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

KGNet: Knowledge-Guided Networks for Category-Level 6D Object Pose and Size Estimation

Qiwei Meng[1,2], Jason Gu[3], Shiqiang Zhu[1,2],
Jianfeng Liao[1,2], Tianlei Jin[1,2], Fangtai Guo[1,2], Wen Wang[1,2] and Wei Song[1,2]

***Abstract*— Despite the giant leap made in object 6D pose estimation and robotic grasping under structured scenarios, most approaches depend heavily on the exact CAD models of target objects beforehand, thereby limiting their wide applications. To address this, we propose a novel knowledge-guided network - KGNet to estimate the pose and size of category-level unseen objects. This network includes three primary innovations: knowledge-guided categorical model generation, pointwise deformation probability matrix and synergetic RGBD feature fusion, with the former two leveraging categorical object knowledge for unseen object reconstruction and the latter one facilitating pose-sensitive feature extraction. Extensive experiments on CAMERA25 and REAL275 verify their effectiveness, and KGNet achieves the SOTA performance on these two acknowledged benchmarks. Additionally, a real-world robotic grasping experiment is conducted, and its results further qualitatively prove the practicability and robustness of KGNet.**

I. INTRODUCTION

Category-level object 6D pose and size estimation aims to calculate the size, 3D translation and 3D rotation of unseen object instances [1], [2]. This line of work is increasingly focused on and studied recently for its importance to various real-world applications, such as augmented reality (AR) [3], virtual reality (VR) [4], 3D scene reconstruction [5] and robotic manipulations [6], [7], [8]. Though instance-level 6D pose estimation methods have been explored in depth and achieved reliable performances [8], [9], [10], even under severe occlusion, their applicability in practical terms is still questioned [2], [11]. This is mainly because under unstructured household and office scenarios, creating high-quality CAD models for every potential target in advance could be extremely time-consuming and unpractical [2], [12], limiting the uses of instance-level methods. Different from them, category-level approaches can estimate the 6D pose and size for previously unseen objects from known classes, and they do not require the exact CAD models of target objects beforehand, which greatly enhances their generalization ability and practicability [2], [12], [13].

Despite the advantages of category-level approaches, they are acknowledged to be much more challenging than traditional instance-level methods [2], [12]. Apart from common difficulties such as lighting variations and cluttered backgrounds [9], [10], [14], [15], the lack of prior CAD models will cause the failure of voting-based [8] and correspondence-based [16] methods, which have achieved great performances in instance-level object pose estimation. Additionally, due to intra-class variations and distribution shifts, end-to-end deep neural networks (DNNs) approaches are considered to be less reliable and robust [17], [18]. Accordingly, [2] introduces normalized object coordinate space (NOCS) to represent all possible objects within a category, and DNNs are trained to map observed pixels to the NOCS. With canonical representation, the pose and size of unseen instances can be calculated.

The proposal of NOCS significantly promotes the development of category-level pose estimation, and lots of relevant works are coming forth [11], [12], [13], [19]. Though these data-driven approaches achieve remarkable improvements on benchmarks, their performances are still far from satisfactory due to the insufficient leverage of prior knowledge of category-level objects. In addition, previous RGBD-based methods commonly apply separate DNN branches to extract RGB and depth features, which neglect their complementarity, thereby lowering the representativity and sensitivity of extracted features.

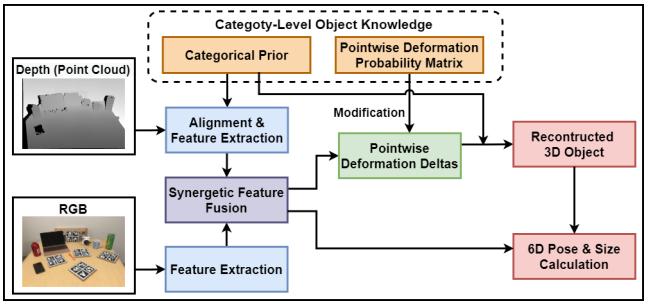


Fig. 1: Semantic Illustration of KGNet.

In this paper, we present KGNet (shown in Fig.1), a novel knowledge-guided network for category-level object pose and size estimation from a single RGBD image. Its primary innovations include the representation and application of category-level object knowledge. To be specific, we propose the category-level empirical CAD model and pointwise deformation probability matrix as object knowledge, which will be incorporated into model training and inference. Such prior geometric knowledge can enhance the effects of mapping

This research is supported by Key Research Project of Zhejiang Lab (No. G2021NB0AL03).

[1]Research Center for Intelligent Robotics, Research Institute of Interdisciplinary Innovation, Zhejiang Lab, Hangzhou 311100, China, {mengqw, zhusq, jfliao, jtl, guofangtai, wangwen, weisong}@zhejianglab.com.

[2]Zhejiang Engineering Research Center for Intelligent Robotics, Hangzhou 311100, China.

[3]Department of Electrical and Computer Engineering, Dalhousie University, Halifax, NS Canada B3H 4R2, jason.gu@dal.ca.

Zhu, Gu and Song are corresponding authors.

real-world objects to the NOCS by bringing in categorical object characteristics. Another innovation of our work is the layerwise synergetic RGB and depth feature fusion, and this module is conducive to pose-sensitive feature extraction.

To summarize, our major contributions are as follows:

1) We introduce the concept of category-level object geometric knowledge, including empirical CAD model and deformation probability matrix, which can improve the model performance by a considerable margin.

2) We propose an embeddable synergetic RGB and depth feature fusion module. Together with the categorical object knowledge, our KGNet achieves the SOTA performance on acknowledged CAMERA25 and REAL275 [2] benchmarks.

3) We conduct real-world robotic grasping experiments, and qualitative results well prove the effectiveness and practicability of our KGNet.

II. RELATED WORK

A. Instance-Level Pose Estimation

For instance-level object pose estimation, the exact CAD models of target objects are available in advance. It is a well-established research field and its technical routes are fairly clear. Depending on the usage of known object models, this line of work can be roughly divided into four types: end-to-end, template-based, correspondence-based and voting-based methods. Posecnn [17] is a typical example of the end-to-end approach, it directly applies multi-branch networks for object 6D pose calculation, and the CAD model is used in iterative closest point (ICP) algorithm for post-refinement. Template-based methods [20], [21] use the CAD model to generate multiple templates for predefined landmark object positions, and pose estimation can be realized through a classification-like network subsequently. Correspondence-based methods [22], [23] apply handcrafted feature descriptors or DNNs to locate paired points between the scene image and object model, and accordingly calculate pose using Perspective-n-Point (PnP) algorithms. Voting-based techniques [8], [9], [10] leverage observed pixels to vote object keypoints in the input image, and then 6D pose can be calculated based on voted keypoints - CAD model keypoints correspondence. Compared with correspondence-based methods, they reduce the number of paired points in PnP equations, thus decreasing computational time. It is manifest that all instance-level approaches depend strongly on the exact object models, so they are less effective when directly transferring to category-level problems.

B. Category-Level Pose Estimation

Category-level object pose estimation is a developing frontier research field, and only a few pioneering works are focusing on it. The exact CAD model of target object is unavailable beforehand in this task, so it will estimate its size together with 6D pose in general. Before the proposal of NOCS [2], handcrafted feature descriptors [1], [24] and end-to-end DNNs [25] are two mainstream methods for this problem, but they tend to be less robust under cluttered scenes. To improve it, Wang [2] maps real-world categorical objects to NOCS for pose estimation, which is an important landmark for reconstructing the target object while including category-level features. Following this, CASS [19] optimizes the feature extraction and mapping for better performance. [11], [13] emphasize the semantic and geometric similarity of objects within the same category, so they apply mean latent embedding to construct categorical shape priors and then incorporate them into NOCS mapping. [26] utilizes 3D graph convolution (3DGC) to extract shift-invariance features for rotation estimation, and then the size and translation can be calculated using other branch networks. Similarly, [12] propose geometric consistency in 3DGC for more robust rotation-sensitive feature extraction. Nevertheless, none of these works fully harness the category-level object knowledge, resulting in the mitigation of overall performances.

C. Object Grasping and Robotic Manipulation

Object grasping and manipulation under unstructured scenarios are fundamental robotic applications [27]. Classical vision-based grasping can be divided into four key steps [28]: object localization, pose estimation, grasp detection and motion planning. Object localization involves target detection and segmentation to precisely locate the region of interest in the input image. Pose estimation aims to calculate the 3D rotation and translation of target objects. With abundant 3D information, grasp detection could be realized through analytical, empirical or DNN methods. The last step is motion planning, which mainly applies probabilistic road map (PRM) approaches to design the path from robot hand to grasp points generated before. Recently, some studies try to accomplish several steps jointly, and even in an end-to-end fashion [29], [30]. However, it is widely concerning that such methods might not be robust and flexible enough for real-world manipulation tasks [31], [32].

In this paper, to verify the effectiveness of KGNet, we deploy it with other acknowledged grasping algorithms for real-world object manipulations.

III. METHODS

A. Model Overview

In this paper, we develop the KGNet for category-level object pose and size estimation. As shown in Fig.1, the primary characteristics of this model are the application of category-level prior knowledge and synergetic feature fusion, and a more detailed architecture is illustrated in Fig.2. KGNet can be mainly divided into three parts: a). feature extraction and fusion; b). knowledge-guided target object reconstruction; c). pose and size estimation.

Given the input RGBD image, the off-the-shelf Mask-RCNN [33] is firstly applied for object detection and segmentation. Afterward, convolutional neural networks (CNN) and PointNet++ are respectively leveraged to extract RGB and depth features from the region of interest (ROI). During it, the synergetic fusion module is applied on layerwise RGB and depth features to improve their representativity (Section III.C). In the meantime, a transformer-like structure is designed to evaluate the dense correspondence between

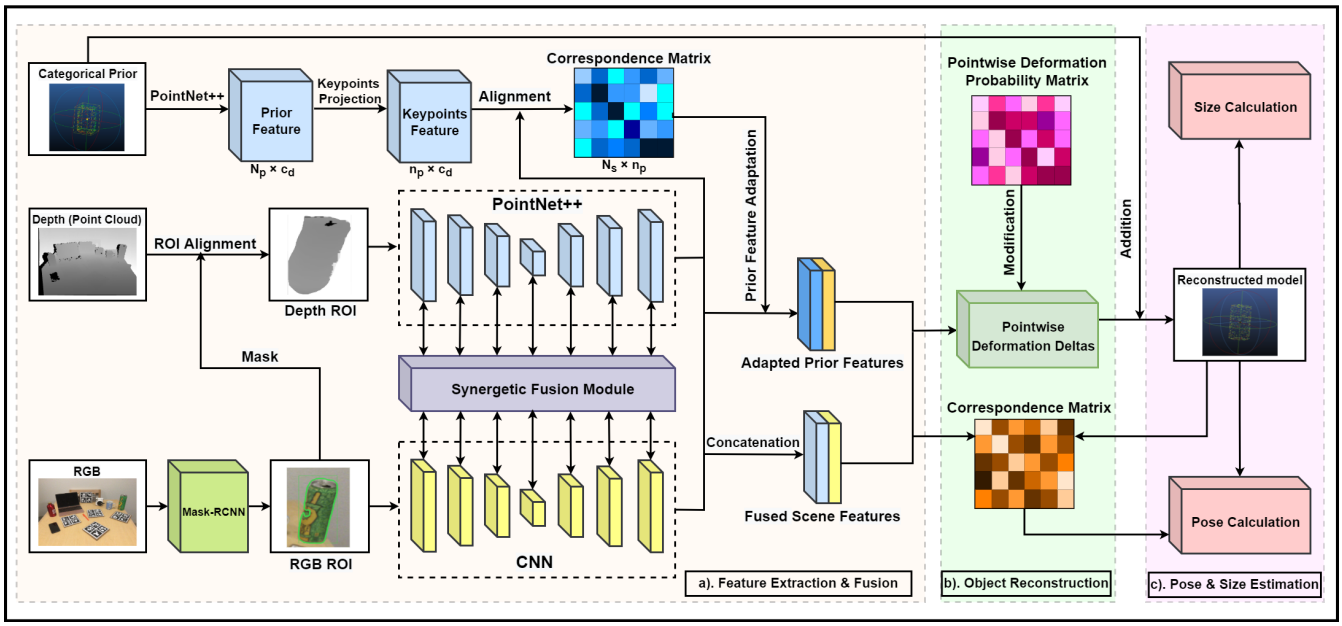


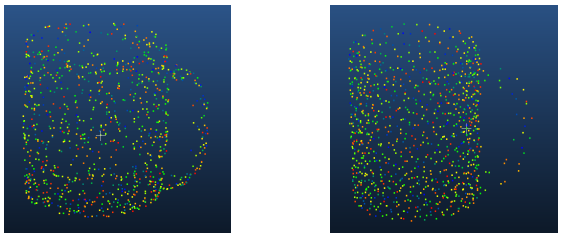
Fig. 2: Architecture of KGNet.

category-level prior model (Section III.B) and scene object point cloud, and this correspondence is then used for adapted categorical feature calculation. Combining it with observed fused features, scene object can be accordingly reconstructed under the supervision of deformation probability matrix (Section III.D), together with its 6D pose and size (Section III.E). Major model components are discussed below.

B. Knowledge-Guided Categorical Model Generation

As shown in Fig.2, the categorical prior model performs an important role in feature extraction and target object reconstruction. Therefore, it is crucial to build representative and universal categorical models. In previous studies, [11], [13] have attempted to apply mean latent embedding methods for prior model generation. Specifically, all available instance models within a certain category are fed into an autoencoder to calculate the mean latent embedding features, and these features are then passed into the decoder to get the mean shape priors. Though these approaches are considerably more robust than simple averaging, it is concerned about their deficiency in handling intra-class shape variations.

Therefore, we propose knowledge-guided methods for categorical model generation. It is common sense that objects are mainly classified by their functions, and functions are closely relevant to geometric shapes. Consequently, despite intra-class variations, instances within a certain category will have pointwise correspondence, which is the theoretical basis of our approach. For available instances in each respective category, we map them to the canonical space for points alignment and clustering. For example, mugs with different sizes are firstly normalized for dimensionless shape mapping, and subsequently a closest points searching algorithm is applied to densely align points among candidate mug models. Finally, the aligned points are clustered to obtain the categorical mug model (shown in Fig.3(a)). Compared with the categorical model generated by mean latent embedding (Fig.3(b)), the most significant characteristic of our model is the non-uniform and clustered distribution of model points, especially around the handle part. This is primarily because mean latent embedding applies global features (mainly the body part) for modeling, causing the neglect of handle part information. While in our approach, model points are generated through geometric alignment and clustering, so features for each model part are retained and represented explicitly, which improves its robustness and generalizability when adapting to unseen objects.



(a) Generated by Prior Knowledge.

(b) Generated by Mean Latent Embedding.

Fig. 3: Categorical Mug Model.

For a fair comparison, we follow [11], [13] to use the same available instances for categorical prior model generation, and the ablation study in Section IV.E could verify the effectiveness of our innovation.

C. Synergetic RGB and Depth Feature Fusion

Given the input RGBD image, off-the-self Mask-RCNN is applied for ROI selection, with bounding boxes for RGB and segmentation masks for depth. Subsequently, the depth ROI is transformed into point cloud format using camera

intrinsics, and the cropped RGB image and selected point cloud are fed into CNN and PointNet++ respectively for feature extraction. Conventionally, the extracted RGB and depth features are fused through concatenation or DenseFusion [8] in the final layer, but we consider this fusion strategy could be deficient in fully leveraging the complementarity of RGB and depth features. Accordingly, a synergetic fusion module is designed for layerwise feature interaction during extraction phases, and its detailed structure is demonstrated in Fig.4.

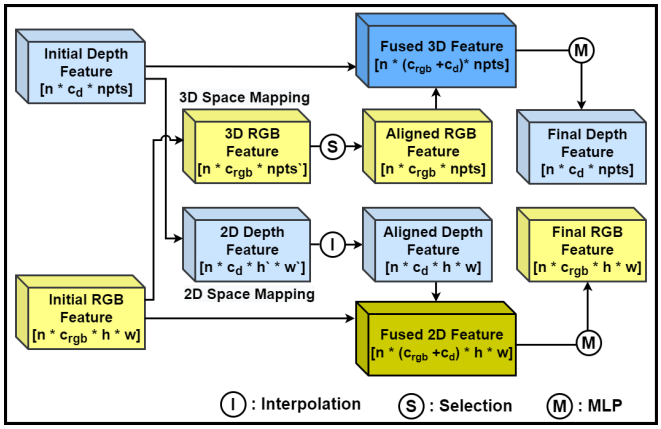


Fig. 4: Synergetic RGB and Depth Feature Fusion.

The initial RGB features are pixel-wise and in 2D space, while the depth features are pointwise and in 3D space, so we firstly map them to 3D and 2D space respectively. To be specific, the 3D space mapping means the RGB features are aligned to the point cloud to obtain the 3D RGB features; while 2D space mapping is to project pointwise depth features to RGB pixels in 2D coordinates. Afterward, though mapped to the same space, the size of 3D RGB features is still different from the initial depth features due to the inconsistency in ROI for RGB and depth, so it is with 2D depth features and the initial RGB features. Therefore, a selection module is designed to gather 3D RGB features for depth ROI, and similarly, the 2D depth features are expanded to RGB ROI using a nearest neighbor interpolation module. Finally, we concatenate the initial and aligned features for fused 2D/3D features, and two separate multi-layer perceptrons (MLPs) are applied on them to obtain the final RGB/depth features respectively.

As shown in Fig.2, the synergetic fusion module is applied to each corresponding layer during RGB and depth feature extraction. Its major advantage is to bridge the geometric and appearance features during extraction phases, thereby facilitating representation learning of pose-sensitive features.

D. Deformation Probability Matrix & Object Reconstruction

After synergetic fusion, the fused scene features are combined with categorical models for object reconstruction. Firstly, we extract pointwise geometric features from the categorical model through PointNet++, and it is noteworthy that the PointNet++ used for prior and scene depth feature extraction are weight-sharing, so they have the same feature channels. Subsequently, assume that the shapes of prior and scene depth features are N_p*c_d and N_s*c_d respectively (N_p, N_s are the number of points on categorical model and scene object; c_d represents the feature channels), we multiply a keypoints projection matrix (n_p*N_p, n_p is the number of selected keypoints on prior model) with prior features to obtain the prior keypoints features (n_p*c_d), and then a low-rank transformer is deployed to quantify their correspondences with scene features (N_s*n_p). The advantage of squeezing prior features is to improve its representativity while reducing computational complexity, and more details about its design can be found in [13]. The feature correspondence matrix (N_s*n_p) is then normalized by row, and we multiply it with prior keypoints features to obtain the adapted prior geometric features (N_s*c_d). Afterwards, for the scene depth features (N_s*c_d) and adapted prior geometric features (N_s*c_d), a similarity matrix (N_s*N_s) is computed to represent their pointwise feature similarity. Assume the scene RGB features to be N_s*c_r (c_r represents feature channels), we multiply similarity matrix (N_s*N_s) with it to obtain the adapted prior appearance features (N_s*c_r). Together with the adapted prior geometric features and fused scene features, an MLP is developed to calculate the pointwise deformation deltas for target object reconstruction.

However, directly applying MLP for deltas regression is considered to be less robust to intra-class variations, so we introduce the deformation probability matrix, as another form of categorical knowledge, to facilitate this process. The construction method for deformation probability matrix is as follows: for all available instances within a certain category, we first align them with the categorical prior model and accordingly calculate their pointwise deformation deltas (n*N_p*3, n is the number of available instances). Since n for different categories varies widely, we calculate the mean and standard deviation to represent deformation deltas and compose the deformation probability matrix (2*N_p*3). Apart from standardization among different categories, its another advantage is that the values in this matrix can be continuously updated without influencing the interfaces of subtasks. The initially regressed pointwise deltas matrix (1*N_p*3) is concatenated with deformation probability matrix and then fed into an additional layer for final pointwise deltas (1*N_p*3) computation. The deformation probability matrix brings in the statistical tendency of pointwise deformation for categorical models, thereby facilitating deltas regression, and the reconstructed object is the sum of categorical model and final pointwise deltas.

E. Object Pose and Size Estimation

With the reconstructed model, the size of target object can be estimated accordingly, while for its 6D pose, we mainly utilize the corresponding-based method [34] to calculate it. Given two sets of points on reconstructed model and scene object, an MLP is designed to regress their correspondence matrix based on pointwise feature similarity, and the dense correspondence can then be applied for 6D pose estimation through least-squares fitting algorithms.

F. Loss Function Design

As demonstrated in Fig.2, the technical route of KGNet is applying categorical model and scene features for object reconstruction and dense correspondence, so we accordingly design loss functions to supervise its training. The total loss shown in Eq.1 is composed of 5 parts: keypoints loss (L_{kp}) to guide the selection of keypoints for prior model representation; pointwise deformation deltas loss (L_{def}) to discourage large deformations and preserve the semantic consistency; object reconstruction loss (L_{cd}) to minimize chamfer distance between the reconstructed model and ground truth; correspondence loss (L_{corr}) for better alignment between reconstructed model and target object; cross-entropy loss ($L_{entropy}$) to encourage the peaked distribution of correspondence matrix. Details about loss calculation can be found in [11], [13], and we apply the same hyperparameter configurations as them for a fair comparison.

$$L_T = \alpha L_{kp} + \beta L_{def} + \gamma L_{cd} + \delta L_{corr} + \varepsilon L_{entropy} \quad (1)$$

IV. EXPERIMENTS

A. Implementation Details

The KGNet is coded on PyTorch framework, and hardware environments for model training are 24 Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz and 1 Tesla V100S-PCIE-32GB GPU. In the grasping experiment, we apply RealSense D435i as "eye-to-hand" camera for sensing data acquisition and KINOVA Gen3 as robot arm for object manipulation. Additionally, laptop with 6 Intel(R) Core i7-10750H CPU @ 2.60GHz and 1 NVIDIA GeForce GTX 1650Ti GPU is utilized to deploy KGNet and control robot arm.

For hyperparameters and general settings in our model, we mainly follow the previous works [2], [11], [13] for a fair comparison. Mask-RCNN is applied to generate object detection and segmentation results offline; the available models for categorical knowledge extraction are from ShapeNet dataset [35]; the number of points on categorical model (N_p) and scene object (N_s) are both 1024; the number of projected object keypoints (n_p) is 256; weighting factors α, β, γ, δ, ε in Eq.1 are 1, 0.01, 5, 1, 1e-4 respectively.

B. Datasets

To evaluate the effectiveness of KGNet, two acknowledged category-level pose estimation benchmarks are applied, **CAMERA25** and **REAL275** [2]. These two datasets both cover 6 object categories: bottle, bowl, camera, can, laptop and mug. CAMERA25 is a synthetic dataset that contains 300K virtual RGBD images generated by compositing synthetic objects into real backgrounds, and following previous works [2], [12], [26], 25K images among them are set aside for testing. As for REAL275, it consists of 8K annotated real-world RGBD images, where 4.3K for training, 0.95K for validation, and the remaining 2.75K for testing.

C. Evaluation Metrics

Following widely adopted evaluation schemes, 3D intersection over union (3DIoU) and rotation/translation errors (n°m *cm*) are calculated to quantitatively represent the model performance. More specifically, 3DIoU is related to object reconstruction and size estimation, it computers the overlap ratio of estimated and ground truth 3D bounding box, and we report $3DIoU_{50}$, $3DIoU_{75}$ for the percentage of objects with overlap ratio larger than 50% and 75%. As for rotation/translation errors, we calculate the ratio of objects with pose errors smaller than specific thresholds: 5°2 *cm*, 5°5 *cm*, 10°2 *cm* and 10°5 *cm*.

TABLE I: The Performance of KGNet on CAMERA25.

Model	CAMERA25 Dataset					
	$3D_{50}$	$3D_{75}$	5°2 *cm*	5°5 *cm*	10°2 *cm*	10°5 *cm*
NOCS [2]	83.9	69.5	32.3	40.9	48.2	64.6
SPD [11]	**93.2**	83.1	54.3	59	73.3	81.5
DualPose [36]	92.4	86.4	64.7	70.7	77.2	84.7
SGPA [13]	**93.2**	88.1	70.7	74.5	82.7	88.4
Ours	92.9	**88.6**	**73.1**	**77.1**	**84.5**	**89.5**

TABLE II: The Performance of KGNet on REAL275.

Model	REAL275 Dataset					
	$3D_{50}$	$3D_{75}$	5°2 *cm*	5°5 *cm*	10°2 *cm*	10°5 *cm*
NOCS [2]	78	30.1	7.2	10	13.8	25.2
SPD [11]	77.3	53.2	19.3	21.4	43.2	54.1
DualPose [36]	79.8	62.2	29.3	35.9	50.0	66.8
SGPA [13]	80.1	61.9	35.9	39.6	61.3	70.7
FSNet [26]	**92.2**	63.5	/	28.2	/	60.8
GPV-Pose [12]	83	64.4	32	42.9	/	73.3
Ours	81.6	**67.7**	**40**	**44.3**	**64.2**	**73.9**

D. Quantitative Results

TABLE I demonstrates the quantitative performances of KGNet on the CAMERA25 dataset, manifesting that our model achieves the SOTA results among most evaluation metrics. Additionally, the performance improvement is more significant on REAL275 dataset shown in TABLE II. We exceed the previous SOTA GPV-Pose [12] by a considerable margin, especially among $3D_{75}$ and 5°2 *cm* metrics, they are increased by 3.3% and 8.0% respectively. It needs to be noted that FSNet [26] is less comparable with other works for applying different detection and segmentation results. Compared with CAMERA25, REAL275 is a realistic and more challenging dataset, thereby valued in lots of works. The outstanding performance of KGNet on this dataset not only generally proves the effectiveness of categorical knowledge and synergetic feature fusion, but also demonstrates its robustness and practicability in real-world object manipulation tasks.

E. Ablation Study

From quantitative results, the overall performance of the KGNet could be proven. In addition to this, an ablation study is conducted to further verify the effectiveness of three major innovations in our paper: knowledge-guided categorical model generation (KGM); deformation probability matrix (DPM) and synergetic feature fusion (SFF). The results are shown in TABLE III, and we could notice that with the introduction of categorical knowledge in prior model generation and pointwise deformation tendency, the model performances on size (3DIoU metrics) and 6D pose estimation (n°m *cm*

TABLE III: Ablation Study on CAMERA25 and REAL275.

KGM	DPM	SFF	CAMERA25						REAL275					
			$3D_{50}$	$3D_{75}$	$5°2\ cm$	$5°5\ cm$	$10°2\ cm$	$10°5\ cm$	$3D_{50}$	$3D_{75}$	$5°2\ cm$	$5°5\ cm$	$10°2\ cm$	$10°5\ cm$
			93.2	88.1	70.7	74.5	82.7	88.4	80.1	61.9	35.9	39.6	61.3	70.7
✓			93	88.5	70.9	74.6	83.6	88.8	80.9	65.2	36.8	40.8	60.9	71
✓	✓		92.8	88.4	72.1	75.8	83.7	88.8	81.5	65.4	38.7	43	60.5	71.1
✓	✓	✓	92.9	**88.6**	**73.1**	**77.1**	**84.5**	**89.5**	**81.6**	**67.7**	**40**	**44.3**	**64.2**	**73.9**

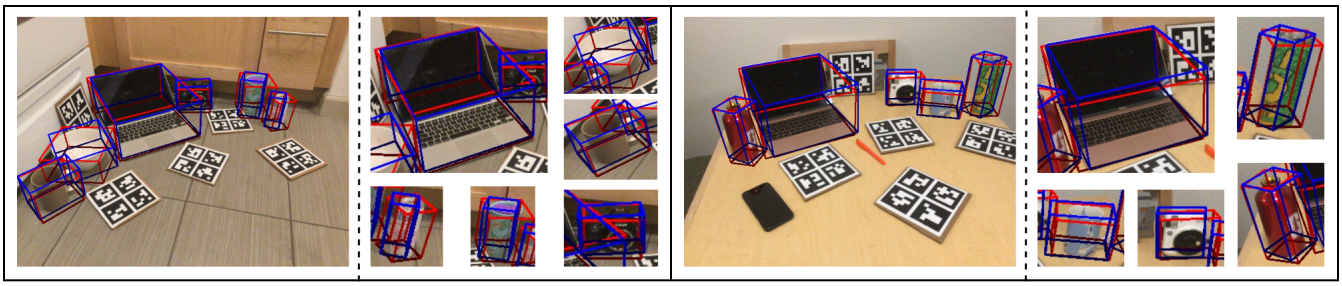


Fig. 5: Visualization on REAL275.

metrics) both have a noticeable improvement, especially on the REAL275 dataset, the $3D_{75}$ and $5°2\ cm$ are improved from 61.9% to 65.4% and 35.9% to 38.7% repressively. While the novel feature fusion module is mainly conducive for 6D pose estimation, verifying its ability to extract more pose-sensitive scene features. Generally, compared with the baseline method which applies mean latent embedding for prior model generation and fuses RGBD features by simple concatenation, the application of KGM, DPM and SFF can jointly enhance model performances in pose and size estimation. Besides, the SFF module is also embeddable and can be adaptively generalized in other relevant works to improve the representativity of RGBD features.

F. Visualization and Robotic Application

In this section, we demonstrate the qualitative evaluations of KGNet on REAL275. As shown in Fig.5, the red and blue 3D bounding boxes represent the predicted and ground truth poses respectively, and rotational error around the axis of symmetry for symmetrical objects (bowl, bottle, can) should be ignored. It is apparent that the KGNet achieves competitive performances on rotation and translation estimation, even for complex objects like mugs and cameras.

Moreover, intelligent robots are expected to not only be able to perceive the environment, but also interact with it [28]. Therefore, we perform a real-world unseen objects grasping experiment to further verify the effectiveness of visual perception. Following classical approaches in II.C, the off-the-shelf Mask-RCNN is applied for object localization, and then our KGNet can accordingly estimate the pose of target objects together with their 3D models. Afterward, grasp points are generated on target objects through analyzing their 3D models, and finally RRT-Connect [37] algorithm is leveraged for motion planning and grasping. Fig.6 demonstrates partial results of visual perception, and it is noticeable that our model can precisely calculate the pose of target objects, even under heavy occlusion. Additionally, on NVIDIA GeForce GTX 1650Ti, the inference speed of a single object reaches 20fps, meeting the requirements of real-time robotic manipulation.

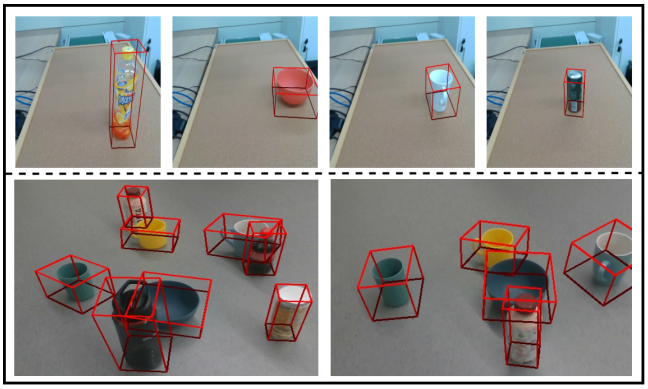


Fig. 6: Visualization for Real-world Object Pose Estimation.

V. CONCLUSIONS

To conclude, we propose a novel knowledge-guided network - KGNet for category-level object pose and size estimation. The primary characteristics of this network are the introduction of categorical object knowledge and synergetic RGBD feature fusion. For the former one, it includes the categorical prior model to represent universal features of objects within a certain category, and the pointwise deformation probability matrix to demonstrate intra-class variations during prior model deformation. While the latter one bridges the complementary RGB and depth features in extraction phases, thereby improving feature representativity and sensitivity. An ablation study has proven their effectiveness, and our KGNet achieves the SOTA performances on category-level object pose estimation benchmarks - CAMERA25 and REAL275, especially the REAL275, the previous SOTA $3D_{75}$ and $5°2$ *cm* are improved by 3.3% and 8.0% respectively. Additionally, we also conduct a real-world grasping experiment based on KGNet, demonstrating its robustness and practicability in robotic applications.

REFERENCES

[1] C. Sahin and T.-K. Kim, "Category-level 6d object pose recovery in depth images," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 0–0, 2018.

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2642–2651, 2019.

[3] E. Marchand, H. Uchiyama, and F. Spindler, "Pose estimation for augmented reality: a hands-on survey," *IEEE transactions on visualization and computer graphics*, vol. 22, no. 12, pp. 2633–2651, 2015.

[4] J. Yang, Y. Lei, Y. Tian, and M. Xi, "Deep learning based six-dimensional pose estimation in virtual reality," *Computational Intelligence*, vol. 38, no. 1, pp. 187–204, 2022.

[5] Z. Tang, R. Gu, and J.-N. Hwang, "Joint multi-view people tracking and pose estimation for 3d scene reconstruction," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, IEEE, 2018.

[6] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," *arXiv preprint arXiv:1809.10790*, 2018.

[7] 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)*, pp. 3665–3671, IEEE, 2020.

[8] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3343–3352, 2019.

[9] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4561–4570, 2019.

[10] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3003–3013, 2021.

[11] M. Tian, M. H. Ang, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation," in *European Conference on Computer Vision*, pp. 530–546, Springer, 2020.

[12] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6781–6791, 2022.

[13] K. Chen and Q. Dou, "Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2773–2782, 2021.

[14] Y. Ma, D. Lin, B. Zhang, Q. Liu, and J. Gu, "A novel algorithm of image gaussian noise filtering based on pcnn time matrix," in *2007 IEEE International Conference on Signal Processing and Communications*, pp. 1499–1502, IEEE, 2007.

[15] A. Bais, R. Sablatnig, and J. Gu, "Single landmark based self-localization of mobile robots," in *The 3rd Canadian Conference on Computer and Robot Vision (CRV'06)*, pp. 67–67, IEEE, 2006.

[16] F. Wang, "Simulation of registration accuracy of iterative closest point (icp) method for pose estimation," in *Applied Mechanics and Materials*, vol. 475, pp. 401–404, Trans Tech Publ, 2014.

[17] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[18] Y. Bukschat and M. Vetter, "Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach," *arXiv preprint arXiv:2011.04307*, 2020.

[19] D. Chen, J. Li, Z. Wang, and K. Xu, "Learning canonical shape space for category-level 6d object pose and size estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11973–11982, 2020.

[20] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Asian conference on computer vision*, pp. 548–562, Springer, 2012.

[21] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *2011 international conference on computer vision*, pp. 858–865, IEEE, 2011.

[22] S. Gold, A. Rangarajan, C.-P. Lu, S. Pappu, and E. Mjolsness, "New algorithms for 2d and 3d point matching: pose estimation and correspondence," *Pattern recognition*, vol. 31, no. 8, pp. 1019–1031, 1998.

[23] L. Zhou, S. Wang, and M. Kaess, "A fast and accurate solution for pose estimation from 3d correspondences," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1308–1314, IEEE, 2020.

[24] W. Zhang, W. Zhang, K. Liu, and J. Gu, "A feature descriptor based on local normalized difference for real-world texture classification," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 880–888, 2017.

[25] T.-T. Do, M. Cai, T. Pham, and I. Reid, "Deep-6dpose: Recovering 6d object pose from a single rgb image," *arXiv preprint arXiv:1802.10367*, 2018.

[26] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1581–1590, 2021.

[27] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, 2018.

[28] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021.

[29] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11444–11453, 2020.

[30] M. Gou, H.-S. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, "Rgb matters: Learning 7-dof grasp poses on monocular rgbd images," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 13459–13466, IEEE, 2021.

[31] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6401–6408, IEEE, 2022.

[32] K. Fang, Y. Zhu, A. Garg, A. Kurenkov, V. Mehta, L. Fei-Fei, and S. Savarese, "Learning task-oriented grasping for tool manipulation from simulated self-supervision," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 202–216, 2020.

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

[34] S. Salti, F. Tombari, and L. Di Stefano, "Shot: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251–264, 2014.

[35] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[36] J. Lin, Z. Wei, Z. Li, S. Xu, K. Jia, and Y. Li, "Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3560–3569, 2021.

[37] J. J. Kuffner and S. M. LaValle, "Rrt-connect: An efficient approach to single-query path planning," in *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, vol. 2, pp. 995–1001, IEEE, 2000.