Interactive Shape Estimation for Densely Cluttered Objects

Jiangfan Ran[1], Zhenyu Wu[1], Ziwei Wang[2] and Haibin Yan*[1]

***Abstract*— Accurately recognizing the shape of objects in dense and cluttered scenes is important for robots to perform a variety of manipulation tasks, such as grasping and packing. However, the performance of previous shape estimation methods is not satisfactory due to the heavy occlusion between objects in dense clutter. In this paper, we propose an interactive exploration framework to estimate the shape of densely cluttered objects. Our framework utilizes pixel-wise uncertainty to generate efficient interactions, allowing to achieve a better trade-off between the shape estimation accuracy and the interaction cost. Specifically, the extracted features are utilized as network weights to predict the confidence of each proposal being located on the surface of the objects. Proposals with higher confidence are considered reliable results for shape estimation. Meanwhile, we obtain the uncertainty of shape and scale estimation based on the confidence of each proposal, and further propose the adaptive fusion strategy to construct the pixel-wise estimation uncertainty height map. In addition, our proposed interaction strategy leverages the uncertainty height map to generate effective interaction actions to significantly improve the shape estimation accuracy for severely occluded objects. Therefore, the optimal accuracy-efficiency trade-off for shape estimation in dense clutter is achieved by iterating the shape estimation and interaction actions. Extensive experimental results verify the effectiveness of the proposed approach. Under challenging scene, the proposed approach has 66.7% and 52.0% less average Chamfer distance than direct estimation and random interaction, respectively.**

I. INTRODUCTION

With the rise of deep learning technology, there has been rapid development in robot manipulation tasks. However, cluttered environments such as warehouses and homes can be heavily occluded and highly crowded, which results in challenging robot manipulation tasks. Accurate estimation of the shapes of all objects in a dense scene is a prerequisite requirement for various robot manipulation tasks, such as grasping [1]–[7], packing [8]–[11], and rearranging [12]–[16]. Therefore, it is necessary to design frameworks which can accurately estimate the shapes of all the objects in the clutter.

The conventional shape estimation approaches ignore the occlusion of objects in practical robot deployment scenes, which leads to poor robustness to object point cloud completeness in deployment. To improve the performance of object shape estimation in obscured environments, visual detection and shape estimation models for occlusion environments have been proposed [17], [18], which obtain visual clues about heavily occluded objects by enhancing the discriminative features of localized or overlapping regions. However, objects in dense cluttered scenes are missing significant visual clues due to partial or even complete occlusion. Existing shape estimation methods for occluded objects primarily focus on feature extraction to enhance estimation performance, which results in previous approaches performing poorly in dense and cluttered scene deployment.

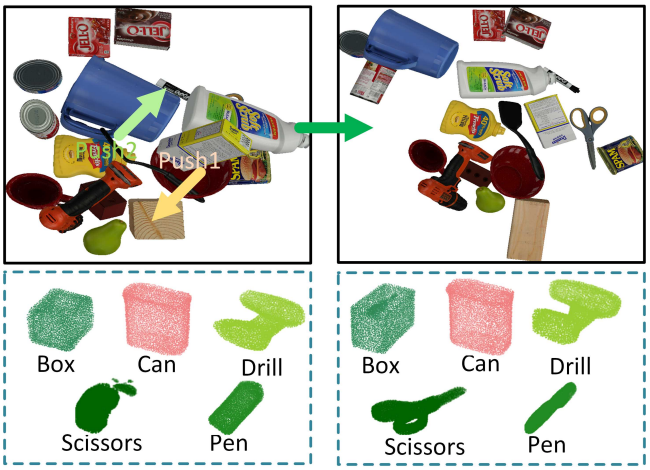

Fig. 1. Example of interactive exploration in dense clutter. The goal of shape estimation in clutter is to estimate the point cloud of each object, as shown at the bottom. Left and right sides demonstrate the scene before and after the interaction. Interacting with the clutter through iterative actions can significantly improve the performance of the framework.

In this paper, we propose an active exploration framework for shape estimation of dense cluttered scene objects. Compared to conventional methods that enhance shape estimation performance through feature extraction, our approach utilizes pixel-wise uncertainty to generate efficiently interacting actions that maximize the accuracy of shape estimation. Specifically, we utilize the extracted features as network weights for predicting the confidence of each estimated proposal obtained through Gaussian distribution sampling. Higher confidence proposals provide accurate representations of object shapes and output as reliable shape estimates. We propose to measure the uncertainty of shape and scale estimation by the bias of the estimation results from the actual shape with confidence of each output proposal. Meanwhile, we adaptively fuse the uncertainties of shape and scale proposals into pixel-wise uncertainty height maps based on the distance to the center point of the objects, which intuitively responds to the effect of estimating the shape of the objects in the dense clutter scenes. The starting point

*Corresponding author.

[1]Jiangfan Ran, Zhenyu Wu and Haibin Yan are with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, 100876, China. {ran-jf06, wuzhenyu, eyanhaibin}@bupt.edu.cn

[2]Ziwei Wang is with tthe Department of Automation, Tsinghua University, and Beijing National Research Center for Information Science and Technology (BNRist), Beijing, 100084, China. wang-zw18@mails.tsinghua.edu.cn

of the interaction action is generated with the uncertainty height map by reducing the uncertainty in the clutter, with the direction pointing to the region with the least uncertainty in the expectation of further disturbing the occlusion situation to improve the efficiency of the interaction. Fig. 1 demonstrates the exploration of our framework in clutter scenes. Compared to the direct estimation, our approach reduces the Chamfer distance for shape estimation from 3×10^{-3} to 1×10^{-3} using only 6.3 interactions on average in challenging cases.

- We propose an active exploration framework for object shape estimation in dense cluttered scenes, which provides the necessary object shape information for various robotic manipulation tasks.
- We present an interactive strategy that generates interactive actions based on adaptively fused uncertainty, significantly improving estimation accuracy by altering clutter structure.
- Extensive experiments are completed, which illustrate that our framework outperforms not only the direct estimation but also the randomly generated interaction strategy.

II. RELATED WORK

Shape Estimation: The primary objective of shape estimation is to acquire detailed 3D information about an object. Currently, shape estimation can be classified into two main categories: RGB-D shape reconstruction methods [19]–[22] and point cloud-based completion methods [17], [23]–[25]. RGB-D shape reconstruction methods combine object depth information and color information to output point clouds, voxels, etc. to accomplish object perception. Irshad *et al.* [19] utilizes a structure similar to CenterNet [26] to predict the 3D information representation vectors to obtain information such as the shape of the object. The point cloud-based approach infers a complete shape representation by using a partial point cloud as input. Rosasco *et al.* [17] represent the object shape as an implicit function that generates a confidence for each point and reconstructs the point cloud from the implicit function species by means of a gradient-based sampling algorithm. Wu *et al.* [23] predicted the deformation parameters of a template for shape estimation by dense fusion of feature maps of visual information from multi-view RGB images and pixels learned from a cluttered point cloud. Compared with previous shape estimation methods, our proposed approach effectively reduces the occlusion of objects in clutter through interactive actions and significantly improves the accuracy of object shape estimation in cluttered scenes.

Active exploration in visual perception: Active exploration of robots is a methodology employed to extract scene perception information through adaptive observation and interaction with objects. As a result, this approach has a broad range of applications and is extensively researched in vision-guided robotic autonomous manipulation tasks, including perception [27]–[30], target search [31]–[35], and mapping [36]–[40]. In the perception task, Wu *et al.* [27] proposed an interactive exploration framework to model object uncertainty through instance segmentation entropy and multi-view object inconsistency and generate interactive actions based on uncertainty and spatial physical constraint relationships. For mapping, Asgharivaskasi *et al.* [39] used a Bayesian multiclass mapping algorithm for range category measurements to derive a computationally efficient closed lower bound for the Shannon mutual information between multiclass maps and measurements for evaluating potential robot trajectories. In order to locate and extract known targets from clutter, researchers have proposed target search. Ye *et al.* [31] explored the environment through an intrinsically rewarding subgoal-driven agent low-level strategy and further learned hierarchical strategies through exploration experience to achieve optimization of both high-level and low-level strategies. However, previous active exploration methods ignore object geometric information and measure only the regional uncertainty, which leads to low performance of scale estimation in severely occluded scenes. Our proposed framework fully utilizes object geometric information to generate pixel-wise uncertain height maps, achieving an improved accuracy-efficiency trade-off.

III. APPROACH

In this section, we will begin by introducing the problem of shape estimation for objects in dense clutter and providing an overview of the overall pipeline. Subsequently, we will delve into the details of the shape estimation implementation, followed by a discussion on the definition of uncertainty. Finally, we will present the design of an interaction strategy aimed at generating valid interaction actions.

A. Problem Statement and Overall Pipeline

The objective of accuracy shape estimation in clutter can be expressed as follows:

$$\hat{Y} = f(P) \quad s.t. \quad m(\hat{Y}, Y) \leq \delta \tag{1}$$

where P is the input partial point cloud, f denotes shape estimation function, $\hat{Y}$ and Y represent the output shape estimation and the ground truth, respectively, m and δ respectively represent a shape estimation metric and the threshold. The performance of the framework is significantly affected by the lack of visual clues caused by mutual occlusion between objects and high object crowding in cluttered environments. To address the visual clues loss caused by occlusion, we leverage a hyper-network to map the partial point cloud to an implicit representation of densely cluttered objects from which we sample shape, scale, and confidence. We define shape scale uncertainty utilizing confidence and adaptively merge the two uncertainties, depending on the distance to the central point. Since uncertainty reflects occlusion between objects, we generate interactive actions based on uncertainty to alter the clutter structure and enhance the performance of framework.

The overall pipeline of our framework is illustrated in Fig. 2. The clutter is observed solely through a top-view RGB-D camera. Our framework first takes the partial point cloud as input and generates implicit representations. Candidate points are sampled from a Gaussian distribution, and flow towards

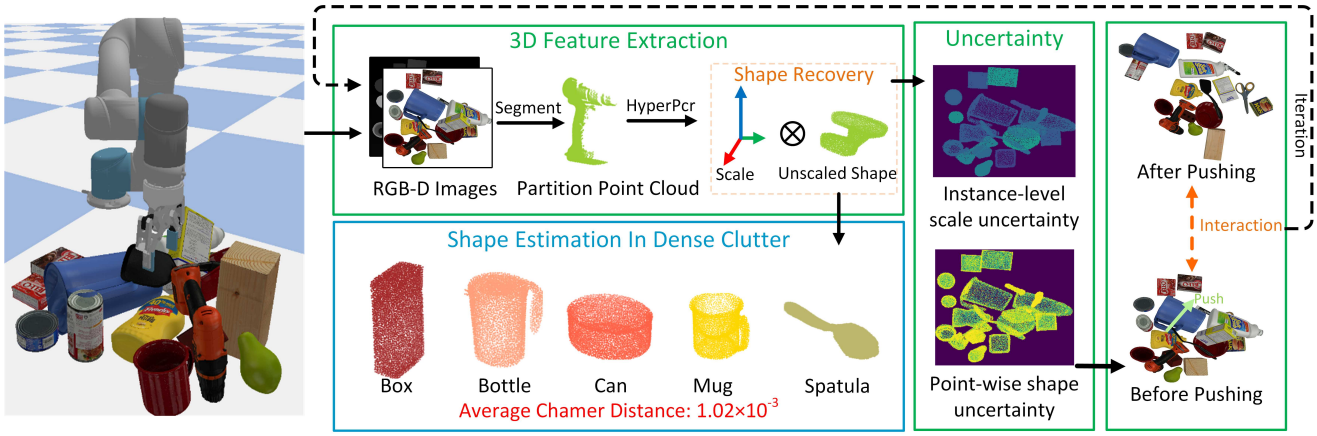

Fig. 2. The pipeline of our interactive exploration framework. Observation of dense clutter through a top camera and estimation of 3D information of scene objects, such as scale, and shape. Specifically, the framework first utilizes the scale factor estimation results to obtain the unscaled deformed partial point cloud, and the shape information. Our framework generates uncertainty for each point of the object, fuses the two uncertainties and projects them into an uncertainty height map, and finally generates effective interaction operations based on the uncertainty height map. The shape estimation is realized by predicting the scale deformation factor and the original point cloud based on the scale estimation and point cloud completion through object 3D information perception and physical interaction actions.

the point cloud surface or scale boundary is obtained for each candidate point using gradient descent. The actual point cloud of each object is generated by integrating the shape and scale estimation. During combining scale uncertainty and shape uncertainty, we consider the distance of the point from the centroid because the farther the scale-deformed objects are from the centroid, the greater the difference becomes.

In our approach, we exclusively utilize pushes parallel to the ground as action primitives. Since the merged uncertainty represents estimation bias, the start point, pushing direction and distance are generated based on the estimation uncertainty to provide the informative visual clues with high motion efficiency.

B. Scale and Shape Estimation

Our framework aims to generate uncertainty for all points of shape estimation while performing shape estimation. Inspired by the shape uncertainty of HyperPcr [17], we propose scale uncertainty. Additionally, the secondary network utilizes a deeper MLP, with its weights generated by the backbone network. Moreover, we have adapted the training process for shape estimation to enable it to sense the degree of scale deformation of the object and produce uncertainty regarding the scale. In this subsection, we describe the whole scale shape estimation process in detail.

To estimate the scale deformation factor and generate the corresponding uncertainty, we have employed an approach similar to shape estimation. Specifically, we define a parameter function $g_\phi = \mathbb{R}^3 \to \{0, 1\}$, such that:

$$g_\phi(p) = \begin{cases} 1, & \text{if IoU}(p \cdot \hat{Y}, Y) \geq t \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\hat{Y}$ is the unscaled deformed point cloud, Y is the scaled deformed point cloud and $p = (\alpha_x, \alpha_y, \alpha_z)$ is the scale deformation factor under the object coordinate system, t respectively a threshold. The output of the implicit function is 1 when the IoU between Y and $\hat{Y}$ is greater than the threshold, and 0 otherwise. We generate training data by applying randomly generated scale deformation factors to the CAD model. Specifically, we introduce noise to randomly generated scale factors and apply the noisy scale factors to the original CAD model. We obtain the IoU score between the CAD model with noisy scale factors applied and the original CAD model without noisy scale factors applied. Subsequently, we assign appropriate labels based on Eq. 2.

Our training process is as shown in Fig. 3. We input the 3D vector proposal sampled from a Gaussian distribution to the implicit function and obtain values between 0 and 1 via the sigmoid activation function. The parameters of the backbone network are updated according to the loss between the prediction confidence and the label, using the binary cross entropy as a supervisory signal.

During inference, we apply a backward gradient propagation based approach to capture the actual output from the implicit function. To be able to perform backward gradient propagation, we construct an L1 loss function $\mathcal{L}$ subtracting confidence from a probability of 1:

$$\mathcal{L}(\hat{y_i}, 1) = |1 - \hat{y_i}|, \ \hat{y_i} = g_\phi(p_i) \tag{3}$$

where $\hat{y_i}$ is the confidence of the implicit function with respect to the i_{th} 3D vector. We can define uncertainty through:

$$u_i = 1 - \hat{y_i} \tag{4}$$

if p_i is the i_{th} 3D vector sampled from a Gaussian distribution, then we optimize p_i using the following formula:

$$p_{i,j} = p_{i,j-1} - \eta \nabla_p \mathcal{L}(\hat{y_i}, 1) \tag{5}$$

where j refers to the j_{th} iteration, η is the learning rate. Optimizing all points obtained by sampling from the Gaussian distribution as described above. After n iterations, we select the 3D vector with the biggest confidence as the scale estimation:

$$s = \arg\max_{i \in N} \ g_\phi(p_{i,n}) \tag{6}$$

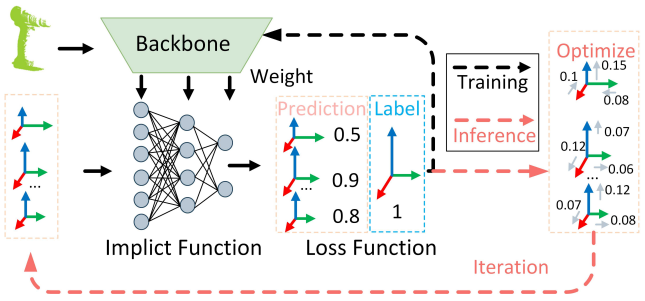

Fig. 3. The training and inference process of scale estimation. We employ a hyper-network architecture for scale estimation, where the secondary network leverages a MLP as an implicit representation and is provided with weights by the primary network. In prediction, the numbers to the right of each proposal represent confidence. During inference, gray arrows represent gradient directions, and the numbers represent specific values. All solid-line segments will be executed in both training and inference.

where N is the number of 3D vectors sampled from the Gaussian distribution.

C. Uncertainty of Object in Clutter

The framework has the ability to generate fine-grained uncertainty for generating interaction actions. However, shape uncertainty is only related to the original shape and does not take into account the scale deformation. Since the scale deformation is performed in the object coordinate system, it leads to a greater difference in the shape of the deformed point cloud region further away from the center point, compared to the undeformed point cloud region. Taking these factors into consideration, we define the fusing uncertainty of the object as follows:

$$U_{i,j} = \lambda \cdot U_{i,j}^{shape} + (1-\lambda) \cdot U_j^{scale} e^{-\frac{r-d}{r}} \tag{7}$$

where $U_{i,j}$ and $U_{i,j}^{shape}$ respectively represent the fusing uncertainty and shape uncertainty of the i_{th} point of the j_{th} object, U_j^{scale} denotes the scale uncertainty of the j_{th} object, λ is the hyperparameter that balances scale uncertainty and shape uncertainty, and r and d denote the radius of the outer circle of the object and the distance between the i_{th} point and the center point of the object, respectively.

The two uncertainties are generated by the bias between the implicit representation based on object shape labels and the actual prediction. Therefore, the uncertainty defined by Eq. 7 reveals the bias in shape estimation during scale deformation. By employing the uncertainty fusion approach described, we effectively integrate instance-level scale uncertainty with pixel-wise shape uncertainty to model the overall uncertainty of the object.

D. Interaction Strategy

A reasonable interaction strategy not only needs to efficiently explore the scene but also aims to minimize the number of interactions as much as possible. Therefore, we need to design an efficient interaction strategy based on the occlusion information provided by the uncertainty height map between objects. As shown in Fig. 4, our framework assigns a higher uncertainty to the invisible region, taking into account that the invisible region tends to have higher

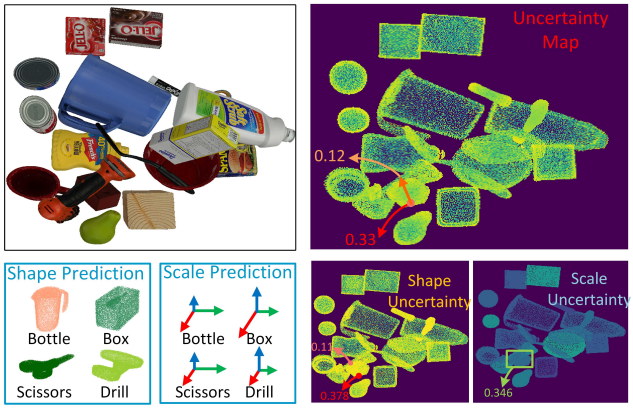

Fig. 4. Top-down view of the clutter and the uncertainty fusion map. After fusing the two uncertainty height maps, our interaction framework will generate interaction actions accordingly. The interaction action will prioritize the interaction with the object with the highest average uncertainty and select the point with the lowest uncertainty as the starting point and the lowest point pointing to the highest point as the direction. A fixed constant is chosen as the distance to ensure that small objects can be fully interacted with.

visual information. Therefore, our interaction action selects the object with the highest average uncertainty as the interaction object and picks the object with the maximum value in the uncertainty height map as the starting point:

$$A_s = \underset{i \in O_{max}}{\arg\max}\ U_{map}(i) \tag{8}$$

where A_s represents the starting point of interaction, U_{map} is the uncertainty height map, i is the i_{th} pixel point in the uncertainty height map and O_{max} is the object with the highest average uncertainty.

Furthermore, considering the feature of uncertainty, the visible region tends to have lower uncertainty and contains less visual information. Taking this into account, the interaction direction is selected as follows:

$$\vec{d} = \underset{i \in O_{max}}{\arg\min}\ U_{map}(i) - \underset{i \in O_{max}}{\arg\max}\ U_{map}(i) \tag{9}$$

where $\vec{d}$ denotes the direction of interaction. By combining the expression of the starting point, the interaction direction is determined as the direction with the highest uncertainty pointing towards the point with the lowest uncertainty. Regions with lower uncertainty contain fewer visual clues and are also sparser in terms of object distribution. Therefore, the physical meaning of the interaction direction is to point from dense regions toward sparse regions, allowing for a more efficient alteration of cluttered structures and exposing more visual clues from the object.

Regarding the distance, we take into account the limitations of the working area and the effectiveness of the interaction. As a result, we choose a fixed distance as the distance for our interaction strategy.

IV. EXPERIMENTS

In this section, we conduct extensive experiments in the simulator Pybullet [41]. We first present the details of the

Fig. 5. The selected YCB subset of objects in our experiments.

Fig. 6. Visualization of random scenes, from left to right, corresponding to easy, normal, and hard, with corresponding number of instances of 10, 15, and 20.

experiments. Second, the evaluation metrics of our framework are presented. After that, we compare it with randomly generated interactions. Overall, we experimentally verify : 1) the performance of the framework can be significantly improved by interaction exploration 2) our interaction strategy surpasses randomized interactions in terms of accuracy and efficiency.

A. Implementation Details

In the simulation environment, our chose a rectangular region of size $1m \times 1m \times 1m$ as the constraint space for robot exploration. In world coordinates, we set the resolution of each pixel to $2mm$, so our working area is discretized into a square area of 500×500, which is the resolution of our uncertainty height map and height map.

We leverage 100K samples collected in Pybullet for training. For the shape estimation, w set the thresholds for positive labeling to IoU_{75} and 0.005, respectively, and update the model parameters by the Adam optimizer, with the learning rate set to 1×10^{-4} and batch-size of 32. During inference, the 3D vectors are also updated by the Adam optimizer with the learning rate set to 0.1, the uncertainty threshold added to the output list is set to 0.2, and 20 iterations of optimization are performed. In terms of uncertainty fusion, we set the hyperparameter λ, which balances the two uncertainties, to 0.8. For the interaction strategy, we set the fixed distance to 15 cm, the uncertainty threshold to 0.15, and limit only 2 interactions to the object. To avoid trivial interactions, we set the total number of interactions to 10. Thus, the interaction ends when one of the following two conditions is met: 1) the number of interactions reaches the upper limit 2) all object uncertainties are less than the uncertainty threshold once.

TABLE I

PERFORMANCE IN 10 TO 20 INSTANCES OF RANDOM CLUTTER, WHERE MOTION DENOTES THE AVERAGE NUMBER OF MOVEMENTS.

Method	CD($\times 10^{-4}$)	Better Rate	Motion
Baseline	25.3	-	-
Random	19.4	62.4	6.7
Ours	12.8	84.2	2.5
Random	20.0	61.7	10.0
Ours	10.6	84.0	4.5
Random	19.3	54.5	10.0
Ours	9.1	81.2	6.3

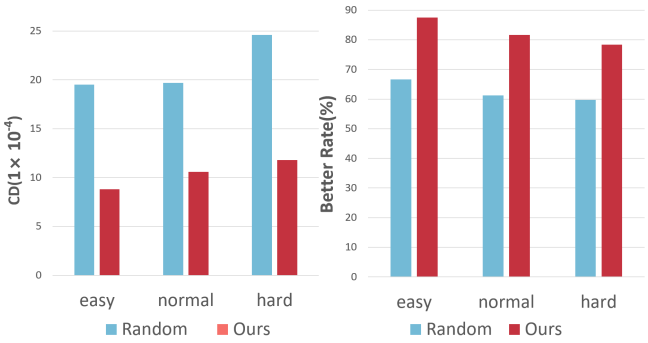

Fig. 7. Average Chamfer distance(the lower the better) and better rate(the higher the better) at different levels of difficulty in shape estimation.

The objects involved in our experiment are shown in Fig. 5. All object CAD models in our simulator are from the YCB dataset. Based on the number of instances in the scene, we classify the scene difficulty into easy, normal, and hard, with the number of instances corresponding to 10, 15, and 20. All training experiments are accelerated by NVIDIA GeForce RTX 3090 GPU.

B. Evaluation Metrics

We evaluated the interactive exploration framework in terms of estimation accuracy, better rate, and motion efficiency. We measure the estimation performance of the framework through Chamfer distance. For motion efficiency, we evaluate by the average number of interactions. The better rate is assess by comparing the number of objects with reduced Chamfer distance before and after the interaction as a percentage of the total. It defines as follows:

$$BR = \frac{N_b}{T} \tag{10}$$

Where BR is the better rate, N_b is the number of Chamfer distance that get reduced after the comparison interaction, and T is the total number of objects in the scene.

C. Results and Discussions

In our experiments, we utilize a UR5 robot with a robotiq85 gripper to carry out active exploration tasks. To assess the effectiveness of our interactive exploration framework, we compare its performance against an uninteracted shape estimation method. Additionally, we evaluate the significance of generating interaction strategies based on the uncertainty by comparing our method with randomized interaction actions. Furthermore, we conduct experiments using different uncertainty thresholds to evaluate the influence of these

Fig. 8. Visualization of six challenging scene that make clutter more messy by reducing the range of drop points. Each scene has 20 instances.

TABLE II

PERFORMANCE IN CHALLENGING CLUTTER WHERE EACH SCENE CONTAINS 20 INSTANCES OF HEAVILY OCCLUDED.

Method	CD($\times 10^{-4}$)	Better Rate	Motion
Baseline	35.4	-	-
Random	25.6	58.3	5.3
Ours	14.8	85.5	2.5
Random	25.1	60.2	10.0
Ours	12.4	83.1	5.3
Random	24.6	59.7	10.0
Ours	11.8	78.4	7.0

thresholds on the execution accuracy and efficiency of the framework.

1) Random clutters: Considering that the number of objects in the clutter is positively correlated with the estimation difficulty, we generate the randomized clutter by randomly selecting 10, 15, and 20 objects, as shown in Fig. 6. The performance of our framework in random clutter is shown in Table I. Comparing the no-interaction action approach under three uncertainties, our interactive exploration framework reduces the Chamfer distance by 49.4%, 57.0%, and 64.4%, respectively. While the random interaction actions only reduced 23.3%, 21.0%, and 23.7%, and in terms of the better rate, our interaction strategy achieved a large advantage over the random interaction. Fig. 7 also demonstrates the Chamfer distance and improvement rate at various levels of difficulty. Meanwhile, in terms of the number of interaction actions, our approach has fewer actions than random interactions in all cases.

As a result, our experimental results demonstrate that our method outperforms randomly generated interaction actions in both accuracy and efficiency. By comparing different uncertainty thresholds, we can find that as the uncertainty threshold decreases, the estimation accuracy of cluttered objects will rise, but at the same time the number of interacting actions will also increase.

2) Challenging clutters: To further verify the effectiveness of our method in complex scenes, we evaluated our method in 6 challenging scenes. These six scenes were generated in a similar manner as described above, although we narrowed down the generation range of random drop points and

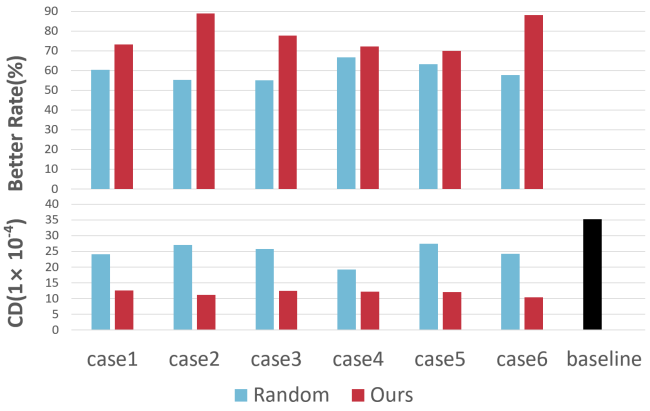

Fig. 9. Performance metrics for six challenging cases, where the black bars represent performance without interaction.

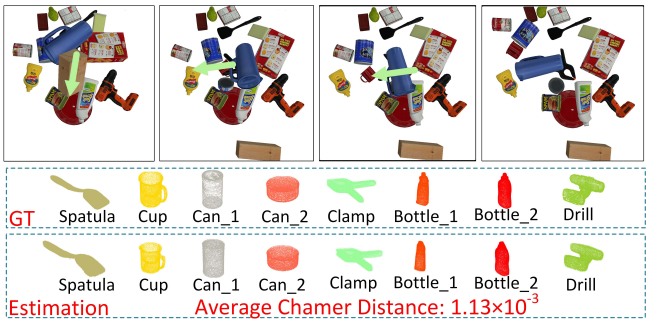

Fig. 10. Visualization of the interaction process and results, with each row representing a time series. The green arrow depicts the start point, direction and distance of the generated pushing actions.

exacerbated the mutual occlusion between objects to make the scenes more cluttered. Fig. 8 demonstrates each of these six scenes. As shown in Table II, We evaluated accuracy and motion efficiency on several methods. And Fig. 9 shows the various scenes for each scene in detail. Similar to the way of comparing the randomized scenes, our method reduces 57.9%, 65.0%, and 66.7% compared to Baseline. And the randomized interaction approach reduces 27.7%, 29.1%, and 30.5%.

Obviously, due to the lack of uncertainty perception of the scene objects, the random interaction strategy can only improve the estimation performance in a very limited way. Fig. 10 illustrates the interaction process of our framework, in which the cluttered structure is altered to obtain visual clues.

V. CONCLUSIONS

In this paper, we propose an interactive exploration framework for estimating densely cluttered objects. We sample from the implicit representation of shape to obtain the point cloud of object for shape estimation, and generate interaction for effective interactive exploration to reduce the estimation uncertainty for informative visual clue discovery. Therefore, our approach can effectively perceive objects in dense clutter and significantly improve estimation results. Extensive experiments have demonstrated the effectiveness of our framework, providing reliable object shape information for downstream tasks.

REFERENCES

[1] Z. Liu, Z. Wang, S. Huang, J. Zhou, and J. Lu, "Ge-grasp: Efficient target-oriented grasping in dense clutter," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 1388–1395.

[2] D. Son, "Grasping as inference: Reactive grasping in heavily cluttered environment," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7193–7200, 2022.

[3] X. Lou, Y. Yang, and C. Choi, "Learning object relations with graph neural networks for target-driven grasping in dense clutter," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 742–748.

[4] Y. Yang, Y. Liu, H. Liang, X. Lou, and C. Choi, "Attribute-based robotic grasping with one-grasp adaptation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6357–6363.

[5] T. Chen, A. Shenoy, A. Kolinko, S. Shah, and Y. Sun, "Multi-object grasping – estimating the number of objects in a robotic grasp," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021, pp. 4995–5001.

[6] Y. Yang, H. Liang, and C. Choi, "A deep learning approach to grasping the invisible," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2232–2239, 2020.

[7] J. Wu, H. Wu, S. Zhong, Q. Sun, and Y. Li, "Learning pre-grasp manipulation of flat objects in cluttered environments using sliding primitives," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 1800–1806.

[8] S. Huang, Z. Wang, J. Zhou, and J. Lu, "Planning irregular object packing via hierarchical reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 81–88, 2023.

[9] F. Wang and K. Hauser, "Dense robotic packing of irregular and novel 3d objects," *IEEE Transactions on Robotics*, vol. 38, no. 2, pp. 1160–1173, 2022.

[10] A. Yasuda, G. A. G. Ricardez, J. Takamatsu, and T. Ogasawara, "Packing planning and execution considering arrangement rules," in *2020 Fourth IEEE International Conference on Robotic Computing (IRC)*, 2020, pp. 100–106.

[11] F. Wang and K. Hauser, "Stable bin packing of non-convex 3d objects with a robot manipulator," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 8698–8704.

[12] S. B. Bayraktar, A. Orthey, Z. Kingston, M. Toussaint, and L. E. Kavraki, "Solving rearrangement puzzles using path defragmentation in factored state spaces," *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4529–4536, 2023.

[13] P. Miguel Uriguen Eljuri, L. El Hafi, G. Alfonso Garcia Ricardez, A. Taniguchi, and T. Taniguchi, "Neural network-based motion feasibility checker to validate instructions in rearrangement tasks before execution by robots," in *2022 IEEE/SICE International Symposium on System Integration (SII)*, 2022, pp. 1058–1063.

[14] C. Song and A. Boularias, "Object rearrangement with nested nonprehensile manipulation actions," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 6578–6585.

[15] K. Ren, P. Chanrungmaneekul, L. E. Kavraki, and K. Hang, "Kinodynamic rapidly-exploring random forest for rearrangement-based nonprehensile manipulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 8127–8133.

[16] W. Wang, Z. Zhao, Z. Jiao, Y. Zhu, S.-C. Zhu, and H. Liu, "Rearrange indoor scenes for human-robot co-activity," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 11 943–11 949.

[17] A. Rosasco, S. Berti, F. Bottarel, M. Colledanchise, and L. Natale, "Towards confidence-guided shape completion for robotic applications," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022, pp. 580–586.

[18] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation," 2020.

[19] M. Z. Irshad, T. Kollar, M. Laskey, K. Stone, and Z. Kira, "Centersnap: Single-shot multi-object 3d shape reconstruction and categorical 6d pose and size estimation," 2022. [Online]. Available: https://arxiv.org/abs/2203.01929

[20] 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 *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[21] M. Tian, M. H. A. J. au2, and G. H. Lee, "Shape prior deformation for categorical 6d object pose and size estimation," 2020.

[22] N. Heppert, M. Z. Irshad, S. Zakharov, K. Liu, R. A. Ambrus, J. Bohg, A. Valada, and T. Kollar, "Carto: Category and joint agnostic reconstruction of articulated objects," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 21 201–21 210.

[23] Z. Wu, Z. Wang, J. Lu, and H. Yan, "Category-level shape estimation for densely cluttered objects," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 3846–3852.

[24] X. Wen, P. Xiang, Z. Han, Y.-P. Cao, P. Wan, W. Zheng, and Y.-S. Liu, "Pmp-net++: Point cloud completion by transformer-enhanced multi-step point moving paths," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 1, pp. 852–867, 2023.

[25] X. Deng, J. Geng, T. Bretl, Y. Xiang, and D. Fox, "icaps: Iterative category-level object pose and shape estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1784–1791, 2022.

[26] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019.

[27] Z. Wu, Z. Wang, Z. Wei, Y. Wei, and H. Yan, "Smart explorer: Recognizing objects in dense clutter via interactive exploration," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 6600–6607.

[28] A. Eitel, N. Hauff, and W. Burgard, "Self-supervised transfer learning for instance segmentation through physical interaction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 4020–4026.

[29] K. Xu, H. Huang, Y. Shi, H. Li, P. Long, J. Caichen, W. Sun, and B. Chen, "Autoscanning for coupled scene reconstruction and proactive object analysis," *ACM TRANSACTIONS ON GRAPHICS*, vol. 34, no. 6, NOV 2015.

[30] R. Yoshino, T. Takano, H. Tanaka, and T. Taniguchi, "Active exploration for unsupervised object categorization based on multimodal hierarchical dirichlet process," in *2021 IEEE/SICE International Symposium on System Integration (SII)*, 2021, pp. 176–183.

[31] X. Ye and Y. Yang, "Efficient robotic object search via hiem: Hierarchical policy learning with intrinsic-extrinsic modeling," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4425–4432, 2021.

[32] A. Wolek, S. Cheng, D. Goswami, and D. A. Paley, "Cooperative mapping and target search over an unknown occupancy graph using mutual information," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1071–1078, 2020.

[33] C. Nam, S. H. Cheong, J. Lee, D. H. Kim, and C. Kim, "Fast and resilient manipulation planning for object retrieval in cluttered and confined environments," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1539–1552, 2021.

[34] T. Novkovic, R. Pautrat, F. Furrer, M. Breyer, R. Siegwart, and J. Nieto, "Object finding in cluttered scenes using interactive perception," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 8338–8344.

[35] J. Park, T. Yoon, J. Hong, Y. Yu, M. Pan, and S. Choi, "Zero-shot active visual search (zavis): Intelligent object search for robotic assistants," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 2004–2010.

[36] Q. Gao, F. Chang, L. Ma, and Q. Yi, "Active mapping method for agricultural robot using frontier-based exploration," in *2022 China Automation Congress (CAC)*, 2022, pp. 334–339.

[37] L. Liu, S. Fryc, L. Wu, T. L. Vu, G. Paul, and T. Vidal-Calleja, "Active and interactive mapping with dynamic gaussian process implicit surfaces for mobile manipulators," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3679–3686, 2021.

[38] T. Duong, M. Yip, and N. Atanasov, "Autonomous navigation in unknown environments with sparse bayesian kernel-based occupancy mapping," *IEEE Transactions on Robotics*, vol. 38, no. 6, pp. 3694–3712, 2022.

[39] A. Asgharivaskasi and N. Atanasov, "Active bayesian multi-class mapping from range and semantic segmentation observations," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 1–7.

[40] B. J. Julian, S. Karaman, and D. Rus, "On mutual information-based control of range sensing robots for mapping applications," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2013, pp. 5156–5163.

[41] E. Coumans, Y. P. Bai, and A. PyBullet, "a python module for physics simulation for games, robotics and machine learning. 2016."