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

On Improving Boundary Quality of Instance Segmentation in Cluttered and Chaotic Scenarios

Biqi Yang, Xiaojie Gao, Xianzhi Li✉, Yun-Hui Liu, Chi-Wing Fu, and Pheng-Ann Heng

***Abstract*— Instance segmentation is a long-standing task for supporting robotic bin picking. However, objects of diverse classes can be closely packed with occlusions in cluttered and chaotic scenes, hence, even recent methods could have difficulty in locating clear and precise boundaries to distinguish nearby objects. In this work, we aim to improve the boundary quality of the instance masks for robust and precise instance segmentation in these challenging scenarios. Technical-wise, we first formulate an IoU-based Boundary-aware Mask head (IBM head) for predicting the instance-level mask, boundary, and their corresponding IoU scores. With this core module, we then follow the coarse-to-fine strategy and design our pipeline with two stages: an IoUNet to learn localization-based objectness cue and a hierarchical mask refiner to produce sharper and cleaner boundaries. We deploy the IBM head throughout the framework. Extensive experimental results on three grasping benchmarks manifest that our method attains the best instance segmentation performance, compared with the state-of-the-art approaches. Practically, we conduct real-world picking tests to show that with the objectness and boundary IoU scores as guidance, we are able to filter invalid (occluded) instances and select high-fidelity (exposed) instances for grasping.**

I. INTRODUCTION

For robotics grasping, instance segmentation is a fundamental task to recognize and help locate object instances. In cluttered and chaotic scenarios, where objects of diverse categories are randomly stacked with heavy occlusion, we need high-quality instance masks with **precise boundaries** to distinguish nearby instances for accurate grasping.

With recent progress in deep learning, many instance segmentation solutions are developed to aid robotics grasping [2], [3], [4], [5]. However, these methods do not pay much attention to boundaries. We verify the importance of boundary for challenging environments. Fig. 1 shows the instance segmentation results on some GraspNet [6] scenarios. For the recent method Transfiner [1], basic instance contours can be described when objects are isolated (pointed by a green arrow), but the boundary precision largely drops when objects stack together; as pointed by two red arrows. Without boundary-awareness, these low-quality object boundaries and inaccurate masks containing the connection curve between two close objects can lead to grasping failure, if the robot gripper targets this connection edge.

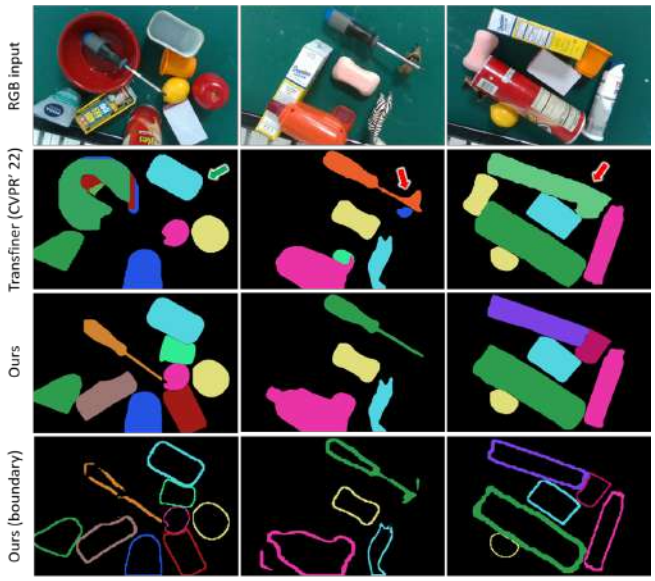

Fig. 1. Comparing results from the recent Transfiner [1] and our method. Clearly, our method can more precisely segment closely-packed objects.

Learning instance-level boundary matters from two aspects. First, for cluttered and chaotic scenes, boundary information can provide a strong guidance to learn the spatial relationships between instances, helping to produce more accurate instance-level masks. Second, compared to masks, boundaries can better reveal the overall object structure. Hence, an estimated boundary with high visibility to the camera means less occlusion, thus helping to indicate whether the associated object instance is easy to pick.

With these motivations, in this work, we propose a new instance segmentation approach that specifically learns and attends to use the instance boundaries for predicting more accurate instance masks in cluttered and chaotic scenes. Our key contributions include a novel IoU-based Boundary-aware Mask head (abbr. IBM head), which predicts instance-level mask, boundary, and their associated IoU scores. Based on this mask head, we further design a segment-then-refine pipeline. In the first stage, coarse instance-level masks and boundaries are produced by an IoU-based instance segmentation network, named IoUNet. We deploy it with IoU awareness for both detection and segmentation. In the second stage, we use a hierarchical mask refiner aided by a parallel semantic segmentation branch. We gradually upsample and

This work is supported by InnoHK of the Government of Hong Kong via the Hong Kong Centre for Logistics Robotics. This work is partially supported by the following grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project Reference Number: 14201321 and 14201620). This work is supported by the China National Natural Science Foundation No. 62202182.

B. Yang, X. Gao, C.-W. Fu and P.-A. Heng are with the Department of Computer Science and Engineering, The Chinese University of Hong Kong. P.-A. Heng is also with Guangdong-Hong Kong-Macao Joint Laboratory of Human-Machine Intelligence-Synergy Systems, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. X. Li is with the School of Computer Science and Technology, Huazhong University of Science and Technology. Y.-H. Liu is with the Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong.

Corresponding author: Xianzhi Li (xzli@hust.edu.cn)

refine the boundary-aware masks with more details.

Extensive experiments show that our method can better recognize the boundary contours and separate closely-packed instances with significant segmentation improvement over recent methods on various grasping environments; see Sec. IV-B to Sec. IV-E. We practically verify the importance of the IoU scores to guide object grasping in Sec. IV-F.

Our contributions can be summarized as follows:

- We state that boundary awareness is critical for accurate segmentation on heavily-occluded grasping scenarios.
- We propose an IoU-based Boundary-aware Mask head, which helps our method gradually sharpen the boundary with precise masks in a coarse-to-fine manner.
- Extensive experiments on three benchmarks demonstrate the strong capability of our method in segmenting objects with high boundary precision, compared with the state-of-the-art approaches.
- We conduct robotic demonstrations on bin-picking tasks to show the practical effectiveness of our method. Our predicted IoU scores provide strong guidance to select good instances with high visibility for grasping.

II. Related Work

A. Instance segmentation for robotics grasping

Existing instance segmentation works [7], [8], [9], [10] for robotic mainly focus on improving robustness in cluttered scenes. A direct and simple solution [11], [12], [5] is to leverage both RGB and depth then fuse them together towards accurate masks. Wada et al. [3] jointly train semantic and instance segmentation for better performance, and Schwarz et al. [2] design a pipeline specifically for autonomous robotic manipulation in cluttered scenes. Towards the cases with severe occlusion, [13], [14] model the visible and occluded regions separately. Another challenging topic is to segment unseen object [15], [16] to generalize the method for the open world. Danielczuk et al. [17] fed depth images into Mask R-CNN [18] for bin picking. Xie et al. [4], [19] first use depth for initial segmentation and then refine the coarse masks with RGB. Further studies included segmentation [14] and test-time domain adaptation [20].

Meanwhile, various datasets are released to facilitate studies on robotic instance segmentation, including both real-world annotated data and synthetic training data produced by simulation. Schwarz et al. [2] compile WISDOM with both synthetic and real data, Suchi et al. [21] create a semi-automatic annotation tool for RGB-D data. Xie et al. [19] build a large-scale synthetic tabletop dataset TOD, which can be used for pre-training, and Back at al. [14] further generate another simulated dataset with amodal annotation.

B. Boundary-aware instance segmentation

With the development of 2D image instance segmentation [18], [22], [23], [24], researchers found that mask precision highly depends on the boundary quality. To obtain clean and sharp contours for object instances, many recent studies [25], [26], [27], [28], [29], [30], [31], [32], [33], [34] paid more attention to the object boundary. Cheng et al. [28], [30] predict both instance-level masks and boundaries, the method gains improvements on the subtle regions. Tang et al. [31] sample small patches on the contour and use semantic segmentation to refine the patches. Kim et al. [32] learn a global boundary representation for high-frequency details; their results outperform the BlendMask [35] baseline. Zhang et al. [33] propose Boundary-Aware Refinement and repeatedly leverage boundary information to upsample the mask. Cheng et al. [29] state that boundary IoU is significantly more sensitive than mask IoU, so using boundary IoU for evaluation generates mask predictions of higher fidelity. We further show that driving the network to predict boundary IoU is an effective way to improve the mask quality, and the IoU can provide strong guidance to indicate if an instance is suitable for picking, as shown in Sec. IV-F.

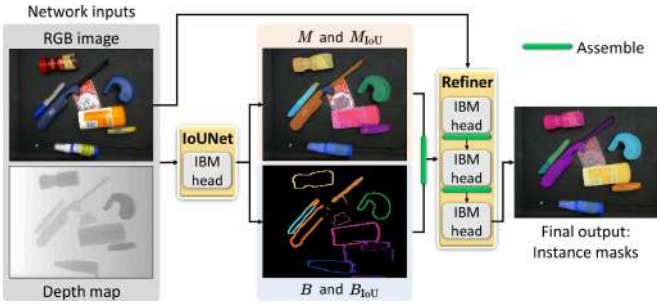

Fig. 2. Overview of our pipeline. Given an RGB-D input, we first design the IoUNet composed mainly of the IBM head to predict instance mask M, boundary mask B, and their associated mask scores M_{IoU} & B_{IoU}. Next, we design a hierarchical refiner with multiple IBM heads to produce masks of higher precision. Note that, "assemble" means the process of assembling a re-weighted probabilistic mask from $\{M, M_{\text{IoU}}, B, B_{\text{IoU}}\}$.

III. Method

A. Overview

Given an RGB-D image of a grasping scenario, our method provides accurate category-agnostic instance masks by delving boundary cues. The core of our instance segmentation method is the IoU-based Boundary-aware Mask head (IBM head) that predicts both instance-level masks M and boundaries B, as well as their associated IoU scores M_{IoU} and B_{IoU}, respectively. Note that, all these predictions can be assembled into a re-weighted probabilistic mask result M_{ins} for further refinement. Equipped with the IBM head, we propose the coarse-to-fine pipeline shown in Fig. 2.

In the first stage, we design the IoUNet composed mainly of the IBM head to predict M, M_{IoU}, B, and B_{IoU}. Benefited by the IBM head with boundary awareness, we can already separate nearby instances in this stage. However, considering that some scenarios may be rather cluttered and chaotic, the current results can be coarse. We thus assemble the coarse predictions and design the hierarchical mask refiner to further enhance the fidelity of each predicted mask. As shown on the right-hand side of Fig. 2, besides stacking a series of IBM heads with assembling to gradually refine instance masks for higher quality and cleaner boundaries, we further feed the original RGB image into our refiner for semantic segmentation, which provides rich texture and semantic features to guide the refinement. Below, we shall

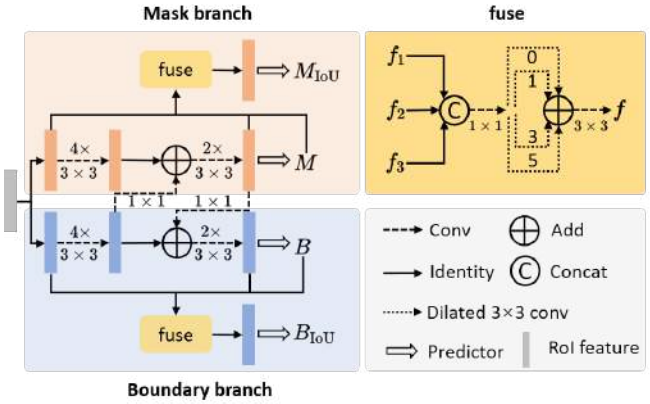

Fig. 3. Illustration of the IBM head. Given the RoI feature, we design a mask branch and a boundary branch to predict masks and boundaries, respectively. Considering the strong relationship between an instance mask and its boundary, the two branches pass the intermediate features to each other to enrich the information. Besides, we design a fuse module with dilated convs to fuse features of different levels for accurate IoU prediction.

elaborate on the details of IBM head, assembling, IoUNet, and refiner.

B. IoU-based Boundary-aware Mask Head

Typical two-stage instance segmentation methods [18], [36] first generate RoI (Regions of Interest) features via region proposal network (RPN) and RoI Align, which are then fed into the bbox head and the mask head for bounding box and instance mask regression, respectively. However, they ignore the importance of boundary for instance mask prediction. To this end, we propose the IoU-based Boundary-aware Mask head (IBM head) to predict not only instance mask but also the boundary and their associated IoU values for each RoI. Our IBM head can be plugged in arbitrary RoI-based methods for consistent boundary improvement.

Fig. 3 illustrates the detailed architecture of our IBM head, which consists of two branches: mask branch and boundary branch. The two branches take the extracted RoI feature as input. Considering that the boundary contains rich information of the inter-instance spatial relationship, it can guide the mask feature to focus on the distinct edge region and avoid confusion between nearby instances. We thus first add the boundary feature to the mask feature to get the fused mask feature. On the other hand, considering that the mask information describes a coarse boundary signal, we then add the fused mask feature to the boundary feature to get the fused boundary feature. Finally, we predict mask M and boundary B from the two fused features via a predictor [18], which contains two 3×3 convs and a 1×1 conv.

Apart from the segmentation results, we also predict IoU scores of both masks and boundaries, since the IoU value directly reflects the prediction quality compared to the ground truths, thus effectively improving the precision. The two branches share the same process of regressing IoU scores. Specifically, as shown in the yellow box of Fig. 3, we concatenate the input RoI feature, the fused mask/boundary feature, as well as the segmentation logits together and pass through parallel 3×3 dilated convs with dilations $\{0, 1, 3, 5\}$ to obtain multi-level feature. At last, we also employ the

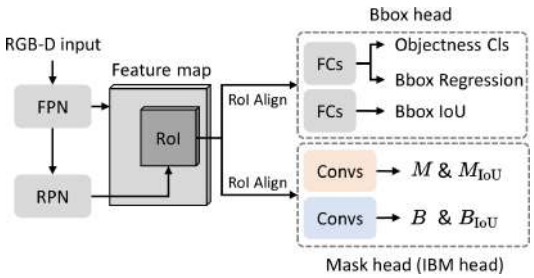

Fig. 4. The detailed architecture of IoUNet.

predictor to obtain M_{IoU} and B_{IoU}.

To evaluate the quality of M, we use cross entropy loss [18] as the pixel-level mask loss. To evaluate the quality of B, we follow [28] to use Laplacian operator to generate soft boundaries from the mask ground truths $\hat{M}$, then use Dice loss and binary cross entropy loss as the pixel-level boundary loss. To evaluate the accuracy of M_{IoU} and B_{IoU}, we calculate the mean squared error between our predicted and ground-truth IoU values as the IoU loss, and the ground-truth IoU values $\hat{M}_{\text{IoU}}$ and $\hat{B}_{\text{IoU}}$ are generated following [29]:

$$\hat{M}_{\text{IoU}} = \frac{\left|\hat{M} \cap M\right|}{\left|\hat{M} \cup M\right|}, \hat{B}_{\text{IoU}} = \frac{\left|\left(\hat{M}_d \cap \hat{M}\right) \cap \left(M_d \cap M\right)\right|}{\left|\left(\hat{M}_d \cap M^{gt}\right) \cup \left(M_d \cap M\right)\right|}, \tag{1}$$

where $\hat{M}_d$ and M_d represent the set of the original masks' pixels within distance d from the contour.

C. IoUNet

As shown in Fig. 2, the purpose of IoUNet is to generate coarse instance masks and boundaries $\{M, M_{\text{IoU}}, B, B_{\text{IoU}}\}$ given an RGB-D input. Fig. 4 shows its detailed architecture. Similar to the two-stage Mask R-CNN [18], we first extract features through Feature Pyramid Network (FPN) and detect object RoIs through Region Proposal Network (RPN). Here, we use two ResNet backbones separately extracting features from RGB and depth, then fuse C2-C5 level features to use FPN and RPN for detection. Then we use a bbox head to regress the precise location of bounding box (denoted as BB) and a mask head (i.e., our IBM head) to segment a pixel-level mask for each RoI.

To delve localization-based objectness cues for bounding box, unlike Mask R-CNN, we design the bbox head to further score each bounding box regarding its IoU value. As shown on top-right of Fig. 4, apart from a bunch of box locations, we also build a branch with two fully-connected (FC) layers to predict IoU per bounding box. Doing so gives us an intuitive cue of how well the predicted box is. Empirically, as we shall show in Sec. IV-F, bbox IoU is an effective signal to support grasping, since it somehow indicates whether an instance is occluded or exposed. Hence, to guide the learning of bbox head, besides the common Smooth L1 loss in box regression, we use an extra Distance IoU loss $\mathcal{L}_{\text{DIoU}}$ [37]:

$$\mathcal{L}_{\text{DIoU}} = \frac{\left|\hat{BB} \cap BB\right|}{\left|\hat{BB} \cup BB\right|} - \frac{\text{Dist}(\hat{BB}, BB)}{c^2}, \tag{2}$$

TABLE I

COMPARING INSTANCE SEGMENTATION PERFORMANCE OF OUR METHOD AGAINST OTHERS ON THE GRASPNET DATASET.

Methods	GraspNet Seen Split						GraspNet Unseen Split					
	P_m	R_m	F_m	P_b	R_b	F_b	P_m	R_m	F_m	P_b	R_b	F_b
Mask R-CNN [18]	93.4	73.7	80.4	74.0	57.6	63.9	90.3	62.9	72.5	68.4	50.6	55.7
BMask R-CNN [28]	95.2	75.7	83.5	79.7	57.0	65.5	89.5	66.2	75.0	70.8	52.3	59.0
RefineMask [33]	95.0	76.9	84.7	81.4	59.3	67.8	**90.7**	64.4	74.3	70.8	51.8	58.6
Transfiner [1]	**96.6**	74.2	84.0	81.1	57.3	66.8	90.6	66.2	74.8	70.0	51.6	57.9
Ours (w/o refiner)	94.2	82.2	87.2	81.7	63.8	70.9	89.8	65.2	75.5	71.3	54.7	61.5
Ours	95.9	**82.2**	**88.1**	**82.5**	**65.1**	**72.2**	90.3	**67.6**	**76.1**	**71.9**	**56.1**	**62.4**

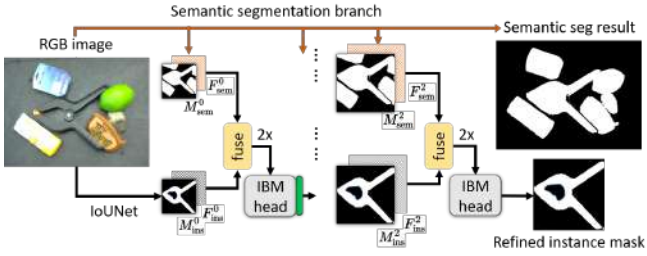

Fig. 5. We hierarchically refine the coarse mask with semantic context. In each stage, we fuse instance information with semantic context, and feed the upsampled 2x feature map into the IBM head to get the refined mask.

where $\hat{BB}$ is the ground-truth bbox, Dist is the Euclidean distance between two box centers, and c is the diagonal length of the smallest enclosing box covering the two boxes. Compared to the standard IoU loss $\frac{|\hat{BB}\cap BB|}{|\hat{BB}\cup BB|}$, $\mathcal{L}_{\text{DIoU}}$ also minimizes the center distance of two boxes towards faster convergence and more accurate localization.

D. Assembling Operation

Given a set of predictions $\{M, M_{\text{IoU}}, B, B_{\text{IoU}}\}$, we design an assembling operation to combine them into a re-weighted probabilistic map M_{ins} that indicates the per-pixel objectness. Specifically, we first get scored mask $M_s = M \times M_{\text{IoU}}$ and scored boundary $B_s = Blur(B \times B_{\text{IoU}})$, where $Blur$ indicates a Gaussian blurring operator of kernel size 10 to coarsely extend the boundary. We then adjust the mask M with the boundary information to obtain M_{ins}:

$$M_{\text{ins}} = (1 - M_s) \otimes B_s + M_s \quad (3)$$

$$= (1 - B_s) \otimes M_s + B_s. \quad (4)$$

Intuitively, the function has plane of symmetry at $M_s = B_s$, which means the mask and boundary are equally important to determine objectness. When closer to the center of the mask, pixel objectness is mainly controlled by M_s; when closer to the outer boundary, B_s plays a critical role to increase the probability of this distinctive region, and is able to correct potential mistakes of M_s for a robust M_{ins}. In this way, we combine the complementary strengths of mask and boundary.

E. Hierarchical Mask Refiner

As shown in Fig. 2, once we obtain $\{M, M_{\text{IoU}}, B, B_{\text{IoU}}\}$ by the IoUNet, we employ our proposed assembling process to generate a coarse instance mask M_{ins}. The current M_{ins} yet may have coarse contours in cluttered and chaotic scenes, we thus propose a hierarchical mask refiner to further detail the boundary regions with the key idea of incorporating an RGB semantic segmentation branch [24] to explore more contextual information; see Fig. 5 as an illustration.

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$, as illustrated by the orange arrow, we design a semantic segmentation branch using ResNet50-FCN8 [38] to generate semantic segmentation result $M_{\text{sem}} \in \mathbb{R}^{H\times W\times 1}$, where the third dimension indicates the pixel as foreground or background. We also collect the Res2 layer feature $F_{\text{sem}} \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C_{\text{sem}}}$, which has the highest resolution and gains most spatial context. Specifically, in stage 0, we acquire the semantic information through RoI Align to obtain $M^0_{\text{sem}} \in \mathbb{R}^{14\times 14\times 1}$ and the corresponding feature $F^0_{\text{sem}} \in \mathbb{R}^{14\times 14\times C_{\text{sem}}}$. Meanwhile, from IoUNet and assembling, we also have the coarse mask map $M^0_{\text{ins}} \in \mathbb{R}^{14\times 14\times 1}$ and its corresponding RoI feature $F^0_{\text{ins}} \in \mathbb{R}^{14\times 14\times C_{\text{ins}}}$ (i.e., the feature fed into mask branch in Fig. 3). We fuse $\{M^0_{\text{ins}}, F^0_{\text{ins}}, M^0_{\text{sem}}, F^0_{\text{sem}}\}$ with the same architecture as the fuse module in Fig. 3 (concatenate then dilated convs), then upsample the fuse feature to get $F^1_{\text{ins}} \in \mathbb{R}^{28\times 28\times C_{\text{ins}}}$. This feature is then passed through the IBM head and assembling to obtain the next-level mask map $M^1_{\text{ins}} \in \mathbb{R}^{28\times 28\times 1}$. We repeat the above procedure with hierarchical mask size $\{14, 28, 56\}$ for stage $\{0, 1, 2\}$, until we get the final refined mask of size 112×112. For the last IBM head, we directly use the mask branch outputs as the final results instead of using assembling to combine boundary and mask information. At inference, the final mask score per instance is its bbox IoU multiplied by its mask IoU.

IV. EXPERIMENTS AND RESULTS

A. Implementation Details and Evaluation Metrics

We follow the PyTorch detection framework [39] to train our framework for 12 epochs with the SGD [40] optimizer with a learning rate of 0.01. Using the GraspNet dataset [6] with 25,600 images, the training takes about 30h and the inference runs at around 4.2fps on a single GeForce GTX 1080 Ti GPU. To evaluate the quality of our predicted instance masks and boundaries, we adopt precision, recall, and F-measure as the evaluation metrics. Specifically, we use $P_m, R_m \& F_m$ to denote the three metric values calculated on masks, and $P_b, R_b \& F_b$ on boundaries. For instance boundary, we calculate P_b and R_b following [19], [14]. Please refer to their original papers for more details.

B. Evaluation on GraspNet Dataset

GraspNet [6] provides 190 cluttered and complex scenes from RealSense D435 covering 88 object categories: 100 scenes for training, 30 for seen object testing, and 60 for

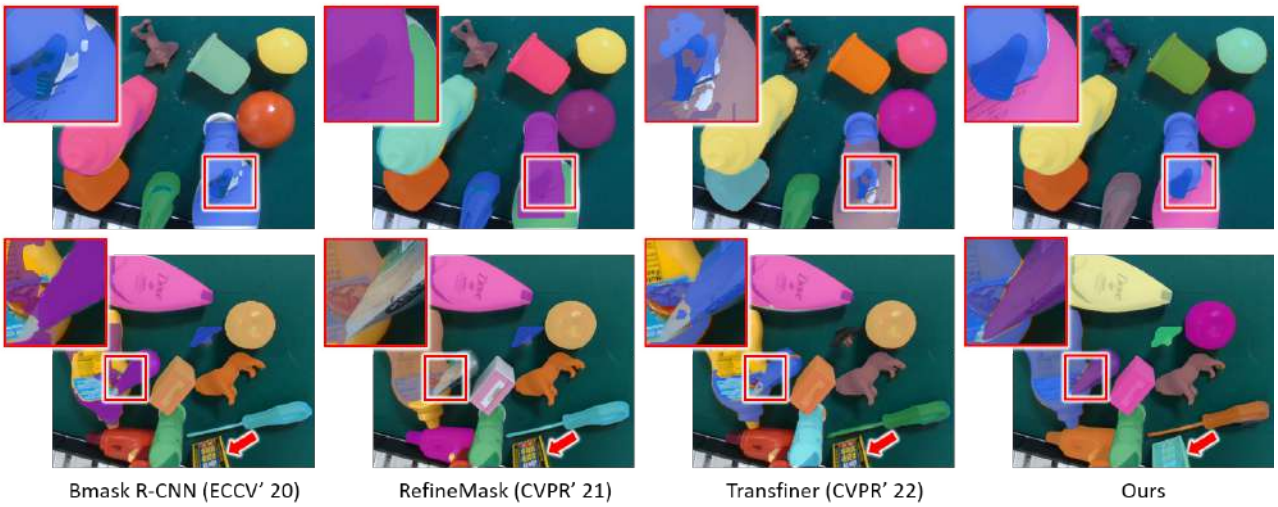

Fig. 6. Comparing instance segmentation results with the SOTA methods on the GraspNet. We emphasize the cluttered regions in red boxes. Our method generates precise masks with clean and sharp boundaries for these regions, whereas other methods fail to accurately distinguish the nearby instances.

TABLE II

COMPARING INSTANCE SEGMENTATION PERFORMANCE OF OUR METHOD AGAINST OTHERS ON WISDOM-REAL.

Methods	P_m	R_m	F_m	P_b	R_b	F_b
Mask R-CNN [18]	74.6	76.0	75.1	45.5	59.4	50.9
BMask R-CNN [28]	78.5	76.7	77.5	49.4	56.5	52.3
RefineMask [33]	84.5	78.1	80.0	57.7	60.2	58.5
Transfiner [1]	85.3	77.4	79.7	57.4	**61.2**	58.9
Ours (w/o refiner)	85.2	77.8	80.6	61.1	58.1	58.9
Ours	**86.8**	**78.5**	**81.7**	**65.0**	60.8	**62.2**

TABLE III

COMPARING INSTANCE SEGMENTATION PERFORMANCE OF OUR METHOD AGAINST OTHERS ON OCID.

Methods	P_m	R_m	F_m	P_b	R_b	F_b
Mask R-CNN [18]	80.8	73.9	76.1	68.2	58.4	61.8
UOIS-2D [4]	88.3	78.9	81.7	82.0	65.9	71.4
UOIS-3D [19]	86.5	86.6	86.4	80.0	73.4	76.2
UCN [41]	86.0	**92.3**	**88.5**	80.4	78.3	78.8
UOAIS [14]	70.7	86.7	71.9	68.2	**78.5**	68.8
Ours†	**89.0**	84.8	86.7	**84.5**	76.4	**79.0**

unseen object testing. Each scene has 256 different images. We compare our method with Mask R-CNN [18] and recent boundary-related instance segmentation methods [28], [33], [1]. The initial input to all methods is an RGB-D image and the input to the refiner (if any) is an RGB image. Tab. I summarizes the results. Clearly, our method gains significant improvements on both seen and unseen splits. Particularly, our method consistently outperforms all others on boundary metrics, even without our designed refiner, indicating the effectiveness of our designed IBM head. Some typical results are shown in Fig. 6. With boundary-awareness, our method precisely segments nearby objects with clear boundaries; see the overlapping regions marked by red boxes and arrows.

C. Evaluation on WISDOM Dataset

WISDOM [17] is a warehouse bin-picking dataset, in which the WISDOM-Real split contains 400 real-world RGB-D images with an average of 4.8 object instances per image. Following [14], we train all networks on the UOAIS-Sim bin-picking dataset [14], with 22,500 synthetic RGB-D images, then directly test on WISDOM-Real split.

TABLE IV

COMPARING INSTANCE SEGMENTATION PERFORMANCE OF OUR IBM HEAD AGAINST OTHER COMMON MASK HEADS.

Methods	P_m	F_m	AP_l	AP_m	AP_s	P_b	F_b
Mask R-CNN [18]	84.9	70.1	48.7	44.0	10.3	59.7	53.1
MS R-CNN [36]	85.8	71.2	50.6	44.0	9.8	61.7	54.0
BMask R-CNN [28]	88.5	74.9	56.7	46.5	12.2	63.8	56.0
Ours‡	**90.0**	**77.6**	**57.8**	**47.4**	**14.9**	**66.5**	**57.6**

TABLE V

ABLATION STUDY OF EACH PROPOSED MODULE ON WISDOM-REAL.

	IBM head	IoUNet (Det)	Refiner	P_m	R_m	F_m	P_b	R_b	F_b
1	-	-	-	74.5	72.9	74.0	45.5	56.7	48.3
2	✓	-	-	82.9	73.7	77.6	56.1	57.0	56.2
3	✓	✓	-	85.2	77.8	80.6	61.1	58.1	58.9
4	✓	✓	✓(S)	85.5	77.8	80.6	63.5	60.0	61.6
5	✓	✓	✓(M)	**86.8**	**78.5**	**81.7**	**65.0**	**60.8**	**62.2**

Hence, the challenge of this experiment is the sim-to-real generalization. Tab. II summarizes the comparison results. Similarly, our method achieves almost the best performance on both the mask and boundary predictions. Please refer to our supplemental video for more visual results.

D. Evaluation on OCID Dataset

To further evaluate the generalization ability of our method, we conduct experiments for unseen object instance segmentation (UOIS) on the OCID [21] dataset with 2,346 RGB-D images and 7.5 objects per image on average. We use TOD [19], a synthetic tabletop dataset with 40,000 RGB-D images for training. Towards better perception for occluded regions on the UOIS task, we modified our method to an amodal instance segmentation pipeline inspired by UOAIS [14], which separately predicts the visible and amodal parts with a boundary-aware manner. The modified version is noted with †. Tab. III summarizes the comparison results against SOTA UOIS methods. We can see that even segmenting on unseen objects, our method still achieves the highest values on three metrics, especially on precision.

E. Ablation Studies

a) Head comparison: To verify the effectiveness of our IBM head, we compare it with three different mask

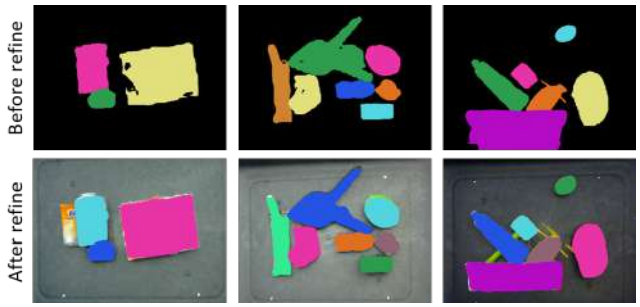

Fig. 7. Illustration of the effect of our hierarchical mask refiner; see particularly the boundary quality of larger objects.

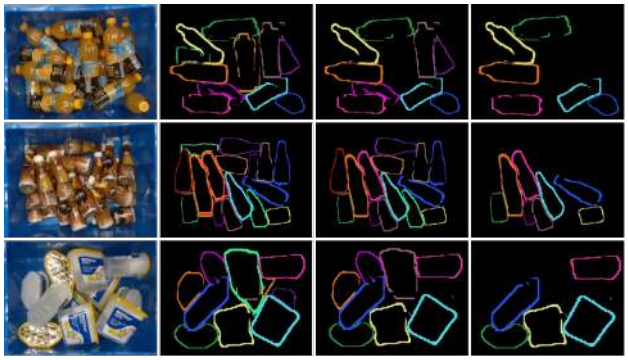

Fig. 8. Filtered results using different bbox IoU thresholds.

heads from [18], [36] and [28]. For fairness, we omit the refiner and use the detection in [18] instead of our design in IoUNet (Sec. III-C). We only use the coarse predictions for comparison, noted with ‡. Tab. IV shows the results. We use WISDOM-Sim with 50,000 synthetic depth images for training and WISDOM-Real with 400 real depth images for testing. The ablation only uses depth images without the RGB modality. We use the evaluation metrics in [39] for analysis, where AP_l, AP_m, and AP_s indicate average precision for large, medium, and small objects, respectively. Clearly, our method achieves the best performance for all metrics, especially on AP_l for objects of large sizes. Equipped with boundary-awareness, both BMask R-CNN [28] and our method significantly strengthen the model to describe fine contours. Yet, our method with both mask and boundary IoU predictions further outperforms BMask R-CNN on AP for all objects of diverse sizes.

b) Module ablation: In Tab. V, we verify the effectiveness of our proposed modules following the setting in Sec. IV-C. By comparing the first three rows, we can see that our IBM head gains significant improvement for precision, while our IoU-based detection (Sec. III-C) contributes for the recall. In addition, our hierarchical mask refiner further improves the segmentation performance, where (S) and (M) indicate single-stage and multi-stage refinement, respectively. Fig. 7 shows the visual effect of our refiner. We can see that our refiner plays an important role for improving the boundary quality, particularly for large objects.

F. Real-world Picking Demonstrations

We deploy our method for warehouse bin picking of automatic supermarkets. In this scenario, multiple objects for

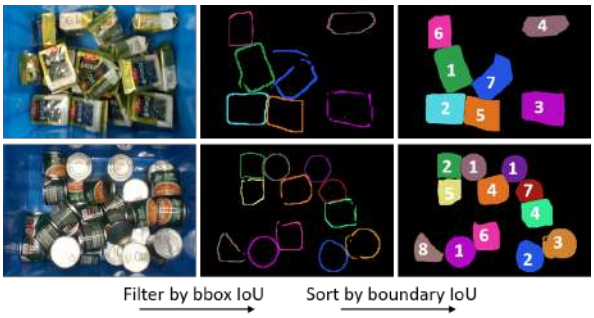

Fig. 9. The sorted results according to the boundary IoU scores.

Fig. 10. A picking sequence according to our mask selection strategy, where the object(s) painted white has the highest boundary IoU value.

sales are randomly cluttered in the container. We annotate 1,000 RGB images of 50 different objects for training, then deploy our method in the storehouse scene for bin-picking testing. Note that our input to the IoUNet is the RGB modality instead of RGB-D in this demonstration, thus we use one ResNet50 as backbone accordingly. To enable sequential grasping, we propose a reliable evaluation scheme based on our IoU predictions. Specifically, as mentioned in Sec. III-C, bbox IoU is a strong guidance whether the instance is exposed. This is verified in Fig. 8, where we show the influence using different thresholds. We set a bbox IoU threshold of 0.75 to filter out the occluded instances. Given the remaining candidates of high objectness, we further find that their boundary IoU varies w.r.t. texture distinctiveness. A higher boundary IoU indicates the salient local distinctiveness as well as a relatively clean surrounding. As illustrated in Fig. 9, we sort these instances by their boundary IoU scores to guide the gripper for picking, , thereby avoiding failures caused by picking occluded objects. Following the above procedures, Fig. 10 shows an example of grasping, where the object(s) marked in white is the best one to pick at the moment. Video demonstrations are provided in our supplemental file.

V. Conclusion

In this work, we propose a new instance segmentation method based specifically on improving the mask boundary. Towards boundary learning, we develop the novel IBM head to predict instance-level mask, boundary, as well as their associated IoU scores. The boundary IoU learning facilitates the network to concentrate on contour details and separate closely-packed instances. Equipped with the IBM head, we design a coarse-to-fine framework to first generate coarse masks and then refine them gradually with semantic contexts. Our method achieves excellent instance segmentation results on three benchmarks. For real-world grasping, we propose an IoU-based strategy to select the grasping target with the highest visibility. Experimental results show that our predicted bbox and boundary IoU values can provide a strong guidance on whether an instance is easy to pick.

REFERENCES

[1] L. Ke, M. Danelljan, X. Li, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask transfiner for high-quality instance segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4412–4421.

[2] M. Schwarz, A. Milan, A. S. Periyasamy, and S. Behnke, "RGB-D object detection and semantic segmentation for autonomous manipulation in clutter," *Int. Jour. Robotics Research*, vol. 37, no. 4-5, pp. 437–451, 2018.

[3] 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 *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2019, pp. 9558–9564.

[4] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging RGB and depth for unseen object instance segmentation," in *Conf. on Robotics Learning*, 2020, pp. 1369–1378.

[5] S. Back, J. Kim, R. Kang, S. Choi, and K. Lee, "Segmenting unseen industrial components in a heavy clutter using RGB-D fusion and synthetic data," in *IEEE Int. Conf. on Image Processing (ICIP)*, 2020, pp. 828–832.

[6] I. C. on Computer Vision and P. R. (CVPR), "Graspnet-1billion: A large-scale benchmark for general object grasping," 2020, pp. 11 444–11 453.

[7] S. Hasegawa, K. Wada, S. Kitagawa, Y. Uchimi, K. Okada, and M. Inaba, "Graspfusion: Realizing complex motion by learning and fusing grasp modalities with instance segmentation," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2019, pp. 7235–7241.

[8] K. Fang, Y. Bai, S. Hinterstoisser, S. Savarese, and M. Kalakrishnan, "Multi-task domain adaptation for deep learning of instance grasping from simulation," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2018, pp. 3516–3523.

[9] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang, and L. Li, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3571–3578.

[10] G. Xu, Y. Tao, B. Jiang, P. Wang, Y. Luo, and J. Zhong, "Pois: Policy-oriented instance segmentation for ambidextrous robot picking," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021, pp. 743–749.

[11] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *European Conf. on Computer Vision (ECCV)*, 2014, pp. 345–360.

[12] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6D pose estimation in the amazon picking challenge," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2017, pp. 1386–1383.

[13] K. Wada, S. Kitagawa, K. Okada, and M. Inaba, "Instance segmentation of visible and occluded regions for finding and picking target from a pile of objects," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2048–2055.

[14] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee, "Unseen object amodal instance segmentation via hierarchical occlusion modeling," *arXiv preprint arXiv:2109.11103*, 2021.

[15] W. Boerdijk, M. Sundermeyer, M. Durner, and R. Triebel, ""what's this?"-learning to segment unknown objects from manipulation sequences," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021, pp. 10 160–10 167.

[16] M. Durner, W. Boerdijk, M. Sundermeyer, W. Friedl, Z.-C. Márton, and R. Triebel, "Unknown object segmentation from stereo images," in *Int. Conf. on Intell. Robots and Systems (IROS)*, 2021, pp. 4823–4830.

[17] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3D objects from real depth images using mask r-cnn trained on synthetic data," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2019, pp. 7283–7290.

[18] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2961–2969.

[19] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Trans. on Robotics (T-RO)*, pp. 1–17, 2021.

[20] L. Zhang, S. Zhang, X. Yang, and Z. Liu, "Unseen object instance segmentation with fully test-time rgb-d embeddings adaptation," *arXiv preprint arXiv:2204.09847*, 2022.

[21] M. Suchi, T. Patten, D. Fischinger, and M. Vincze, "Easylabel: A semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6678–6684.

[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8759–8768.

[23] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *IEEE Int. Conf. on Computer Vision (ICCV)*, 2019, pp. 9157–9166.

[24] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang, *et al.*, "Hybrid task cascade for instance segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4974–4983.

[25] Z. Hayder, X. He, and M. Salzmann, "Boundary-aware instance segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5696–5704.

[26] A. Kirillov, E. Levinkov, B. Andres, B. Savchynskyy, and C. Rother, "Instancecut: from edges to instances with multicut," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5008–5017.

[27] R. S. Zimmermann and J. N. Siems, "Faster training of mask r-cnn by focusing on instance boundaries," *Computer Vision and Image Understanding*, vol. 188, p. 102795, 2019.

[28] T. Cheng, X. Wang, L. Huang, and W. Liu, "Boundary-preserving mask r-cnn," in *European Conf. on Computer Vision (ECCV)*, 2020, pp. 660–676.

[29] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary iou: Improving object-centric image segmentation evaluation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 15 334–15 342.

[30] L. Ke, Y.-W. Tai, and C.-K. Tang, "Deep occlusion-aware instance segmentation with overlapping bilayers," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 4019–4028.

[31] C. Tang, H. Chen, X. Li, J. Li, Z. Zhang, and X. Hu, "Look closer to segment better: Boundary patch refinement for instance segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 13 926–13 935.

[32] M. Kim, S. Woo, D. Kim, and I. S. Kweon, "The devil is in the boundary: Exploiting boundary representation for basis-based instance segmentation," in *IEEE Winter Conf. on Applications of Computer Vision (WACV)*, 2021, pp. 929–938.

[33] G. Zhang, X. Lu, J. Tan, J. Li, Z. Zhang, Q. Li, and X. Hu, "Refinemask: Towards high-quality instance segmentation with fine-grained features," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 6861–6869.

[34] S. Chennupati, V. Narayanan, G. Sistu, S. Yogamani, and S. A. Rawashdeh, "Learning panoptic segmentation from instance contours," in *IEEE Int. Conf. on Robotics and Automation (ICRA)*, 2021, pp. 9586–9593.

[35] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "BlendMask: Top-down meets bottom-up for instance segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8573–8581.

[36] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask Scoring R-CNN," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 6409–6418.

[37] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: Faster and better learning for bounding box regression," in *AAAI Conf. on Artificial Intell. (AAAI)*, vol. 34, no. 07, 2020, pp. 12 993–13 000.

[38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[39] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[40] M. Zinkevich, M. Weimer, L. Li, and A. Smola, "Parallelized stochastic gradient descent," *Conf. and Workshop on Neural Information Processing Systems (NeurIPS)*, vol. 23, 2010.

[41] 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.