2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

GSNet: Model Reconstruction Network for Category-level 6D Object Pose and Size Estimation

Penglei Liu[1,2,3], Qieshi Zhang [1,2,3], Jun Cheng[1,2,3*]

Abstract— **Category-level 6D pose and size estimation is to estimate the rotation, translation and size of the observed instance objects from an arbitrary angle in a cluttered scene. Compared with instance-level 6D pose estimation, there are two main challenges for category-level 6D pose estimation. One is that the algorithm needs to estimate the 6D pose and size of unseen objects, and no 3D models are available. Another is that different instance objects of the same class of objects differ greatly in shape. This paper propose a novel method to estimate the 6D pose and size of unseen objects from an RGB-D image. To handle intra-class shape variation, we propose an autoencoder-decoder that is trained on a set of object models to learn structural feature-invariant and shape-variant features of intra-class objects, and constructs a category-level priori model containing the structure feature and shape feature. To solve the problem of 3D model, this paper proposes a model reconstruction network including 3D graph convolution and spherical convolution (GSNet), which can reconstruct the 3D model of the observed instance object from the input RGB-D image and the priori model, and establish a dense correspondence between the 3D model and the observed instance object. Finally, random sample consensus (RANSAC) algorithm and Umeyama algorithm are used to estimate the 6D pose and size of the object. Extensive experiments on benchmark datasets show that the proposed method achieves state-of-the-art performance in category-level 6D object pose estimation. In order to prove that our method can be applied to the grasping and operation tasks of robots in industry and life, we deploy our method to a physical UR5 robot to perform grasping tasks on unseen but category known instances, and the results validate the efficacy of our proposed method.**

I. INTRODUCTION

Accurate 6D object pose estimation plays an important role in robotic grasping tasks [1, 2, 3, 4, 5, 6, 7, 8, 10, 28, 29, 30, 31]. In recent years, instance-level 6D object pose estimation has developed rapidly and achieved good performance. Unfortunately, these methods [9, 11, 12, 23] cannot be used without 3D models and cannot be directly generalized to unseen instance objects, which greatly limits their usefulness in practical applications. Consequently, the category, 6D pose and size of the objects have to be concurrently estimated, and this task is also called category-level 6D pose and size estimation [13, 14, 32] . At present, the key challenges of category-level 6D object pose and size estimation task mainly include two aspects. One is that there are no corresponding 3D models to use when estimating the 6D pose and size of unseen objects. Another is that there are huge color and shape differences between different objects of the same category.

To solve the above two problems, some recent popular approaches [13, 14] map different objects in the same category into a uniform model to solve this problem. For example, Wang *et al.* [13] proposed a data-driven solution for the category-level 6D pose estimation problem, they introduced a normalized object coordinate space (NOCS) to represent different object instances within a category in a unified manner. They train a deep neural network to infer correspondences from the object pixel to the point in the NOCS, and at the same time obtain the class label and instance mask of each object, and then use these predictions together with the depth map to estimate the 6D pose and size of the object through point matching. However, the lack of explicit representation of shape and structural changes limits their performance. Some other methods [15] [16] solve the above two problems by reconstructing the complete 3D model or reconstructing part of the 3D model. For example, Tian *et al.* [15] first proposed a method for 3D model reconstruction to address the problem of instance objects without 3D models, and trained a deformable domain network to fine-tune model reconstruction for different instances in the class. However, this method only considers the uniform shape when learning the shape, and does not consider the structural information of the objects in the category, so there are still limitations in 3D model reconstruction.

In this work, our research idea is to learn the structural features and shape features of objects within the class, and then reconstruct the complete 3D model of the object in NOCS space to solve the problem that unseen object without model, and finally establish the dense correspondence between the observed object and the reconstructed 3D model to estimate the 6D pose and size of the object. Existing research shows that 3D graph convolutional networks (GCN) [19] can effectively extract the key structure information of objects in the image, which is helpful to extract the feature invariant structure from category-level objects with similar structure but changed shape. At the same time, point cloud networks are also widely used in geometric feature extraction of objects. Therefore, in order to address the

This work was supported in part by the National Natural Science Foundation of China (U21A20487, U1913202), in part by the Shenzhen Technology Project (JCYJ20220818101206014, JCYJ20180507182610734), in part by the Shenzhen Engineering Laboratory for 3D Content Generating Technologies (No. [2017]476), in part by the National Natural Science Foundation of Guangdong Province (No. 2022A1515140119), and in part by the CAS Key Technology Talent Program.(*Corresponding author: Jun Cheng.*)

[1]Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China.

[2]Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, China.

[3]The Chinese University of Hong Kong, Hong Kong, China. (pl.liu, qs.zhang, jun.cheng)@siat.ac.cn

shape and structure variation of objects within category-level, we design a novel autoencoder-decoder based on 3D graph convolutional network and point cloud network (PCN). The autoencoder is trained on a collection of objects models from various categories to learn the structural and geometric features of different category-level objects. These features are then fed to the autodecoder to generate category-level shape and structure priors.

In order to solve the problem that there is no 3D model available for the instance object in 6D pose estimation task, we propose a model reconstruction network including 3D graph convolutional network and spherical convolutional network (SCN) [20] to reconstruct 3D models of observed instance objects, which is called GSNet for short. GSNet considers the shape, structure, and rotation changes of the object when reconstructing the 3D model of the object, so that the reconstructed model contains the shape, structure and rotation invariant features of the object. At the same time, GSNet also estimates the correspondence between the observed objects and the reconstructed 3D model, and then converts the corresponding points on the 3D model to points in NOCS space. Finally, the Umeyama algorithm [21] and the RANSAC algorithm [22] are used to recover the 6D pose and size of objects from the point cloud in NOCS and the point cloud of observed objects.

Extensive experiments conducted on the category-level dataset [13] demonstrate that our approach outperforms the state-of-the-art (SOTA) methods. In order to verify the effect of our method in the real robot scene, the proposed method is deployed on a real robotics platform for evaluation. The experimental results show that our method can effectively assist the robot to grasp and operate unseen objects. Our technical contributions are summarized as follows:

- We propose a novel depth network for category-level 6D object pose and size estimation. The network combines the 3D graph convolutional network and spherical convolutional network (GSNet), which can reconstruct the 3D model of the object and solve the problem that there is no 3D model available for unseen objects.
- We propose a novel autoencoder-decoder based on 3D graph convolution network, which is used to learn the shape and structural features of Intra-class objects and generate a priori of category-level 3D model containing these features.
- Our approach outperforms state-of-the-art methods on category-level benchmark datasets. At the same time, we deploy and verify the proposed algorithm on a real robot platform.

II. Related work

A. Instance-level 6D Object Pose Estimation

Existing 6D pose estimation methods for instance-level objects are mainly divided into the following categories. The first category of methods is template matching, which estimates the pose of an object by matching image features [7]. While these methods perform well at inferring textured objects, they perform poorly when inferring weakly textured or untextured objects. The second category of methods is based on deep networks, estimate the 6D pose of objects directly from RGB images [4,9] or RGB-D images [5,6,10,11,12]. However, 3D models of objects are required for both network training and inference stages. The third category of method is to perform 6D pose estimation recovery by establishing 2D-3D correspondence between 2D images and 3D models, or 3D-3D correspondences between point clouds and 3D models. However, 3D models of objects are also required in the network training and inference stages. A common problem in instance-level 6D pose estimation is that an accurate 3D object model of the object is required both in the training process and in the reasoning process. Our method can infer the 6D pose and size of unseen objects without 3D models.

B. Category-level 6D Object Pose and Size Estimation

Recent work on category-level 6D pose estimation tasks [13,14,15,16] has greatly alleviated the limitations of previous instance-level 6D pose estimation tasks. Wang *et al.* [13] proposed a NOCS representation to represent objects of the same category with the same 3D model. The method first predicts the NOCS image of the object, and then aligns it with the observed depth of the object to estimate the 6D pose and size of the object. In order to deal with the shape changes within the class, Tian *et al.* [15] improved the prediction of the standard object model by deforming the classification shape a prior. Chen *et al.* [16] extracts shape based features from the point cloud of the target object for pose and size recovery. Tian *et al.* [15] did not consider the structural features of category-level objects when processing the intra-class features of objects, while Chen *et al.* [16] did not reconstruct the complete 3D model in the process of model reconstruction. Our method considers both shape and structural changes when processing category-level object features, and reconstructs a complete 3D model of the object.

III. Approach

A. Overview

As shown in Fig. 1, our method consists of three parts. In the first part, a Mask R-CNN [24] is used to perform instance segmentation on color images. Next, we convert the mask depth map into a point cloud with the intrinsic parameters of each instance of the camera, and crop the image patch according to the boundary box of the mask. In the second part, the GSNet is used to perform 3D model reconstruction on point clouds, image patches and corresponding structure priors (Section III.B). At the same time, the GSNet outputs a set of correspondences that associate each point in the point cloud of the instance object with the point of the reconstructed model, and the reconstructed model can be masked to NOCS coordinates by this set of correspondences (Section III.C). In the third part, the 6D pose and size of the object can be estimated by registering the NOCS coordinates and the point cloud obtained from the observed depth map (Section III.D).

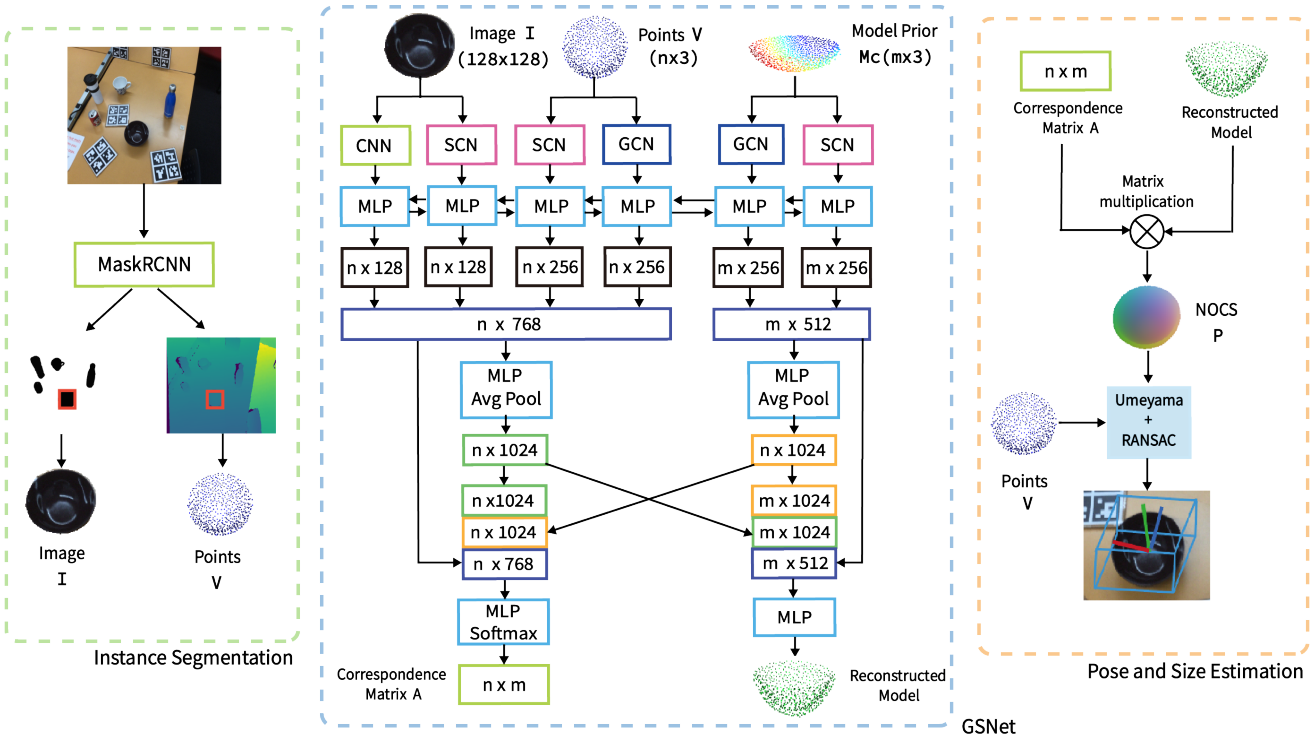


Fig. 1. The network architecture consists of three parts, instance segmentation, GSNet and pose estimation. In order to effectively extract the rotational and structural information of objects, we apply spherical convolution and 3D graph convolution to our network. At the same time, in order to fuse the features of the 3D prior model with the features of the observed instance objects, we share the global features. GSNet performs full 3D model reconstruction of observed instance objects and outputs a set of correspondences A, which are used to map the 3D models to NOCS space.

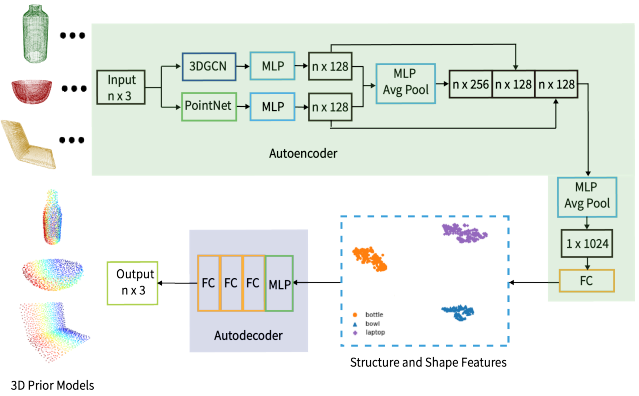


Fig. 2. Autoencoder-decoder architecture. The autoencoder extracts structural and shape features from the input point cloud model, and then we visualize the extracted structural and geometric features in a low-dimensional space. Finally, the decoder is used to reconstruct the prior 3D model containing category-level structural and shape features.

B. Autoencoder-decoder for Category-level Shape and Structure Priors

In this section, we introduce autoencoder-decoder, which is used to extract and generate shape and structure priors for category-level objects. Although there are differences in color, texture and shape among different objects of the same category, objects of the same category have similar structural information. For example, mugs usually consists of a container with a cylindrical structure and a handle with a ring structure; cameras usually consists of a body with a cuboid structure and a lens with a cylindrical structure. These categorical features provide important prior knowledge for model reconstruction of new instance objects. Existing research shows that 3D graph convolutional networks (3DGCN) can effectively extract key structural information of objects in the image, which is helpful to extract feature invariant structure from category-level objects with similar structures but different shapes. Therefore, we design an autoencoder-decoder based on 3DGCN and PCN to extract shape features and structural features of category-level objects, providing prior 3D model knowledge for unseen objects of the same category.

The autoencoder-decoder is trained with all available object models, then the autoencoder extracts the shape features and structural features of each object category, and finally these features are fed into the autodecoder to generate a priori 3D model of the structure and shape of each category. Specifically, given a set of 3D point cloud models aligned in NOCS space $M = \{M_c^i | 1, 2, 3 \cdots, N; c = 1, 2, 3 \cdots, C\}$, where M_c^i is from 3D point cloud model for instance i of category c. The autoencoder Φ takes the point cloud and outputs a set of vectors containing shape and structural features, i.e. z_c^i, and the decoder Ψ takes this set of feature vectors and outputs a reconstructed 3D model $\hat{M}_c^i$ containing category-level shape and structural features:

$$\hat{M}_c^i = \Psi\left(\Phi(M_c^i)\right) = \Psi(z_c^i). \quad (1)$$

Due to the inter-class differences between different category of objects, during the training process of the autoencoder, the structural features of the objects will form clusters

in a low-dimensional space. As shown in Fig. 2, objects of different categories form different clusters. Then these features are fed into the autodecoder, and prior 3D models containing category-level shape and structure features are output.

C. GSNet for 3D Model Reconstruction

After obtaining the segmented object and the prior 3D model, we denote the segmented object instance as (V, I), where $V \in R^{N_v \times 3}$ represents the objects's point clouds and $I \in R^{H \times W \times 3}$ represents the object's image patchs. N_v represents the number of point clouds. The 3D prior model containing shape and structural features is denoted by $M_c \in R^{N_c \times 3}$, where N_c is the number of points in M_c. We input V, I and M_c into the GSNet, and output the 3D reconstruction model M of the instance object and the corresponding matrix $A \in R^{N_v \times N_c}$. Through the correspondence matrix A, we can establish the correspondence between the point cloud of the observed instance object and the corresponding point in the 3D model, and convert the corresponding point in the 3D model to the point in the NOCS space. As shown in Fig. 1, GSNet consists of four parts: (1) feature extraction from object instances; (2) feature extraction from prior 3D models; (3) regression the correspondence matrix A; (4) 3D model reconstruction of the observed object.

In the process of model reconstruction, CNN and SCN are used to extract appearance color features and rotation-invariant features from input RGB images, respectively. At the same time, SCN and GCN are used to extract spatial rotation-invariant and structural features from input point cloud and prior model, respectively. The extraction of RGB image features and the extraction of point cloud features are separated, and then the extracted features are input to the multi-layer perceptron (MLP). Considering that point cloud and color are two different modalities, we follow the feature fusion scheme proposed by [25] to fuse RGB features and point cloud features. In the 3D model reconstruction task, we believe that although the observed instance object (V, I) is partial, it provides the specific structure and shape details of the instance object. At the same time, the prior 3D model M_c provides a priori knowledge of the shape and structure of this category. In order to reconstruct the 3D model of the observed object more accurately, we share and integrate the global features of the observed object and the global features of the prior model M_c, which can ensure that the reconstructed 3D model can not only contain the specific shape and structure information of the observed object, but also contain the structural prior information of this category. Finally, our GSNet network outputs the corresponding matrix A and the reconstructed 3D model M. The coordinate P corresponding to NOCS is:

$$P = A \times M \in R^{N_v \times 3}. \tag{2}$$

D. Category-level 6D Pose and Size Estimation

The task of this paper is to estimate the 6D pose and size of unseen objects. After obtaining the point cloud coordinates V of the instance object and its corresponding NOCS coordinates P, the Umeyama algorithm [21] is used to estimate the 6D pose and size of the instance object, and the RANSAC algorithm [22] is used for robust estimation.

E. Loss Functions

In this section, we define the loss functions used to train our network.

Reconstruction Loss. Assume that ground-truth model M_{gt} is available. In the process of training the Autoencoder-decoder and GSNet, the reconstruction error of the 3D model is measured by Chamfer distance:

$$L_{cd}(M_c^i, M_{gt}) = \sum_{x \in M_c^i} \min_{y \in M_{gt}} \|x - y\|_2^2 + \sum_{y \in M_{gt}} \min_{x \in M_c^i} \|x - y\|_2^2. \tag{3}$$

Correspondence Loss. A is supervised indirectly through the NOCS coordinate P (which is a result of applying the correspondence matrix A on the reconstructed model M) since the ground-truth NOCS coordinates P_{gt} can be obtained easily from the object model and its 6D pose through image rendering. The smooth $L1$ loss function is used:

$$Lcorr(P, P_{gt}) = \frac{1}{N_v} \sum_{i=1,2,\cdots n} \begin{cases} 5(x_i - y_i)^2, & if\ |x_i - y_i| \leqslant 0.1 \\ |x_i - y_i| - 0.05, & otherwise \end{cases} \tag{4}$$

where $x = (x_1, x_2, \cdots, x_n) \in P$, and $y = (y_1, y_2, \cdots, y_n) \in P_{gt}$.

As for the problem of symmetric objects, we follow the loss function proposed in [15] to supervise and train the network.

In summary, the overall goal is the weighted sum of all losses:

$$L = \lambda_1 L_{cd} + \lambda_2 L_{corr}, \tag{5}$$

and for the hyperparameters of the total loss, we set $\lambda_1 = 5.0$ and $\lambda_2 = 1.0$.

TABLE I

RECONSTRUCTION TYPE COMPARISON, AND THE COMPARISON IS ON THE NOCS-REAL DATASET WITH THE CHAMFER DISTANCE METRIC ($\times 10^{-3}$)).

Method	SPD[15]	CASS[14]	FSNet[16]	Ours
Bottle	3.44	**0.75**	1.2	0.77
Bowl	1.21	**0.38**	0.39	0.41
Camera	8.99	0.77	**0.44**	0.45
Can	1.56	**0.42**	0.62	0.46
Laptop	2.91	3.73	2.23	**1.94**
Mug	1.02	0.32	**0.29**	0.31
Average	3.17	1.06	0.86	**0.72**

TABLE II

QUANTITATIVE COMPARISONS OF DIFFERENT METHODS ON CAMERA25 AND REAL275. EVALUATIONS ARE BASED ON BOTH THE METRICS PROPOSED IN [13] (LEFT) AND THE METRICS (RIGHT) PROPOSED IN [17].

Dataset	Method	mAP								
		IoU_{50}	IoU_{75}	5°,2cm	5°,5cm	10°,2cm	10°,5cm	IoU_{75},5°,5%	IoU_{75} ,10°,5%	IoU_{75} ,5°,10%
CAMERA25	NOCS[13]	83.9	69.5	32.2	40.9	48.2	64.6	22.6	29.5	31.5
	SPD[15]	93.2	83.1	54.3	59.0	73.3	81.5	47.5	61.5	52.2
	DualPose[17]	92.4	86.4	64.7	70.7	77.2	**84.7**	56.2	65.1	65.1
	SARNet[18]	86.8	79.0	66.7	70.9	75.3	80.3	-	-	-
	Ours	**93.5**	**87.6**	**69.1**	**73.7**	**79.8**	83.5	**60.4**	**71.2**	**71.9**
REAL275	NOCS[13]	78.0	30.1	7.2	10.0	13.8	25.2	2.4	3.5	7.1
	CASS[14]	77.7	**-**	-	23.5	-	58.0	-	-	-
	SPD[15]	77.3	53.2	19.3	21.4	43.2	54.1	8.6	17.2	15.0
	FSNet[16]	**92.2**	**63.5**	-	28.2	-	60.8	-	-	-
	DualPose[17]	79.8	62.2	29.3	35.9	50.0	66.8	11.2	17.2	24.8
	SARNet[18]	79.3	62,4	31,6	42,3	50.3	68.3	-	-	-
	Ours	85.2	63.3	**34.7**	**45.1**	**52.2**	**67.5**	**15.6**	**22.3**	**28.7**

IV. EXPERIMENTS

Our method is compared with the state-of-the-art (SOTA) methods on two challenging category-level 6D object pose estimation datasets. At the same time, in order to evaluate the robustness and effectiveness of our method in robotic grasping tasks, we deploy our method on a real robotic platform for grasping tasks.

A. Datasets

Category-level Dateset. NOCS dataset [13] contains two parts, the synthetic dataset CAMERA25 and the real dataset REAL275. For CAMERA25, which is a synthetic dataset generated by context-aware mixed reality methods for 6 object categories, there are a total of 300K synthetic images, where 25K are set aside for evaluation. REAL275 is a more challenging real-world dataset with clutter, occlusion, and various lighting conditions, with a total of 7.05K real images, where 2.75K are set aside for evaluation.

B. Evaluation Metrics

The evaluation is divided into two parts: 3D object detection and 6D pose estimation. To evaluate 3D detection and object dimension estimation, we compute the average precision of 3D intersection over union (IoU) with thresholds of 50% and 75% for 3D object detection. For 6D pose estimation, we compute the average precision of object instances for which the error is less than m cm for translation and $n°$ for rotation. Here we choose threshold values of 5°, 10°, 2cm and 5cm ,respectively. Finally, to evaluate pose and size simultaneously, we follow the method of [17] to evaluate the proposed method.

C. Evaluation of Reconstruction

3D point cloud model reconstruction has a close relationship with 6D pose and size estimation performance. We compute the Chamfer Distance of the reconstructed 3D point cloud model with the ground truth 3D point cloud model and compared it with other reconstruction types used by other methods. It can be seen from Table I that the average error of 3D model reconstruction using our method is 0.72, which is 72.3%, 32.1% and 16.3% lower than SPD [15], CASS [14] and FSNet [16], respectively. It shows that 3D models of objects can be more accurately reconstructed using our method.

TABLE III

ABLATION STUDIES ON NOCS-REAL DATASET. WE USE TWO DIFFERENT METRICS TO MEASURE PERFORMANCE. 3DGCN MEANS THE 3D GRAPH CONVOLUTION, SPH MEANS SPHERICAL CONVOLUTION AND 3DP MEANS 3D PRIOR MODEL.

Method	GCN	SCN	3DP	5°,5cm	IoU_{75},5°,10%
Scheme_1	×	×	×	30.2	17.5
Scheme_2	×	✓	×	33.6	18.8
Scheme_3	✓	×	×	34.5	20.2
Scheme_4	×	×	✓	38.9	22.8
Scheme_5	×	✓	✓	42.7	25.3
Scheme_6	✓	×	✓	43.3	26.5
Ours	✓	✓	✓	**45.1**	**28.7**

D. Comparison with the SOTA methods

Compare our proposed method with existing SOTA methods on the CAMERA25 and REAL275 [13] datasets. The quantitative results in Table II show the superiority of our proposed method on both datasets, especially for high accuracy metrics. Specifically, in terms of 6D pose measures metrics 5° 2cm, 5° 5cm and 10° 2cm, our method outperforms SOTA methods on the dataset CAMERA25 and the dataset REAL275. In addition, we evaluate our algorithm with the evaluation scheme proposed by DualPose [17]. In terms of 6D pose measures metrics IoU_{75} 5°5%, IoU_{75} 10°5% and IoU_{75} 5°10%, our method outperforms SOTA methods on the dataset CAMERA25 and the dataset REAL125. Experiments show that our method can effectively estimate the pose and size of category-level objects.

E. Ablation Study

In this section, in order to verify the impact of 3D priori model, 3D graph convolution network (GCN) and spherical convolution network (SCN) on the performance of our method. We conduct ablation experiments with or without 3D prior model, GCN and SCN. The ablation experiments are performed on the NOCS-REAL275 dataset and the results are shown in Table III. By comparing Scheme_1 and Scheme_4, we find that using 3D prior models can effectively improve the performance of 6D pose estimation. Comparing Scheme_1 and Scheme_2, Scheme_1 and Scheme_3 respectively, it can be found that GCN and SCN can improve the performance of the network. The best performance is achieved when using 3D priors, GCN

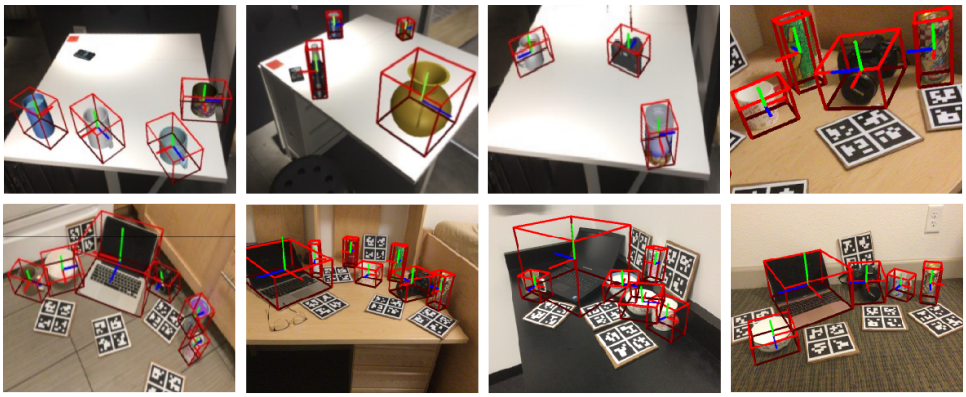
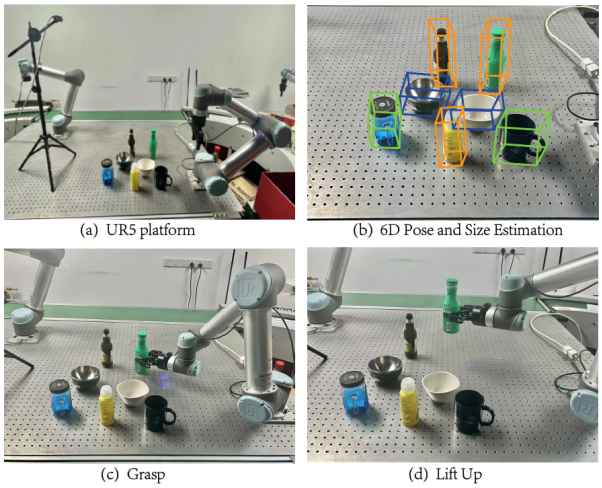


Fig. 3. Visualization results of our method on CAMERA25[13] and REAL275[13] datasets.



(a) UR5 platform (b) 6D Pose and Size Estimation

(c) Grasp (d) Lift Up

Fig. 4. The real experimental platform; (a) the UR5 robot platform; (b) the estimated results given by our algorithm in the real environment; (c) the UR5 robot grasps the object by the algorithm; (d) the UR5 robot lifts the object.

and SCN simultaneously. Overall, the proposed method can significantly improve the performance of the network.

F. Robotic Grasping Experiment

In order to verify the effectiveness of our proposed method, we build a robotic platform. As shown in Fig. 4(a), the robotic platform consists of a Universal Robot 5 (UR5) robotic arm, a RealSense 415 camera, a Robotiq 85 gripper, and a computer configured with an Intel i5 3.7 GHz CPU and a GTX 1080 Ti GPU.

It is well known that robotic grasping is a systematic task, so there are many factors affecting robot grasping, such as suitable motion planning, robot control, and 6D pose estimation algorithms. We use Moveit in the robotic operating system (ROS) to control the motion of the UR5 robotic arm. Based on inverse kinematics, the robotic arm can grasp and manipulate objects according to the 6D pose predicted by the algorithm. Regarding the path planning of grasping, after obtaining the 6D pose of the object, the robotic arm first reaches 0.15 meters above the object, and then approaches the object until it reaches the final grasping pose, grasping and lifting the object. Fig. 4 (b) shows the 6D pose and size estimation results of our algorithm. Three key processes are involved in grasping an object: estimating the 6D pose and size of the object, then grasping or holding

TABLE IV

THE GRASPING SUCCESS RATE (%) OF THE UR5 IN THE REAL ENVIRONMENT.

Method	NOCS[13]	SPD[15]	DualPose[17]	Ours
Bottle	60.2	69.4	83.4	**88.6**
Bowl	53.6	63.2	77.6	**84.6**
Mug	58.2	66.8	79.4	**85.8**
Average	57.3	66.5	80.1	**86.3**
Variance	7.6	6.5	5.9	**2.8**
Standard Deviation	2.8	2.5	2.4	**1.7**

the object, and finally lifting the object. Seven objects are selected as the grasp targets, including 2 cups, 3 bottles and 2 bowls. Since our proposed 6D pose estimation algorithm is used to assist the robotic arm in grasping objects, we evaluate the algorithm by considering the success rate of the robotic arm in grasping objects. We deployed different methods on the robotic arm platform for grasping experiments, and each group of experiments performed 500 grasping tasks.The experimental results are shown in Table IV. Our method has an average success rate of 86.3% in grasping different objects with a variance of 2.8 and a standard deviation of 1.7. Compared with other STOA methods, such as NOCS [13], SPD[15] and DualPose [17], our method has higher average precision, smaller variance and standard deviation, and more stable performance.

V. CONCLUSION

In this paper, we propose a approach for category-level 6D pose estimation. First, we designed an autoencoder-decode to learn and extract the structural features and shape features of category-level objects, and then generated a 3D model containing structural priors, which solved the problem of shape and structure differences between different classes of objects. Then, we propose the GSNet, which can perform 3D model reconstruction of instance objects, so that it can solve the problem that no 3D models are available for category-level objects. Finally, 6D pose estimation and size estimation are performed by Umeyama and RANSAC algorithm. We evaluate the proposed method on benchmark datasets, and experimental results show that our method achieves STOA performance in category-level pose estimation. Furthermore, we deploy the proposed method on the UR5 robot platform for grasping experiments, which verify the practicability of our method in practical robotic applications.

REFERENCES

[1] Mercier Jean-Philippe, Mitash Chaitanya, Giguere Philippe and Boularias Abdeslam, "Learning object localization and 6D pose estimation from simulation and weakly labeled real images," IEEE International Conference on Robotics and Automation (ICRA), 2019, pp.3500-3506.

[2] Bo Cheng, Wanyin Wu, Dapeng Tao, Shibo Mei, Ting Mao and Jun Cheng, "Random cropping ensemble neural network for image classification in a robotic arm grasping system," IEEE Transactions on Instrumentation and Measurement (TIM), 2020, vol. 69, no. 9, pp.6795-6806.

[3] Tommaso Cavallari, Stuart Golodetz, Nicholas Lord, Julien Valentin, Victor Prisacariu, Luigi Di Stefano and Philip HS Torr, "Real-time RGB-D camera pose estimation in novel scenes using a relocalisation cascade," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 2019, vol. 42, no. 10, pp.2465-2477.

[4] Penglei Liu, Qieshi Zhang, Jin Zhang, Fei Wang, Jun Cheng, "MFPN-6D real-time one-stage pose estimation of objects on RGB images," IEEE International Conference on Robotics and Automation (ICRA), 2021, pp.12939-12945.

[5] Porzi Lorenzo, Penate-Sanchez Adrian, Ricci Elisa and Moreno-Noguer Francesc, "Depth-aware convolutional neural networks for accurate 3D pose estimation in RGB-D images," IEEE International Conference on Intelligent Robots and Systems (IROS), 2017, pp.5777-5783.

[6] Umar Asif, Mohammed Bennamoun and Ferdous A. Sohel,"RGB-D object recognition and grasp detection using hierarchical cascaded forests," IEEE Transactions on Robotics, 2017, vol. 33, pp. 547–564.

[7] Park Kiru, Patten Timothy, Prankl Johann and Vincze Markus, "Multi-task template matching for object detection, segmentation and pose estimation using depth images," IEEE International Conference on Robotics and Automation (ICRA), 2019, pp.7207-7213.

[8] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl and Dieter Fox, "PoseRBPF: A rao–blackwellized particle filter for 6D object pose tracking," IEEE Transactions on Robotics, 2021, vol. 37, pp. 1328–1342.

[9] Sergey Zakharov, Ivan Shugurov and Slobodan Ilic, "DPOD: 6D pose object detector and refiner," IEEE International Conference on Computer Vision (ICCV), 2019, pp.1941-1950.

[10] Zhou Teng and Jing Xiao, "Surface-based detection and 6DoF pose estimation of 3D objects in cluttered scenes," IEEE Transactions on Robotics, 2016, vol. 32, pp. 1347–1361.

[11] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu and Fei-Fei Li and Silvio Savarese, "Densefusion: 6D object pose estimation by iterative dense fusion," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp.3343-3352.

[12] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan and JianSun, "PVN3D: A deep point-wise 3D keypoints voting network for 6DoF pose estimation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11632-11641.

[13] 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," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2642–2651.

[14] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu, "Learning canonical shape space for category-level 6D object pose and size estimation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11973–11982.

[15] Meng Tian, Marcelo H Ang Jr and Gim Hee Lee, "Shape prior deformation for categorical 6D object pose and size estimation," arXiv preprint arXiv:2007.08454, 2020.

[16] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen and Ales Leonardis, "FS-net: Fast shape-based network for category-level 6D object pose estimation with decoupled rotation mechanism," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 1581–1590.

[17] 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," IEEE International Conference on Computer Vision (ICCV), 2021, pp. 3560–35691.

[18] Haitao Lin, Zichang Liu, et al, "SAR-Net: Shape alignment and recovery network for category-level 6D object pose and size estimation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 6707–6717.

[19] Zhihao Lin, Shengyu Huang, et al, "Convolution in the cloud: Learning deformable kernels in 3D graph convolution networks for point cloud analysis," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1800–1809.

[20] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis, "Learning so (3) equivariant representations with spherical cnns," In proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 52–68.

[21] Umeyama Shinji, "Least-squares estimation of transformation parameters between two point patterns," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), 1991, vol. 13, pp. 376–380.

[22] Fischler Martin A and Bolles Robert C, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, 1981, vol. 24, pp. 381–395.

[23] Jun Cheng, Penglei Liu, Qieshi Zhang, Hui Ma, Fei Wang and Jin Zhang,"Real-Time and Efficient 6D pose estimation from a single RGB image," IEEE Transactions on Instrumentation and Measurement (TIM), 2021, vol. 70.

[24] Kaiming He, Georgia Gkioxari, Piotr Dollar, and Ross Girshick, "Mask R-CNN," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2961–2969.

[25] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen and Jian Sun, "FFB6D: A full flow bidirectional fusion network for 6D pose estimation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1800–1809.

[26] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige and Nassir Navab, "Model based training, detection and pose estimation of texture-less 3D objects in heavily cluttered scenes," In proceedings of Asian Conference on Computer Vision (ACCV), 2012, pp 548–562.

[27] Xiang Yu, Schmidt Tanner, Narayanan Venkatraman and Fox Dieter, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[28] Muyuan Lin, Varun Murali,and Sertac Karaman1, "6D object pose estimation with pairwise compatible geometric features," IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 10966-10973.

[29] Kilian Kleeberger, Markus Volk, Richard Bormann1, and Marco F. Huber, "Investigations on output parameterizations of neural networks for single shot 6D object pose estimation," IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 13916-13922.

[30] Ge Gao, Mikko Lauri, Xiaolin Hu, Jianwei Zhang and Simone Frintrop, "CloudAAE: Learning 6D object pose regression with online data synthesis on point clouds," IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 11081-11087.

[31] Mengchao Zhang and Kris Hauser, "Non-Penetration iterative closest points for single-view multi-object 6D pose estimation," IEEE International Conference on Robotics and Automation (ICRA), 2022, pp. 1520-1526.

[32] Muhammad Zubair Irshad, Thomas Kollar, Michael Laskey, Kevin Stone and Zsolt Kira, "CenterSnap: Single-Shot multi-object 3D shape reconstruction and categorical 6D pose and size estimation," IEEE International Conference on Robotics and Automation (ICRA), 2022, pp. 10632-10640.