Category-level Shape Estimation for Densely Cluttered Objects

Zhenyu Wu[1], Ziwei Wang[2], Jiwen Lu[2] and Haibin Yan[*1]

***Abstract*— Accurately estimating the shape of objects in dense clutters makes important contribution to robotic packing, because the optimal object arrangement requires the robot planner to acquire shape information of all existed objects. However, the objects for packing are usually piled in dense clutters with severe occlusion, and the object shape varies significantly across different instances for the same category. They respectively cause large object segmentation errors and inaccurate shape recovery on unseen instances, which both degrade the performance of shape estimation during deployment. In this paper, we propose a category-level shape estimation method for densely cluttered objects. Our framework partitions each object in the clutter via the multi-view visual information fusion to achieve high segmentation accuracy, and the instance shape is recovered by deforming the category templates with diverse geometric transformations to obtain strengthened generalization ability. Specifically, we first collect the multi-view RGB-D images of the object clutters for point cloud reconstruction. Then we fuse the feature maps representing the visual information of multi-view RGB images and the pixel affinity learned from the clutter point cloud, where the acquired instance segmentation masks of multi-view RGB images are projected to partition the clutter point cloud. Finally, the instance geometry information is obtained from the partially observed instance point cloud and the corresponding category template, and the deformation parameters regarding the template are predicted for shape estimation. Experiments in the simulated environment and real world show that our method achieves high shape estimation accuracy for densely cluttered everyday objects with various shapes.**

I. Introduction

Robotic packing systems [21], [1], [34], [41], [13], [14] play a key role in warehouse automation with the benefits of reduced uptime, high throughput, and low accident rate compared with the labor-intensive approaches. The goal of robotic packing is to stow objects into constrained space such as shipping boxes. In robotic packing systems, accurate shape estimation of objects in dense clutters is required because the planner has to obtain the shape information of all objects for packing in order to yield the optimal object arrangement in the packing boxes. For example, packing toys with different shapes leads to various placement locations and orientations, and wrong shape estimation of toys may cause packing failure due to object collision and space waste.

*Corresponding author.

[1]Zhenyu Wu and Haibin Yan are with the School of Automation, Beijing University of Posts and Telecommunications, Beijing, 100084, China. {wuzhenyu, eyanhaibin}@bupt.edu.cn

[2]Ziwei Wang and Jiwen Lu are with the 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, lujiwen@tsinghua.edu.cn

This work was supported in part by the National Natural Science Foundation of China under Grant 61976623 and Grant U22B2050.

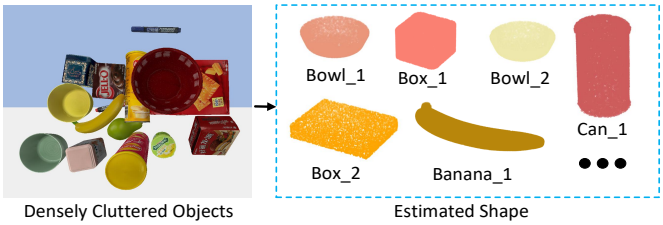

Fig. 1. An example of category-level shape estimation for densely cluttered objects.

Encoder-decoder architectures have been widely employed in other fields [31], [32], and have recently been utilized for object shape estimation. RGB images [25], occupancy voxels [33], depth maps [23] and SDF voxels [10] of objects are embedded into the latent space with object semantics, which is followed by shape reconstruction with the decoder. Since the object size varies across instances in the same category, the object size is predicted by encoding the geometric information with the pre-defined parametric models for fine-grained shape estimation [10], [18], [8], [11], [16]. However, conventional object shape estimation methods face two challenges. First, the objects for packing are usually piled in dense clutters, and the severe occlusion among objects fails to provide informative visual clues for shape recovery. Second, the shape varies significantly for different objects in the same category, and the inaccurate shape recovery on objects with novel appearance decreases the shape estimation precision in deployment.

In this paper, we present a category-level shape estimation method for densely cluttered objects. Our method segments each instance in the clutter by fusing the multi-view visual information, and recovers the object shape by deforming the category templates. Hence, high segmentation precision and high generalization ability are achieved to accurately estimate the shape of all existed objects. More specifically, we collect the multi-view RGB-D images of the clutter and reconstruct the point cloud of the scene, which are utilized as the visual input of the instance segmentation module. The feature maps representing visual information of multi-view RGB images and the pixel affinity learned from the clutter point cloud are fused to generate accurate instance segmentation masks for multi-view RGB images, which are projected to the point cloud in each view for label assignment. By merging the point cloud partitions in each view with similar spatial occupancy, we obtain the observed incomplete point cloud for each object in the clutter. The observed instance point cloud and the corresponding category template are jointly utilized to regress the template deformation parameters for scale and surface transformation. Fig. 1 demonstrates an example of

category-level shape estimation for densely cluttered objects, where the complete point cloud of each existed instance is predicted for the object arrangement planner in robotic packing. Extensive experiments in the simulated environment and real world indicate that our framework accurately recovers the point cloud of objects in dense clutters with diverse appearances.

II. Related Work

Visual segmentation in cluttered scenes: robotic manipulation tasks are usually challenging due to the severe occlusion in dense clutters, and object segmentation in cluttered scenes has aroused extensive interest in robotic visual perception. Existing visual segmentation for densely cluttered objects can be categorized into two types: segmentation based on RGB-D images [35], [37], [36] and point cloud [12], [38]. For the first regard, robotic grasping [28], [39], [24], [40], [7] was usually guided by a visual segmentation module for the planner to generate the optimal grasp pose. In order to segment the invisible objects in the clutter for accurate visual perception, Xie *et al.* [37] acquired initial rough masks according to depth images and then refined the predictions with RGB images. They further mined the relationship among objects via graph neural networks to generate more accurate instance mask refinement [36]. For visual segmentation methods based on the point cloud, Dong *et al.* [12] extracted the point-wise features with the constraint that embedding of points from the same instance should be similar and vice versa, so that the clustered index in the feature space could be leveraged as the segmentation masks. Xu *et al.* [38] inferred the geometric instance center via the learned point-wise features, and the remaining points were clustered into the closest center for segmentation. Nevertheless, the severe occlusion among objects cannot provide informative visual clues for accurate instance segmentation.

Object shape estimation: The goal of object shape estimation is to infer the 3D shape of objects given partial or sparse observations. Early attempts [5], [15] adopted surface reconstruction techniques via shape models to complete point clouds into dense surfaces. However, these methods can only model one object instance at a time with geometric priors, and the generalization ability to objects with different shapes is insufficient. Data-driven approaches [23], [33] for 3D shape estimation were presented which leveraged the encoder-decoder architecture to embed the object geometry and reconstruct the full shape sequentially. Rock *et al.* [23] retrieved similar objects from the database with deformation to recover the original shape. Moreover, simultaneously estimating object shape and pose [2], [22] can benefit each other due to their strong correlation. Since object shape varies significantly across different instances in the same category, category-level shape estimation [8], [18], [6], [30], [11], [17] generates the prediction with the category priors to enhance the generalization ability on unseen objects in deployment. Wang *et al.* [30] learned the canonical shape representation in the normalized object coordinate space to regress the object size regarding the category priors. However, existing methods fail to accurately recover the shape of unseen objects due to the large intra-class variation.

III. Approach

In this section, we first briefly introduce the problem of shape estimation for densely cluttered objects and the overall pipeline, and then detail the instance segmentation of object clutters by fusing the information from multi-view RGB images and point clouds. Finally, we present the category-level shape estimation for partially observed instances via template deformation.

A. Problem Statement and Overall Pipeline

The goal of shape estimation for densely cluttered objects is to predict the point cloud of every existed object in the clutter given the category template, so that the robotic packing system can yield the optimal object arrangement plan with object information. The challenges of achieving precise shape estimation are two-fold. First, the occlusion among cluttered objects disables the visual perception module to accurately recognize the object categories and segment each instance for sequential shape estimation. Second, the object shape varies significantly for instances in the same category, and objects with different shapes in deployment require a high generalization ability of the shape estimation module. To address these, we fuse the information from multi-view RGB images and clutter point cloud by passing the pixel affinity for instance segmentation, and deform the category template with diverse geometric transformation for generalizable shape estimation.

Fig. 2 demonstrates the overall pipeline of our framework. The object clutters are observed by one overhead and four side-view RGB-D cameras, and the side-view cameras are uniformly placed in a horizontal plane. The point cloud of the object clutters is obtained by projecting that converted from the depth image of all cameras, which is combined with the multi-view RBG images to function as the input of our framework. The pixel affinity learned from the clutter point cloud via SoftGroup [27] is fused into the predicted feature maps of multi-view RGB images acquired via Yolact [3], which assigns the instance labels for the point cloud projected inside the mask of each view. The pixel affinity generation process named SSC is proposed by [42]. By merging the point cloud partition across views with similar spatial occupancy, the observed point cloud for each instance is obtained. The instance-wise point cloud partition and the corresponding category template are leveraged to regress the geometric transformation parameters, where the box-cage based deformation is applied for shape estimation.

B. Instance Segmentation of object clutters

Predicting the instance-wise mask of densely cluttered objects makes significant contribution to shape estimation, because categories for different objects and instance-wise point clouds are utilized to regress the geometric transformation parameters regarding category templates. Instead of directly segmenting the clutter point cloud, we employ the

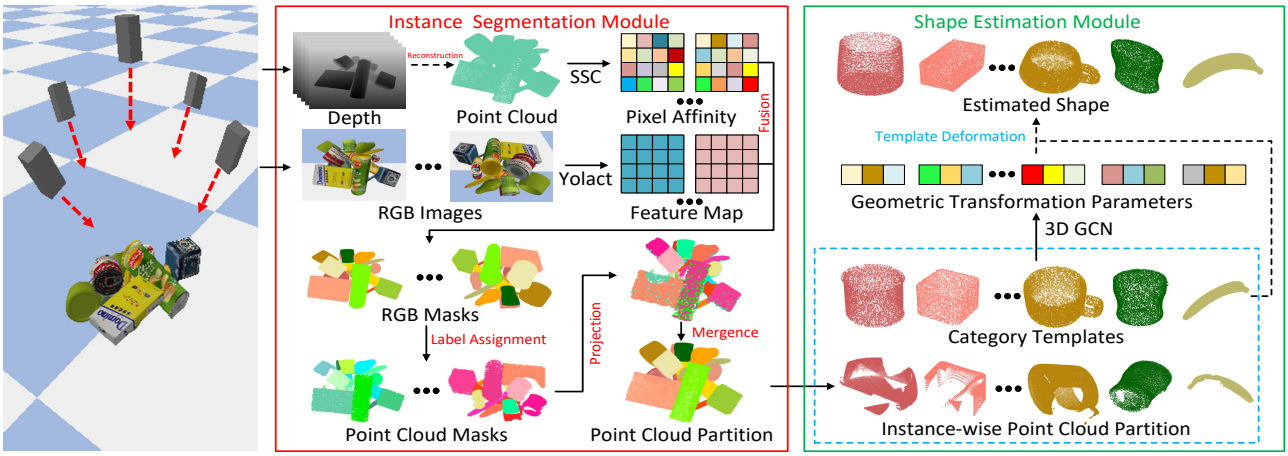

Fig. 2. The overall pipeline of our framework, which consists of the instance segmentation module and the shape estimation module.

instance masks of RGB images across views to assign labels to the point cloud inside the masks, as texture information significantly enhances segmentation masks for cluttered objects compared with geometry information. Since instance masks of RGB images across views may represent the same object, we should verify the object consistency across views based on predicted categories and spatial relationships to avoid false positives and negatives. The point cloud partitions from different views that share the same semantic labels and similar spatial occupancy are iteratively merged to yield the instance segmentation mask of the clutter point cloud. Let us denote the i_{th} instance mask of the point cloud at the t_{th} merging time as $\mathscr{P}_i^t$, the instance mask of the point cloud is updated as follows:

$$\mathscr{P}_i^{t+1} = \mathscr{P}_i^t \cup \{\boldsymbol{S}_m^k | c_m^k = C_i^0, d_{ch}(\boldsymbol{S}_m^k, \mathscr{P}_i^t) < h\}, \quad (1)$$

where $\boldsymbol{S}_m^k$ represents the m_{th} point cloud partition in the k_{th} view, and c_m^k and C_i^0 respectively mean the label of $\boldsymbol{S}_m^k$ and $\mathscr{P}_i^t$ respectively. $d_{ch}(\boldsymbol{x},\boldsymbol{y})$ stands for the chamfer distance between point cloud $\boldsymbol{x}$ and $\boldsymbol{y}$, and h is the threshold where point clouds with chamfer distance less than h are regarded to have similar spatial occupancy. Each point cloud partition is regarded as an instance at the initialization of merging, and the mergence stops to generate the instance segmentation masks for the clutter point cloud until no point cloud partition is enlarged.

Accurate instance segmentation of RGB images is crucial to precisely acquire the observed point cloud of each object for shape estimation. Severe occlusion among objects usually leads to ambiguous masks border of RGB images with incorrect predictions. Rather than directly predicting the masks of the multi-view RGB images, we fuse the pixel affinity learned via the clutter point cloud with the feature maps of RGB images to generate precise instance-wise masks for RGB images. The pixel affinity demonstrates the instance consistency among pixels, where the element in the i_{th} row and j_{th} column is set to one if the i_{th} and j_{th} pixels represent the same object and vice versa. Inspired by [42], we generate the pixel affinity $\boldsymbol{A}^k$ of the RGB image in the k_{th} view based on the point cloud feature of the object clutters. For the k_{th} view, the intra-class feature that fuses the information of pixels within each category and the inter-class feature which considers the visual clues from other categories are defined as follows:

$$\boldsymbol{Y}_{intra}^k = \boldsymbol{A}^k \mathscr{R}(\boldsymbol{X}_{2D}^k), \quad \boldsymbol{Y}_{inter}^k = (\boldsymbol{1} - \boldsymbol{A}^k)\mathscr{R}(\boldsymbol{X}_{2D}^k), \quad (2)$$

where $\mathscr{R}(\boldsymbol{X})$ means reshaping the spatial dimensions of $\boldsymbol{X}$ to match the matrix multiplication, and $\boldsymbol{1}$ is an all-one matrix with the same size as $\boldsymbol{A}^k$. Meanwhile, $\boldsymbol{X}_{2D}^k$ stands for the RGB image feature for the k_{th} view. Finally, we concatenate the RGB feature, the intra-class feature, and the inter-class feature to aggregate the information for accurate instance segmentation of densely cluttered objects. Denoting the element in the i_{th} row and j_{th} column of $\boldsymbol{A}^k$ as a_{ij}^k, we aim to minimize the difference between the predicted pixel affinity matrix and groundtruth via the binary cross-entropy:

$$L_{ce} = -\frac{1}{KN^2}\sum_{k=1}^{K}\sum_{i,j=1}^{N} c_{ij}^k \log a_{ij}^k + (1 - c_{ij}^k)\log(1 - a_{ij}^k), \quad (3)$$

where $c_{ij}^k \in \{0,1\}$ is the groundtruth label of a_{ij}^k, and K is the number of views for visual information collection. In order to learn the correct semantic correlation for pixel affinity, we optimize the precision L_p and recall L_r that reveal the performance of intra-class features, and maximize the specificity L_s that depicts the inter-class feature quality:

$$L_p = \frac{1}{K}\sum_{k=1}^{K}\log\frac{\sum_{i,j=1}^{N} c_{ij}^k a_{ij}^k}{\sum_{i,j=1}^{N} a_{ij}^k}, \quad L_r = \frac{1}{K}\sum_{k=1}^{K}\log\frac{\sum_{i,j=1}^{N} c_{ij}^k a_{ij}^k}{\sum_{i,j=1}^{N} c_{ij}^k},$$
$$L_s = \frac{1}{K}\sum_{k=1}^{K}\log\frac{\sum_{i,j=1}^{N}(1 - c_{ij}^k)(1 - a_{ij}^k)}{\sum_{i,j=1}^{N}(1 - c_{ij}^k)}, \quad (4)$$

The overall learning objective for instance segmentation considers the isolated pixel affinity correctness by binary cross-entropy and the global affinity correctness indicated by precision, recall, and specificity via the following form:

$$L_{seg} = L_{ce} - \lambda(L_p + L_r + L_s), \quad (5)$$

where λ is a hyperparameter that controls the importance of global correctness in the predicted pixel affinity. By fusing the multi-view visual clues of the object clutters,

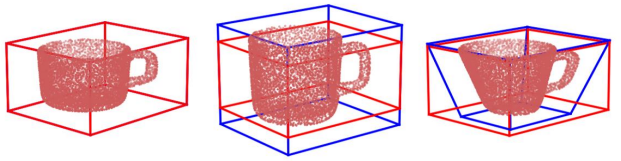

Fig. 3. Examples of scale and surface transformation for category-level templates.

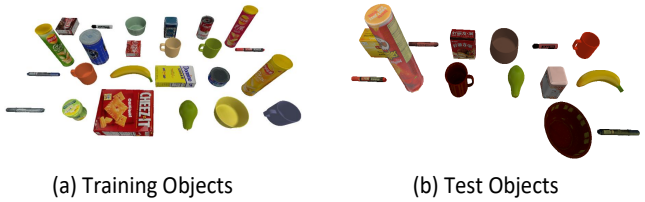

Fig. 4. Selected objects from YCB and OCRTOC datasets for training and test in our experiments.

we strengthen the instance segmentation accuracy of RGB images and assign precise labels for instance-wise point cloud partition for subsequent shape recovery.

C. Category-level Shape Estimation of Cluttered Objects

Acquiring the shape of each object existing in the clutter is necessary for the object arrangement planner in robotic packing systems. The point cloud partition for each instance provides visual clues with partial observation and estimating the shape of each object is equivalent to recovering the complete point cloud. Since objects from the same category share similar geometric structures [26], we apply a box-cage based template deformation method to enhance the generalization ability of the shape estimation module on intra-class variation inspired by Fs-Net [8]. Fig. 3 shows an example of box-cage based deformation techniques including scale and surface transformation. The predicted shape can be obtained by modifying the vertices of the template in the following:

$$\boldsymbol{V}_d = \mathbb{F}_{sur} \circ \mathbb{F}_{sca}(\boldsymbol{V}_0), \tag{6}$$

where $\boldsymbol{V}_d$ and $\boldsymbol{V}_0$ respectively represent vertices of objects after and before deformation. Denoting the i_{th} vertex in object vertices $\boldsymbol{V}$ as $\boldsymbol{V}^i$, the scale transformation function is defined as follows:

$$\mathbb{F}_{sca}(\boldsymbol{V}^i) = [\alpha_x V_x^i, \alpha_y V_y^i, \alpha_z V_z^i], \tag{7}$$

where the V_x^i, V_y^i, V_z^i indicates the component in the x, y, z axis for the vertex $\boldsymbol{V}^i$, and α_x, α_y and α_z are the scaling factors for object size adjustment. The surface transformation function changes the area of the top and bottom surfaces for the box-cage in order to achieve diverse shape variations of symmetrical categories:

$$\mathbb{F}_{sur}(\boldsymbol{V}^i) = [V_x^i + \frac{\varepsilon(V_z^i - V_z^{\downarrow})}{V_z^{\uparrow} - V_z^{\downarrow}} V_x^i, V_y^i + \frac{\varepsilon(V_z^i - V_z^{\downarrow})}{V_z^{\uparrow} - V_z^{\downarrow}} V_y^i, V_z^i], \tag{8}$$

where $V_z^{\downarrow}$ and $V_z^{\uparrow}$ demonstrate the vertical coordinates of vertices with minimal and maximal z value, and ε is the surface factor to change the ratio of top and bottom surface area for the box-cage. We construct the mesh of deformed objects with the adjusted vertices and original triangles from the template, and uniformly sample the object point cloud from the mesh for shape recovery. To regress scaling factors and surface factors in template deformation, we leverage 3D graph convolutional networks [19] to parameterize them for each object, where instance-wise point cloud partition and vertices of corresponding category templates are employed as the input. The objective to train the shape estimation module is to minimize the Chamfer distance between the predicted point cloud and the groundtruth. By learning the correspondence pattern between the partial observation of objects and the category template, the shape estimation module reconstructs the complete point cloud for densely cluttered objects with diverse appearances.

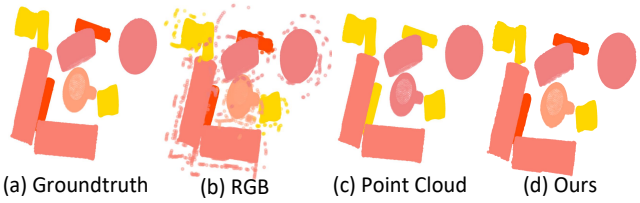

Fig. 5. The visualization of instance segmentation results for (a) groundtruth, (b) RGB-only (c) point cloud-only and (d) our methods, where partitions in different colors represent various classes.

IV. Experiments

In this section, we conduct extensive experiments in simulated environments (Pybullet [9]) and the real world to evaluate our framework. The goal of the experiment is to verify that (1) our shape estimation framework for densely cluttered objects can accurately generate complete point clouds of all existed objects, (2) the multi-view visual information fusion via pixel affinity passing significantly enhance the instance segmentation performance for object clutters, (3) deforming the category template with diverse geometric transformation according to predicted parameters strengthens the generalization ability.

A. Implementation Details and Evaluation Metrics

All objects utilized in our experiments come from the YCB dataset [4] and OCRTOC dataset [20]. We only select a subset of 24 objects for training and 14 objects for testing to construct our scenes including some generic objects such as boxes, cans, markers, sugar, bananas, pears, mugs, and bowls, where their fine-grained category names are replaced by coarse class names in category-level shape estimation. Fig. 4 visualizes the selected objects in our experiments, where most objects in the test scenarios do not appear in the training scenes. We employ the mean shape of all training objects in each category as the template. For simulated experiments, we deform the template with random parameters in scale and surface transformation for object generation to diversify the shape of instances in training and test set. We collected 1,750 and 350 RGB images from different views with pixel annotation and the corresponding clutter point cloud as the training and test set for instance segmentation

TABLE I

THE MAP AND AP WITH DIFFERENT IOU THRESHOLDS OF POINT CLOUD INSTANCE SEGMENTATION, WHERE RANDOM, EASY, NORMAL AND HARD CASES ARE LEVERAGED FOR EVALUATION.

Methods	Random			Easy			Normal			Hard		
	mAP	AP_{25}	AP_{50}	mAP	AP_{25}	AP_{50}	mAP	AP_{25}	AP_{50}	mAP	AP_{25}	AP_{50}
RGB-only [3]	34.16	58.52	51.23	42.63	68.70	58.03	33.17	58.42	49.21	24.19	52.11	43.75
Point-only [27]	32.10	64.50	45.90	32.50	65.40	47.30	30.70	62.40	44.70	27.90	54.40	39.00
Ours	38.75	78.43	58.11	49.18	85.43	71.33	38.53	79.03	58.22	28.15	74.18	49.76

TABLE II

THE CD BETWEEN THE PREDICTED SHAPES AND THE GROUNDTRUTH FOR GIVEN INSTANCE-WISE POINT CLOUD PARTITIONS.

Deformation	Random	Easy	Normal	Hard
None	71.12	39.71	62.17	147.77
Scale-only	67.25	38.00	58.84	139.02
Surface-only	69.39	38.82	60.55	143.54
Ours	63.66	36.27	55.67	131.46

module, and constructed 400 and 30 scenes which include 5-15 objects for training and testing respectively. Moreover, we prepared 24 scenes containing 5-15 objects for evaluation in real-world experiments.

Since our framework consists of the instance segmentation module and the shape estimation module, we respectively present three metrics to evaluate the above two individual modules and the overall performance on category-level shape estimation for densely cluttered objects. For instance segmentation, we leverage the mean average precision (mAP) of the point cloud masks with the IoU$\in$[0.5:0.05:0.95]. To assess the shape estimation, we utilize the Chamfer distance (CD) between the predicted and the groundtruth shape for true positive segmentation predictions. The instance segmentation module influences the precision and recall of the segmentation masks, and the shape estimation module affects the bounding box IoU between the predicted shape and the groundtruth. To measure the overall performance of our framework, we reconstruct the clutter point cloud by placing the estimated object shape with known poses and report the precision and recall of the reconstructed point cloud with various bounding box IoU thresholds. Moreover, we also provide the F1 score of the mean average precision and recall for reference.

B. Simulated Experiments

We first demonstrate the performance of the instance segmentation module with different visual information perception methods. Then we evaluate the shape estimation module with given instance-wise point clouds across various template deformations. Finally, we depict the overall performance of category-level shape estimation for densely cluttered objects.

Results on instance segmentation: The random clutter is constructed by dropping objects into the workspace, where the landing point for each object is selected randomly. Since the difficulty of instance segmentation is positively related to the object density, we set up the object clutters with 5, 10 and 15 objects for easy, normal and hard scenarios of instance segmentation. Table I demonstrates the mAP of the instance segmentation masks of point clouds, where the baseline methods contain instance segmentation only based on multi-view RGB images [3] and clutter point cloud [27]. Compared with the method that directly segments the point cloud, we increase the mAP by 6.65% in random cases because the texture information significantly enhances the segmentation masks for cluttered objects. Meanwhile, our framework also outperforms the baseline, which removes the pixel affinity learned from the clutter point cloud by 4.56% in hard cases, because fusing the scene information via the clutter point cloud alleviates the segmentation errors caused by occlusion. Fig. 5 visualizes the instance segmentation masks of clutter point clouds for different methods. Only leveraging the RGB images for segmentation fails to generate accurate pixel-wise masks, and methods only utilizing point cloud cannot assign the precise label to each partition.

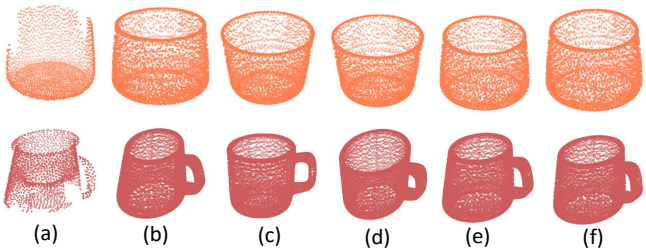

Fig. 6. The visualization of (a) the observed instance point cloud for the shape estimation module, (b) the groudtruth shape, (c) the predicted shape without template deformation, (d) with only scale transformation, (e) with only surface deformation and (f) with our framework.

Results on shape estimation: The baselines for comparison include utilizing the template as the predicted shape without deformation, with only scale transformation, and with only surface transformation. We apply the point cloud partition for each object acquired by our instance segmentation module as the input of the shape estimation module. Table II demonstrates the Chamfer distance (CD) of different shape estimation methods, and Fig. 6 visualizes several examples of recovered shapes given the fixed partial observation of object point cloud. Our framework significantly decreases the CD compared with the baseline methods, which verifies the effectiveness of diverse geometric information in shape recovery including scale and surface transformation in shape estimation of novel objects.

Results on shape estimation for densely cluttered objects: By integrating the instance segmentation and the shape estimation modules, we obtain the overall performance on shape estimation for densely cluttered objects. Table III illustrates the averaged precision and recall of the reconstructed clutter point cloud with different IoU thresholds and the mean ones with IoU from 0.1 to 0.55, where the F1 score of the mean average precision and recall is also provided for reference. For the chosen baseline methods, we only

Fig. 7. An example of the estimated shape for densely cluttered objects via different methods including RGB image based and point cloud based instance segmentation with our template deformation, surface-only and scale-only template deformation with our instance segmentation. The estimation results of our method and the groundtruth are also demonstrated.

TABLE III

COMPARISON OF THE PRECISION AND RECALL WITH DIFFERENT IOU THRESHOLDS FOR THE SHAPE ESTIMATION OF RANDOM CASES. THE F1 SCORE OF THE MEAN AVERAGE PRECISION AND RECALL IS PROVIDED FOR REFERENCE.

Methods		Precision				Recall				F1 Score
Segmentation	Estimation	mAP	AP_{10}	AP_{25}	AP_{50}	mAR	AR_{10}	AR_{25}	AR_{50}	
RGB	None	30.25	50.15	45.84	10.57	36.38	46.97	45.52	13.15	33.03
RGB	Scale	43.33	49.48	50.05	46.58	51.60	47.59	46.95	45.35	47.10
RGB	Surface	40.18	49.53	47.81	44.57	48.51	49.40	47.29	46.37	43.95
Point	None	27.28	42.84	34.65	7.89	23.82	38.60	31.28	6.83	25.43
Point	Scale	41.83	57.10	45.10	29.58	40.10	50.93	40.89	29.24	40.95
Point	Surface	35.79	49.77	39.39	21.70	33.02	44.97	34.93	20.32	34.35
Fsnet [8]		43.73	62.87	53.31	43.95	47.83	66.77	52.16	45.48	45.69
Densefusion [29]		48.32	67.19	56.91	49.87	44.77	63.35	52.71	42.90	46.48
Ours		55.94	65.32	57.96	54.13	61.08	64.97	58.45	56.99	58.40

TABLE IV

COMPARISON ON THE PRECISION AND RECALL WITH DIFFERENT IOU THRESHOLDS FOR SHAPE ESTIMATION OF REAL-WORLD EXPERIMENT.

Methods	Precision		Recall		F1 Score
	mAP	AP_{25}	mAR	AR_{25}	
RGB-only	46.26	55.93	45.63	51.26	45.94
Point-only	38.43	40.61	42.48	49.97	40.35
Scale-only	50.43	56.15	41.43	46.50	45.49
Surface-only	40.14	49.35	36.11	40.58	38.02
Ours	54.26	59.82	50.15	53.22	52.12

employed the final predicted bounding boxes. Our framework outperforms the baseline methods that combines different instance segmentation and shape estimation techniques by a sizable margin, which reveals that both instance segmentation and shape estimation are necessary to achieve practical category-level shape estimation for objects in dense clutters. Our method also achieves higher precision and recall than the state-of-the-art methods in shape estimation because of the accurate partial observation of objects and diverse template deformation.

C. Real-world Experiments

Fig. 7 shows several quantitative examples for estimating the shape of all objects in the dense clutters. RGB-only and point-only represent the methods only leveraging the RGB images and point cloud for instance segmentation following our template deformation techniques. Surface-only and scale-only depict the approaches that utilize our instance segmentation module following the surface and scale transformation for template deformation respectively. Compared with the RGB-only and point-only methods, our framework accurately segments each instance without missing objects because of global information fusion. The surface-only and scale-only methods cannot precisely estimate the shape of each instance due to the limited geometric transformation of templates. For example, the instant noodle bucket in Fig. 7 uses the same bucket template as the potato chip bucket applies in template deformation, and our framework can effectively estimate the shape of the instant noodle bucket with the help of surface transformation even their shapes differ obviously. Table IV illustrates the average precision and recall with different IoU thresholds of our framework in real-world experiments. We also provide the F1 score of the mean average precision and recall for reference. Our framework outperforms baseline methods, which further verifies the effectiveness of both our instance segmentation and shape estimation modules in practical scenarios. The difficulties in the simulated and real-world scenarios are similar due to the same object number setting. The mAP acquired in the real-world experiment is only 1.68% lower than that in the simulated environment, which reveals the high generalization ability of our method to real-world cluttered object shape estimation.

V. CONCLUSION AND FUTURE WORK

In this paper, we have presented a category-level shape estimation framework for densely cluttered objects. We collect the multi-view RGB-D images of the object clutters and reconstruct the point cloud of the whole scene for visual clue representation. The feature maps of multi-view RGB images and the pixel-wise similarity learned from the clutter point cloud are fused via affinity passing for accurate instance segmentation of RGB images, which assigns correct labels for point clouds of each view to acquire the instance point cloud with mergence. The correspondence pattern between the instance-wise point cloud partition and the category template is extracted to predict the parameters of geometric transformation regarding templates for shape estimation. Extensive experiments in the simulated environment and real world demonstrate the effectiveness of the proposed method. In future work, we plan to reduce the computational and storage complexity of pixel affinity prediction and diversify the geometric transformation for template deformation.

REFERENCES

[1] M. Agarwal, S. Biswas, C. Sarkar, S. Paul, and H. S. Paul. Jampacker: An efficient and reliable robotic bin packing system for cuboid objects. *IEEE Robotics and Automation Letters*, 6(2):319–326, 2020.

[2] A. Avetisyan, A. Dai, and M. Nießner. End-to-end cad model retrieval and 9dof alignment in 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2551–2560, 2019.

[3] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9157–9166, 2019.

[4] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv preprint arXiv:1502.03143*, 2015.

[5] J. C. Carr, R. K. Beatson, J. B. Cherrie, T. J. Mitchell, W. R. Fright, B. C. McCallum, and T. R. Evans. Reconstruction and representation of 3d objects with radial basis functions. In *Proceedings of the 28th annual Conference on Computer Graphics and Interactive Techniques*, pages 67–76, 2001.

[6] 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*, pages 11973–11982, 2020.

[7] 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)*, pages 4995–5001, 2021.

[8] 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*, pages 1581–1590, 2021.

[9] E. Coumans and Y. Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016.

[10] A. Dai, C. Ruizhongtai Qi, and M. Nießner. Shape completion using 3d-encoder-predictor cnns and shape synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5868–5877, 2017.

[11] 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*, 2022.

[12] Z. Dong, S. Liu, T. Zhou, H. Cheng, L. Zeng, X. Yu, and H. Liu. Ppr-net:point-wise pose regression network for instance segmentation and 6d pose estimation in bin-picking scenarios. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1773–1780, 2019.

[13] M. Gualtieri and R. Platt. Robotic pick-and-place with uncertain object instance segmentation and shape completion. *IEEE Robotics and Automation Letters*, 6(2):1753–1760, 2021.

[14] S. Huang, Z. Wang, J. Zhou, and J. Lu. Planning irregular object packing via hierarchical reinforcement learning. *IEEE Robotics and Automation Letters*, 8(1):81–88, 2022.

[15] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Transactions on Graphics (ToG)*, 32(3):1–13, 2013.

[16] T. Lee, B.-U. Lee, M. Kim, and I. S. Kweon. Category-level metric scale object shape and pose estimation. *IEEE Robotics and Automation Letters*, 6(4):8575–8582, 2021.

[17] B. Liang, W. Liang, and Y. Wu. Parameterized particle filtering for tactile-based simultaneous pose and shape estimation. *IEEE Robotics and Automation Letters*, 7(2):1270–1277, 2021.

[18] 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*, pages 3560–3569, 2021.

[19] Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang. Convolution in the cloud: Learning deformable kernels in 3d graph convolution networks for point cloud analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1800–1809, 2020.

[20] Z. Liu, W. Liu, Y. Qin, F. Xiang, M. Gou, S. Xin, M. A. Roa, B. Calli, H. Su, Y. Sun, et al. Ocrtoc: A cloud-based competition and benchmark for robotic grasping and manipulation. *IEEE Robotics and Automation Letters*, 7(1):486–493, 2021.

[21] 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)*, pages 1388–1395, 2022.

[22] F. Manhardt, W. Kehl, and A. Gaidon. Roi-10d: Monocular lifting of 2d detection to 6d pose and metric shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2069–2078, 2019.

[23] J. Rock, T. Gupta, J. Thorsen, J. Gwak, D. Shin, and D. Hoiem. Completing 3d object shape from one depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2484–2493, 2015.

[24] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke. Rgb-d object detection and semantic segmentation for autonomous manipulation in clutter. *The International Journal of Robotics Research*, 37(4-5):437–451, 2018.

[25] D. Stutz and A. Geiger. Learning 3d shape completion from laser scan data with weak supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1955–1964, 2018.

[26] H. A. Vlach. How we categorize objects is related to how we remember them: The shape bias as a memory bias. *Journal of Experimental Child Psychology*, 152:12–30, 2016.

[27] T. Vu, K. Kim, T. M. Luu, T. Nguyen, and C. D. Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022.

[28] K. Wada, K. Okada, and M. Inaba. Joint learning of instance and semantic segmentation for robotic pick-and-place with heavy occlusions in clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 9558–9564, 2019.

[29] F. Wang and K. Hauser. Dense robotic packing of irregular and novel 3-d objects. *IEEE Transactions on Robotics*, 2021.

[30] 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*, pages 2642–2651, 2019.

[31] Z. Wang, J. Lu, and J. Zhou. Learning channel-wise interactions for binary convolutional neural networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3432–3445, 2020.

[32] Z. Wang, H. Xiao, Y. Duan, J. Zhou, and J. Lu. Learning deep binary descriptors via bitwise interaction mining. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1919–1933, 2023.

[33] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1912–1920, 2015.

[34] 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)*, pages 6600–6607, 2022.

[35] Y. Xiang, C. Xie, A. Mousavian, and D. Fox. Learning rgb-d feature embeddings for unseen object instance segmentation. *arXiv preprint arXiv:2007.15157*, 2020.

[36] C. Xie, A. Mousavian, Y. Xiang, and D. Fox. Rice: Refining instance masks in cluttered environments with graph neural networks. In *Conference on Robot Learning*, pages 1655–1665, 2022.

[37] C. Xie, Y. Xiang, A. Mousavian, and D. Fox. he best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation. In *Conference on Robot Learning*, pages 1369–1378, 2020.

[38] Y. Xu, S. Arai, D. Liu, F. Lin, and K. Kosuge. Fpcc: Fast point cloud clustering based instance segmentation for industrial bin-picking. *Neurocomputing*, 494:255–268, 2022.

[39] Y. Yang, H. Liang, and C. Choi. A deep learning approach to grasping the invisible. *IEEE Robotics and Automation Letters*, 5(2):2232–2239, 2020.

[40] 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)*, pages 6357–6363, 2021.

[41] Z. Yang, S. Yang, S. Song, W. Zhang, R. Song, J. Cheng, and Y. Li. Packerbot: Variable-sized product packing with heuristic deep reinforcement learning. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5002–5008, 2021.

[42] C. Yu, J. Wang, C. Gao, G. Yu, C. Shen, and N. Sang. Context prior for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12416–12425, 2020.