAP supervises 3D shape information, which can effectively aid in inferring the geometry of objects from RGB images. Subsequently, we only add the D-FAP module to the baseline. It can be found that the metric of 10cm increases from 42.0% to 46.3% and the metric of 10°increases from 39.5% to 44.1%. It should be noted that the D-FAP module not only provides absolute depth supervision but also estimates depth maps for subsequent processes. Therefore, we replaced D-FAP with a depth estimation network [27]. Finally, with the collaboration of S-FAP and D-FAP, our method performs best on all metrics.

TABLE I

COMPARISONS WITH STATE-OF-THE-ART METHODS ON REAL275 AND CAMERA25 DATASETS. OVERALL BEST RESULTS ARE SHOWN IN BOLD, AND THE SECOND-BEST RESULTS ARE UNDERLINED.

Method	REAL275						CAMERA25					
	IoU25	IoU50	IoU75	10cm	10°	10°10cm	IoU25	IoU50	IoU75	10cm	10°	10°10cm
IST-Net [7]	24.9	2.4	0.1	0.6	22.6	0.2	59.5	14.7	0.6	6.2	71.2	6.2
Synthesis [30]	-	-	-	34.0	14.2	4.8	-	-	-	-	-	-
lee et al. [9]	62.0	23.4	3.0	39.5	29.2	9.6	75.5	32.4	5.1	29.7	60.8	19.2
OLD-Net [10]	68.7	25.4	1.9	38.9	37.0	9.8	74.3	32.1	5.4	30.1	74.0	23.4
Ours	**74.2**	**36.8**	**5.2**	**49.7**	**49.6**	**24.5**	**81.4**	**39.2**	**6.7**	**36.0**	**80.4**	**29.8**

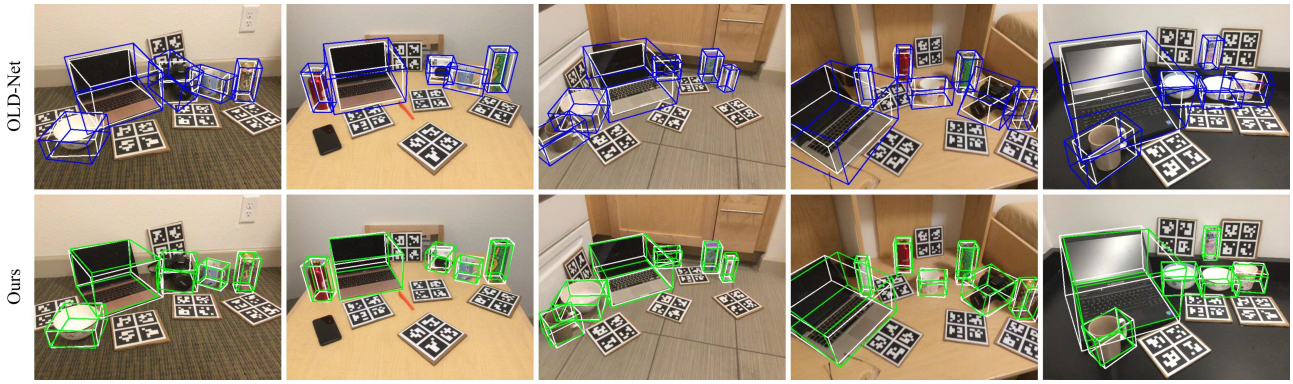

Fig. 4. Qualitative results of the FAP-Net (green) and the OLD-Net (blue). The ground truth bounding boxes are outlined with white lines.

TABLE II

RECONSTRUCTION QUALITY EVALUATION ON CHAMFER DISTANCE.

Method	bottle	bowl	camera	can	laptop	mug
OLD-Net	0.0179	0.0200	0.0136	0.0162	0.0151	0.0129
Ours	0.0510	0.0090	0.0140	0.0215	0.0146	0.0065

TABLE III

ABLATION EXPERIMENTS ON S-FAP AND D-FAP MODULES.

S-FAP	D-FAP	IoU25	IoU50	IoU75	10cm	10°	10°10cm
		65.6	28.2	4.5	42.0	39.5	19.0
✓		71.4	30.1	3.4	42.0	47.4	19.8
	✓	72.9	34.2	4.7	46.3	44.1	20.7
✓	✓	**74.2**	**36.8**	**5.2**	**49.7**	**49.6**	**24.5**

TABLE IV

ABLATION EXPERIMENTS ON P-FAP MODULE. P-FAP-C REFERS TO L_{mse2} AND P-FAP-O REFERS TO L_{mse3}.

P-FAP-C	P-FAP-O	IoU25	IoU50	IoU75	10cm	10°	10°10cm
		69.8	30.7	4.3	45.4	46.3	20.1
✓		71.9	31.6	3.3	45.2	**49.9**	22.1
	✓	72.7	31.7	2.6	46.6	46.4	21.1
✓	✓	**74.2**	**36.8**	**5.2**	**49.7**	49.6	**24.5**

TABLE V

ABLATION ON SIAMESE NETWORK ARCHITECTURE IN P-FAP MODULE. PARAM. REFERS TO THE NUMBER OF PARAMETERS.

Siamese	IoU25	IoU50	IoU75	10cm	10°	10°10cm	Param.
w	**74.2**	**36.8**	**5.2**	**49.7**	49.6	**24.5**	**19M**
w/o	71.5	31.7	3.8	46.2	**50.8**	22.3	21M

TABLE VI

ABLATION ON GEOMETRY FEATURE AUGMENTATION MODULE.

GFA	IoU25	IoU50	IoU75	10cm	10°	10°10cm
w	**74.2**	**36.8**	**5.2**	**49.7**	**49.6**	**24.5**
w/o	73.4	31.0	4.0	45.7	49.0	22.4

Fig. 5. Object pose and size estimation for robot grasping.

Ablation on P-FAP module: We split P-FAP into P-FAP-C and P-FAP-O, representing the supervision of intermediate features F_c and F_o respectively for better analysis. As shown in Table IV, when supervising the features of both coordinate systems simultaneously, we achieve the best performance, with 6.1% improvement on IoU50 and 4.4% on 10°10cm compared to baseline. P-FAP can extract high-level knowledge related to pose and size which facilitates estimation.

Ablation on siamese network architecture: As shown in Table V, with the additional parameters, although there is a 1.2% improvement on 10° metric, the performance deteriorates on other metrics.

Ablation on GFA module: We also validate the effectiveness of the GFA module in Table VI. Without the GFA module, the performance significantly decreases on IoU50 and 10cm metrics by 5.8%.

V. CONCLUSION

In this paper, we present the Feature Auxiliary Perception Network for RGB-based category-level object pose estimation to mitigate the ambiguity in translation and size. We utilize supervision from 3D information modules to assist the main network in coarse-to-fine 3D feature perception. Qualitative results on the REAL275 and CAMERA25 datasets, as well as real-scene robotic grasping, demonstrate the outperformance and robustness of FAP-Net.

REFERENCES

[1] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas, "Normalized object coordinate space for category-level 6d object pose and size estimation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 2642–2651, Computer Vision Foundation / IEEE, 2019.

[2] X. Deng, Y. Xiang, A. Mousavian, C. Eppner, T. Bretl, and D. Fox, "Self-supervised 6d object pose estimation for robot manipulation," in *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pp. 3665–3671, IEEE, 2020.

[3] T. Nie, J. Ma, Y. Zhao, Z. Fan, J. Wen, and M. Sun, "Category-level 6d pose estimation using geometry-guided instance-aware prior and multi-stage reconstruction," *IEEE Robotics Autom. Lett.*, vol. 8, no. 4, pp. 2381–2388, 2023.

[4] Y. Su, J. R. Rambach, N. Minaskan, P. Lesur, A. Pagani, and D. Stricker, "Deep multi-state object pose estimation for augmented reality assembly," in *IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2019 Adjunct, Beijing, China, October 10-18, 2019*, pp. 222–227, IEEE, 2019.

[5] E. Sucar, K. Wada, and A. J. Davison, "Nodeslam: Neural object descriptors for multi-view shape reconstruction," in *8th International Conference on 3D Vision, 3DV 2020, Virtual Event, Japan, November 25-28, 2020* (V. Struc and F. G. Fernández, eds.), pp. 949–958, IEEE, 2020.

[6] J. Lin, Z. Wei, C. Ding, and K. Jia, "Category-level 6d object pose and size estimation using self-supervised deep prior deformation networks," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IX* (S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), vol. 13669 of *Lecture Notes in Computer Science*, pp. 19–34, Springer, 2022.

[7] J. Liu, Y. Chen, X. Ye, and X. Qi, "Prior-free category-level pose estimation with implicit space transformation," *CoRR*, vol. abs/2303.13479, 2023.

[8] B. X. Nie, P. Wei, and S. Zhu, "Monocular 3d human pose estimation by predicting depth on joints," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 3467–3475, IEEE Computer Society, 2017.

[9] T. Lee, B. Lee, M. Kim, and I. S. Kweon, "Category-level metric scale object shape and pose estimation," *IEEE Robotics Autom. Lett.*, vol. 6, no. 4, pp. 8575–8582, 2021.

[10] Z. Fan, Z. Song, J. Xu, Z. Wang, K. Wu, H. Liu, and J. He, "Object level depth reconstruction for category level 6d object pose estimation from monocular RGB image," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part II* (S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), vol. 13662 of *Lecture Notes in Computer Science*, pp. 220–236, Springer, 2022.

[11] C. Wang, D. Xu, Y. Zhu, R. M. Martin, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 3343–3352, Computer Vision Foundation / IEEE, 2019.

[12] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "PVN3D: A deep point-wise 3d keypoints voting network for 6dof pose estimation," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 11629–11638, Computer Vision Foundation / IEEE, 2020.

[13] D. Cai, J. Heikkilä, and E. Rahtu, "OVE6D: object viewpoint encoding for depth-based 6d object pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 6793–6803, IEEE, 2022.

[14] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2938–2946, IEEE Computer Society, 2015.

[15] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *Robotics: Science and Systems XIV, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA, June 26-30, 2018* (H. Kress-Gazit, S. S. Srinivasa, T. Howard, and N. Atanasov, eds.), 2018.

[16] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4561–4570, Computer Vision Foundation / IEEE, 2019.

[17] S. Zakharov, I. Shugurov, and S. Ilic, "DPOD: 6d pose object detector and refiner," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 1941–1950, IEEE, 2019.

[18] T. Tan and Q. Dong, "Smoc-net: Leveraging camera pose for self-supervised monocular object pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 21307–21316, IEEE, 2023.

[19] M. Tian, M. H. Ang, and 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* (A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds.), vol. 12366 of *Lecture Notes in Computer Science*, pp. 530–546, Springer, 2020.

[20] K. Chen and Q. Dou, "SGPA: structure-guided prior adaptation for category-level 6d object pose estimation," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pp. 2753–2762, IEEE, 2021.

[21] Z. Lin, S. Huang, and Y. F. Wang, "Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 1797–1806, Computer Vision Foundation / IEEE, 2020.

[22] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis, "Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 1581–1590, Computer Vision Foundation / IEEE, 2021.

[23] Y. Di, R. Zhang, Z. Lou, F. Manhardt, X. Ji, N. Navab, and F. Tombari, "Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 6771–6781, IEEE, 2022.

[24] R. Zhang, Y. Di, Z. Lou, F. Manhardt, F. Tombari, and X. Ji, "Rbp-pose: Residual bounding box projection for category-level pose estimation," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I* (S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), vol. 13661 of *Lecture Notes in Computer Science*, pp. 655–672, Springer, 2022.

[25] L. Zheng, C. Wang, Y. Sun, E. Dasgupta, H. Chen, A. Leonardis, W. Zhang, and H. J. Chang, "Hs-pose: Hybrid scope feature extraction for category-level object pose estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 17163–17173, IEEE, 2023.

[26] G. Gkioxari, J. Johnson, and J. Malik, "Mesh R-CNN," in *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 9784–9794, IEEE, 2019.

[27] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *IEEE Winter Conference on Applications of Computer Vision, WACV 2019, Waikoloa Village, HI, USA, January 7-11, 2019*, pp. 1043–1051, IEEE, 2019.

[28] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA* (I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, eds.), pp. 5099–5108, 2017.

[29] Z. Song, J. Lu, T. Zhang, and H. Li, "End-to-end learning for inter-vehicle distance and relative velocity estimation in ADAS with a monocular camera," in *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, pp. 11081–11087, IEEE, 2020.

[30] X. Chen, Z. Dong, J. Song, A. Geiger, and O. Hilliges, "Category level object pose estimation via neural analysis-by-synthesis," in *Computer*

Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVI (A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, eds.), vol. 12371 of *Lecture Notes in Computer Science*, pp. 139–156, Springer, 2020.