POIS: Policy-Oriented Instance Segmentation for Ambidextrous Robot Picking

Guangyun Xu[1, 2,*], Yi Tao[3, *], Bowen Jiang[4,*], Peng Wang[1,2,5,✉], Yongkang Luo[1], Jun Zhong[1]

***Abstract*— Robots with a parallel-jaw gripper and suction cup is an adaptive and efficient robotic picking system. This paper proposed Policy-Oriented Instance Segmentation (POIS) for ambidextrous robots. POIS can generate a pair of target masks that allows ambidextrous robots to pick in parallel. It takes a depth image and predicts initial mask, center offset, and policy confidence map through three paralleled branches. We incorporate the initial mask with center offset to obtain candidate instances, from which we select masks of target objects for policy execution (decided with policy confidence map). We also provide a dataset that contains 6k synthetic scenes and 100 real scenes for ambidextrous picking. Trained on synthetic scenes, POIS generalizes well in real scene and is capable of handling novel objects in cluttered scenes. Our dataset and video are available at https://bit.ly/3oJj8Tu.**

I. Introduction

Universal picking has a huge potential in e-commerce logistic management [1], [2], manufacturing [3] and service robots [4]–[6], and its application requires an adaptative and efficient robotic picking system. Using multiple grippers is an effective way to improve adaptability [7]–[9]. Parallel-jaw gripper and suction cup complement each other: Vacuum-based suction-cup grippers can quickly pick larger objects with planar surfaces such as boxes, and the parallel-jaw grippers can easily pick small objects, such as paper clips, or irregular-shaped objects, such as cup [10]–[13]. Therefore, combining them increases robots' adaptability to object with various geometries and materials. In Amazon Picking Challenge [14], many teams have used the gripper with a retractable mechanism that enables quick switching between suction and grasping [14]–[16].

Pick-policy-making is a critical part of a multi-gripper robot. Andy Zeng et al [15] used Suction Affordance ConvNet and Horizontal Grasp Affordance ConvNet to generate pixel-wise confidence maps for grasping and suction from a multi-viewed image and choose the picking pose with the highest confidence score. However, the training takes a massive hand-labelled pick proposal dataset, which is too expensive for industrial applications. Moreover, the multi-functional gripper designed for the competition has a delicate mechanical structure, and frequent switch can cause damage. Dex-Net 4.0 [17] enables ambidextrous robots to switch between two arms and successfully pick a wide range of objects with various materials and geometries. However, only one arm operating at a time is undesired in efficiency.

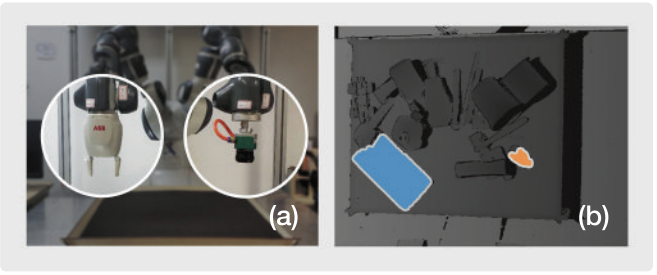

Fig. 1: The method takes a depth image and outputs a pair of masks, one object for suction(blue mask in (b)) and the other for grasping(orange mask in (b)), for ambidextrous picking (a).

Therefore, we propose Policy-Oriented Instance Segmentation (POIS) that outputs a pair of masks that allows ambidextrous robots to pick in parallel (as shown in Fig. 1). The selection of the best target pair depends on the position, pose, and geometry, and we use a neural network to deal with complicated factors. Trained on 5k synthetic scenes, POIS generalize well in real space and is capable of handling novel objects in cluttered scenes. Besides, we propose Mask-based Grasp Pose Detection (Mask-GPD) and Mask-based Suction Pose Detection (Mask-SPD) to generate picking pose as ground truth label for training POIS and evaluating segmentation result. It is noteworthy that we estimate the centroid of the target object in the analysis. With the contact close to the centroid, the gripper holds the object more stably.

The contributions are as follows:

1) A policy-oriented instance segmentation (POIS) method which can generate a pair of masks that enables ambidextrous robots to pick in parallel.
2) A dataset that contains synthetic and real data for policy-oriented instance segmentation.
3) A set of metrics that assess instance segmentation based on ambidextrous picking performance.

II. Related Work

A. Multi-Gripper Picking

Many multi-gripper picking solutions show that combining parallel-jaw gripper and suction cup is an effective way to increase picking adaptability [14]–[17]. In 2017 Amazon

This work was supported in part by the National Natural Science Foundation of China under Grants (91748131, 61771471, and 62006229), and the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050100.

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China.
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China
[3] Columbia University, New York, NY 10027, USA
[4] Harvey Mudd College, Claremont, CA 91711, USA
[5] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China
* Authors contributed equally.
✉ Corresponding author: peng_wang@ia.ac.cn

Picking Challenge, Andy Zeng et al. [15] innovated a multi-functional gripper with a retractable mechanism that enables switching between suction and grasping. By comparing the confidence maps of grasp and suction, the robot switches between the two grippers to take action for good performance in cluttered scenes. However, the training procedure needs a massive hand-labelled picking proposals dataset, which is too expensive for industrial applications. Multi-functional gripper designed for the competition has a delicate mechanical structure, and frequent switch can cause damage.

To solve the above problems, Dex-Net 4.0 [17] utilized ambidextrous robots and can be trained on synthetic data. Dex-Net 4.0 uses Grasp Quality Convolutional Neural Network (GQ-CNN) to predict the grasp qualities of parallel-jaws and suction cups, and select a gripper by maximizing grasp quality. The robot switches between two arms and successfully pick a wide range of objects with various materials and geometries. However, with only one arm moving at a time, it does not take full advantage of ambidextrousness.

B. Segmentation of Unseen Objects

The problem of unseen object instance segmentation (UOIS) [18] is a significant problem in universal picking [19]–[21]. SD Mask R-CNN [22] uses a large amount of synthetic data to train instance segmentation network Mask R-CNN [23] category-agnostically. Through massive randomization over a diverse set of 3D objects, camera poses, and camera intrinsic parameters, SD Mask R-CNN does not need extra techniques to tackle the Sim-to-Real problem.

UOIS-Net [18] is a more sophisticated network that can generate sharper masks by taking advantage of depth images and RGB images separately. It uses the two-stage method: Depth Seeding Network (DSN) uses depth image to generate initial mask; Region Refinement Network (RRN) uses RGB image and initial mask to produce refined mask. Trained on synthetic data, it shows great efficacy on multiple datasets for UOIS in tabletop environments. Inspired by its architecture, we developed a policy-oriented instance segmentation network for ambidextrous picking.

III. Problem Statement

With a depth image as input, our goal is to find a pair of masks, with one object for suction and the other for grasping, that allows parallel picking from the cluttered scene.

A. Definition

We use the following definition in this paper:

- Pick: The overall name of grasp and suction.
- Suction reference point P_s: The farthest point to the initial position of suction gripper.
- Grasp reference point P_g: The farthest point to the initial position of parallel-jaw gripper.
- Null policy: A null policy is a policy that does nothing.
- Target mask: Mask of a selected object (either through grasp or suction).
- Interference-free: Two arms' paralleled motions do not interfere (such as colliding or blocking) with each other.

B. Assumptions

- The moving distance of gripper is inversely proportional to the distance between target ran reference point.
- According [24] [25] [26], the difficulty of path planning is inversely proportional to the distance between the two targets where they locate in their workspace.
- A single overhead depth sensor with known intrinsic, position, and orientation relative to the robot.

IV. Method

We consider the problem of generating a pair of masks for paralleled picking in clutter. The architecture of POIS is shown in Fig. 2. We also provide Mask-GPD and Mask-SPD to generate picking poses to verify the segmentation result.

A. POIS Network

1) Network Architecture: We consider the problem of generating a pair of interference-free picking poses for unseen objects in clutter.

The network takes as input a 3-channel (XYZ coordinate) organized point cloud, $D \in^{H \times W \times 3}$ (H, W denote height and width), and outputs n masks of targets ($n \in \{0, 1, 2\}$). Point clouds D are converted from depth images.

In order to obtain a higher receptive field, we use a modified U-Net [27] used in UOIS [18] as backbone. It takes point cloud D and outputs a feature map with 64 channels. Sitting on top of this are three parallel branches of convolutional layers that produce three outputs: initial mask $F \in^{H \times W \times C}$, where C is the number of semantic classes (target objects, other objects and background), policy confidence $P \in^{H \times W \times G}$, includes confidence of grasp, suction and null, and center offsets to object centers $V \in^{H \times W \times 3}$. Each pixel of V encodes a 3-dimensional offset vector pointing to the object's 3D center, so the predicted object centers for each pixel is $Obj^{co} = D + V$. While we use U-Net for the segmentation architecture, our framework can replace it with other suitable network architectures.

Instance Segmentation: we compute target masks from F,V and P. Firstly, we perform mean shift clustering [28] in 3D space over the center votes Obj^{co}. After clustering, each pixel is assigned to the center vote's cluster ID to generate the instance masks. The clustering is only applied to the target objects pixels for computational efficiency.

Policy decision: we use confidence map P to get C_g^k, C_s^k, the parallel-jaw grasping and suction confidence:

$$C_{\{g,s\}}^k = \frac{1}{N_k} \sum_{i=1}^{N_k} P_{\{g,s\}}^i, \tag{1}$$

where N_k is the number of pixels in the k-th mask, P_g^i, P_s^i is grasping confidence and suction confidence of the ith pixel.

We choose the mask with the largest P_g^i as the grasping target and the mask with the largest P_s^i as the suction target.

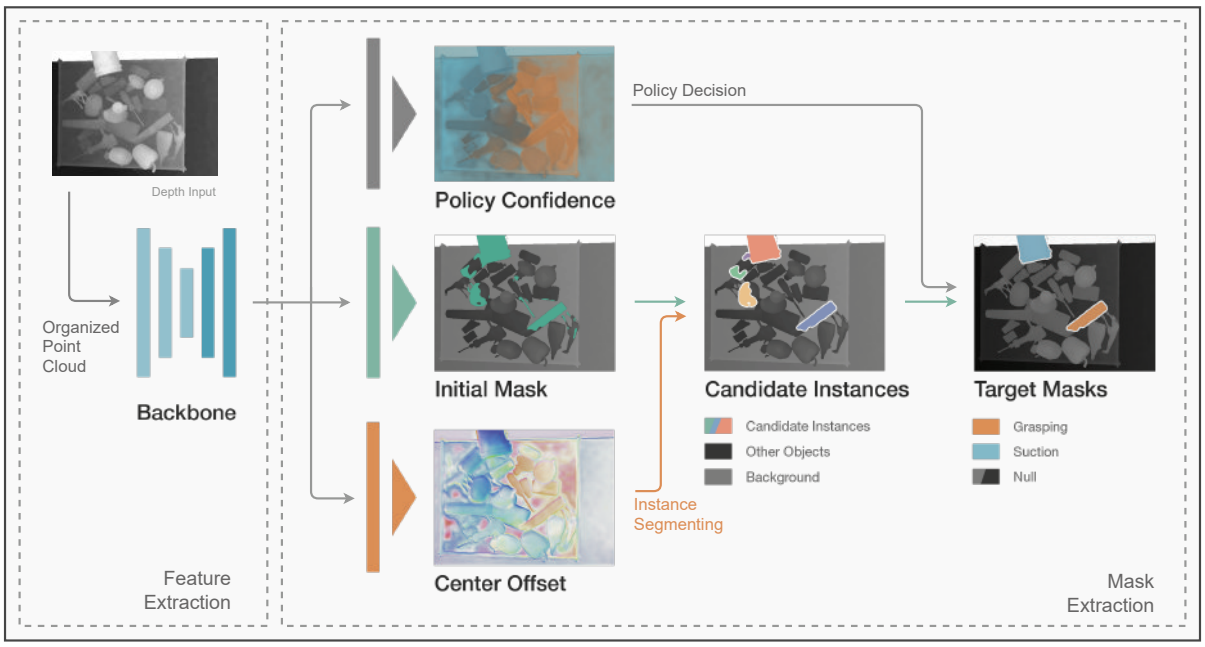

Fig. 2: POIS first converts a depth image to an organized point cloud and uses CNN to extract features. It produces three outputs: initial mask F, center offset V, and policy confidence P with three paralleled convolutional networks. We incooperate the initial mask with V to get instance masks, from which we select target masks (decided with policy confidence).

2) Loss Functions: target segmentation F, center offsets V and grasp confidences P each has its loss function. Consider the case of imbalanced image, we use a weighted cross entropy loss function for target segmentation F :

$$\ell_{ts} = \sum_{i \in \Omega} w_i^{ts} \ell_{CE} \left(\widehat{F}_i, F_i \right), \tag{2}$$

where Ω is the set of all pixels, $\hat{F}_i$ and F_i are the predicted and ground truth probabilities of pixel i, and ℓ_{CE} is the cross-entropy loss. The weight w_i^{ts} is inversely proportional to the number of pixels with labels equal to F_i, normalized to $[0, 1]$.

We apply a weighted Smooth L1 loss ℓ_{SL1} to the center offsets V to minimize the distance of the center votes to their corresponding ground truth object centers:

$$\ell_{co} = \sum_{i \in \Omega} w_i^{co} \ell_{SL1} \left(\widehat{Obj_i^{co}} - Obj_i^{co} \right), \tag{3}$$

where $Object_i$ is the 3D coordinate of the ground truth grasp target object center for pixel i. The weight w_i^{co} is inversely proportional to the number of pixels with the same grasp target object label t_i.

For grasp confidence P, we also use a weighted cross entropy:

$$\ell_{gc} = \sum_{i \in \Omega} w_i^{gc} \ell_{CE} \left(\widehat{P}_i, P_i \right), \tag{4}$$

where $\hat{P}_i$, P are the predicted and ground truth grasp confidence of pixel i, respectively. The weight w_i^{gc} is inversely proportional to the number of pixels with labels equal to P_i, normalized to $[0, 1]$.

In summary, the total loss is given by:

$$L = \lambda_{ts} \ell_{ts} + \lambda_{co} \ell_{co} + \lambda_{gc} \ell_{gc} \tag{5}$$

Where $\lambda_{ts}, \lambda_{co}$, and λ_{gc} represent the balance weights for the different losses.

B. Mask-Based Grasp Pose Detection (Mask-GPD)

Mask-GPD is based on GPD [29], and the major changes are as follow:

1) Instead of sampling picking candidates globally, we constrain the sampled area by the mask of targets.
2) In the evaluation of grasp pose, we use the following formula S_g:

$$S_g = 1 - \left(\frac{d_g}{d_m} + \frac{\theta_g}{\theta_m} \right), \tag{6}$$

where d_g is the distance between the center of target and picking approaching line. d_m is the maximum distance between the center of target and its surface. θ_g is the angle between the approaching line and gravity, θ_m is the maximum limit of θ_g.
3) We choose the pose with the highest S_g as the best grasp pose.

C. Mask-Based Suction Pose Detection (Mask-SPD)

The key of a successful suction is whether a seal can be formed between a suction cup and a target object surface, and Mask-based suction pose detection is based on that. We exclude the points on the uneven area by calculating the angle between norms of point Pi and its neighborhood:

$$\begin{cases} p_i \in \mathbb{C} & \text{if } \alpha > \alpha_m, \\ p_i \in \mathbb{Z} & \text{otherwise,} \end{cases} \tag{7}$$

where $\mathbb{C}$, $\mathbb{Z}$ is the sets of suction candidates and excluded points, α_m is the maximum limit of α. Then, we evaluate each point in $\mathbb{Z}$ as follows:

$$S_s = 1 - \left(\frac{d_s}{d_{sm}} + \frac{d_b}{d_{bm}} + \frac{\theta_s}{\theta_{sm}} \right), \tag{8}$$

where d_s is the distance between center of target and suction pose. d_{sm} is the maximum distance between center for the object and its surface. d_b and d_{bm} are the distances between suction candidates to the closest and furthest excluded points. θ_s is the angle between suction approaching line and gravity, θ_{sm} is the maximum limit of θ_s. Finally, we choose the pose with the highest S_s as the best suction pose.

V. Dataset

To train and test POIS, we build a dataset called POIS Dataset. The training set comprises only synthetic data, and the testing set comprises both synthetic data and real data. Fig. 3 contains some examples in POIS Dataset. Notice that the synthetic data includes RGB, depths, instance segmentation masks, and ground truth picking policy, and the real data only contains depths because our real evaluation does not rely on annotative information.

A. Synthetic Dataset Generation

Based on our task, we generate a synthetic dataset of 6000 scenes that features the following:

1) A variety of clutteredness across scenes
2) An extensive coverage of 289 daily objects
3) A relatively even division between the numbers of graspable and suckable objects in each scene
4) A test set that consists completely of unseen objects

The dataset is prepared in the BOP format [30]: a scene with N objects produces a RGB image Y_c, a depth image Y_d and a set of pairs of visibility mask and non-occlusion mask $S_M = \{(M_v^k, M_{no}^k)\}_{k=1}^N$. Each pair of visibility mask and non-occlusion mask corresponds to an object.

Our dataset generation follows a three-step procedure: object loading, scene rendering, and mask computation. As an overview of the tools used in the process, the first two steps utilize BlenderProc [31], which can generate photorealistic training image. Comparing to similar solutions [32], [33], BlenderProc provides an outstanding RGB rendering quality and sufficient customizability for the task. We conduct the third step using BOP Toolkit.

To configure a single scene, we first load our environment (workstation) model and select x_k objects ($x_k \sim \mathcal{U}(1, 15)$) from our dataset of 3D PLY models, where $x_k = N_{graspable} = N_{suckable}$. Each object ends up with a 3D poses after dynamic simulation. The next step of rendering an RGB image and a depth image applies domain randomization [34] over lighting colour, light source position, object material, and camera pose for robust transfer from simulation to reality. Lastly, we use the object poses and the camera pose configured in the previous steps to calculate the visibility masks and non-occlusion masks for each object in the scene.

B. Training & Test Set Division

The 6000 scenes consist of 5000 training scenes and 1000 test scenes. It is important to note that we adopt different 3D models for the training and test to evaluate the POIS's performance on novel objects. Specifically, 244 out of the 289 objects are only used in the training set generation. And the rest 45 objects are only used in the test set generation.

C. Ground Truth

The masks and poses of the objects generated by BlenderProc is not directly usable for POIS training. The training requires a pair of masks as ground truth labels with a suckable object and a graspable object for this task. Moreover, in order to minimize difficulty of path planning and the displacement of arms, the relative location of targets is critical.

To find the target pair, we use Mask-GPD and Mask-SPD proposed in Sec.IV-B and Sec.IV-C. We first evaluate the scores of grasping S_{mg} and scores of suction S_{ms} for each object as follows:

$$S_{mg} = d_l + r_v \tag{9}$$

$$S_{ms} = d_r + a_m, \tag{10}$$

where d_l is the distance between the center of mask and the left reference point, d_r is the distance between the center of mask and the right reference point, a_m is the visible area of mask, and r_v is the visibility of mask (visible area / top view surface area). d_l, d_r, a_m, r_v are normalized to $[0, 1]$.

We rank S_{mg} and S_{ms} ascendingly to get list L_s and list L_g. Then, we take the pair of the best suction mask and the best grasping target as ground truth for the training process.

D. Real Data

To evaluate POIS's real-world performance and its ability to work on novel objects, we collect 100 real scenes. We use a high-resolution Photoneo PhoXi industrial sensor (1032 × 772 with 0.05 mm depth precision) to obtain depths in real space and crop them (to 640 × 480). With the calculated camera configuration, we can rebuild the 3D structure from the cropped depth image.

The objects in these bins were sampled from a set of 50 novel objects with highly-varied geometries, including fruits, vegetables, toys, and daily necessities. The number of objects is a random integer in $[1, 50]$.

VI. Experiment

POIS is trained for 30K iterations with Adam [35], with an initial learning rate of 1e-4. In this section, we use a batch size of 8, with $\lambda_{ts} = 3$, $\lambda_{co} = 5$, $\lambda_{gc} = 1$. All images have a resolution of H = 480, W = 640.

A. Metrics

Instance segmentation is partitioning each object from each other, and its performance is conventionally largely evaluated by accuracy and sharpness. However, a sharp segmentation does not necessarily lead to a successful pick in our task. The goal of POIS is to find the most appropriate

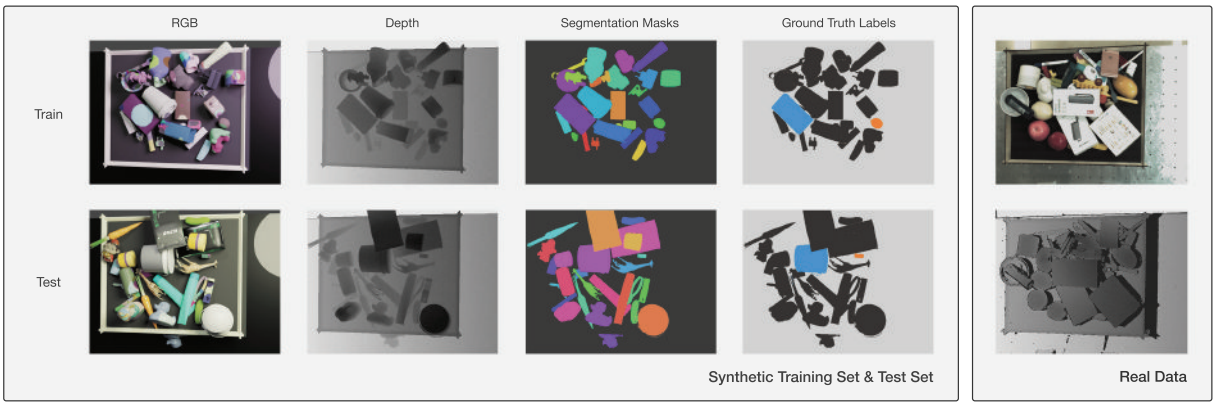

Fig. 3: Some examples in POIS dataset. Training set and testing set contain different set of objects in order to test the segmentation of unknown object. Real data is used only for testing the performance.

targets and the corresponding picking policy. Therefore, we propose metrics that evaluate ambidextrous performance on grasping with either synthetic or real depth input.

1) Omission rate P_0: When the predicted number of objects is less than 2, we believe there is an omission. The omission rate is:

$$P_o = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(\frac{2-n_p}{2}\right), \tag{11}$$

where n_p is the estimated number of objects, N_t is the number of images in testing set.

2) Incorrect detection rate: The rate of recognizing non-graspable to graspable object. After obtaining grasping object masks, we use Mask-GPD to generate the grasping pose and its confidence S'_g. If $S'_g < S_{gm}$ (S_{gm} is the minimum limit of confidence, $S_{gm} = 0.4$ in our experiment), we believe the object is non-graspable. Similarly, S'_s, S_{sm} are the confidence of suction pose (calculated with Mask-SPD) and the corresponding minimum ($S_{sm} = 0.6$ in our experiment). Mathematically, the incorrect detection rate is:

$$P_e = \frac{1}{2N_t}\sum_{i=1}^{N_t}\left(I_g^{(i)} + I_s^{(i)}\right), \tag{12}$$

$$I_{\{g,s\}}^{(i)} = \begin{cases} 1 \text{ if } S_{\{g,s\}}^{\prime(i)} < S_{\{g,s\}m}, \\ 0 \text{ otherwise .} \end{cases} \tag{13}$$

3) Successful paralleled picking rate P_d: A successful paralleled picking requires: a) $n_p = 2$, b) $S'_g > S_{gm}$, c) $d_{gs} > d_m$. Where d_{gs} is the distance between grasping target and suction target, and d_m is the minimum of d_{gs}. The successful paralleled picking rate is defined as:

$$P_d = \frac{N_s}{N_t}, \tag{14}$$

where N_s is the count of successful paralleled picking.

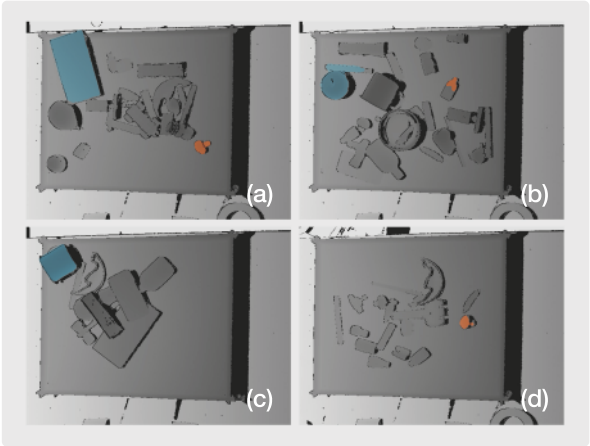

Fig. 4: Four representative segmentation result (a) accurate segmentation of target objects (b) over segmentation of target objects (orange) and under segmentation of target objects (blue) (c) segmentation of suction target when grasp target is not available (d) segmentation of grasp target when suction target is not available.

4) Average Score of paralleled picking P_a:

$$P_a = \frac{1}{N_t}\sum_{i=1}^{N_t}(\alpha_s S'_s + \alpha_g S'_g + \alpha_s\alpha_g\alpha_{gs}d'_{gs}) \tag{15}$$

$$\alpha_s, \alpha_g, \alpha_{gs} = \begin{cases} 1 & \text{if } S'_s > S_{sm}, S'_g > S_{gm}, d'_{gs} > d_m, \\ 0 & \text{otherwise,} \end{cases} \tag{16}$$

where d'_{gs} is the normalized d_{gs}.

B. Performance on synthetic and real images

To prove POIS's performance, we evaluate it quantitatively with POIS synthetic dataset and POIS real dataset. Fig. 4 shows the 4 typical segmentation results. The results similar to Fig. 4(b) have a high occurrence in real testing because of the domain gap between synthetic and real data. However, Mask-GPD and Mask-SPD still generate robust picking poses, which indicates that a successful pick does not rely on sharp segmentation. Fig. 5 demonstrates a quantitative comparison between the segmentation results on synthetic

scenes and real scenes, and it is based on a set of metrics in VI-A. The result shows that POIS has good generalization on Sim-to-Real.

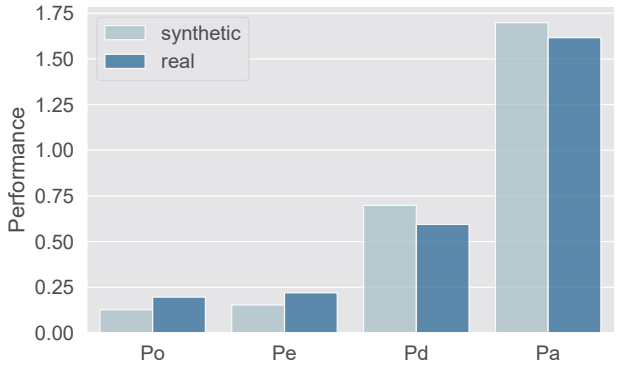

Fig. 5: Performance in synthetic data and real data. Notice that lower values on P_o, P_e and higher values on P_d, P_a are preferable.

C. Comparison with baseline policies

In POIS, we choose target masks from candidate instance masks with a policy confidence map. To prove the advantage of using the policy confidence map in decision making, we build another decision-making algorithm that decides picking policies based only on the spatial distribution of objects (the spatial distribution of objects is almost the determinant factor for paralleled picking policy decision-masking). We use this algorithm as the baseline algorithm. We chose the leftmost mask in the candidate instance masks as the suction target and the rightmost mask as the grasp target.

The experiment shows that POIS with Policy confidence map obtains results that are closer to ground truth, as compared to the baseline algorithm, in both real (Table.I) and synthetic dataset (Table.II).

TABLE I: Comparison in synthetic data

	P_o	P_e	P_d	P_a
baseline	0.196	0.2195	0.594	1.616
POIS	**0.127**	**0.153**	**0.698**	**1.698**
ground truth	0.085	0	0.82	1.761

TABLE II: Comparison in real data

	P_o	P_e	P_d	P_a
baseline	0.447	0.453	0.28	1.259
POIS	**0.227**	**0.237**	**0.546**	**1.678**

D. Robotic Application

We perform policy oriented instance segmentation for unseen objects in real cluttered scenes on an ABB YuMi IRB1400 robot platform with a plastic parallel-jaw gripper and a silicone suction cup (as shown in Fig. 1(b)). The gripper has a maximum width of 40mm and a length of 35 mm, and the suction cup has a radius of 7.5 mm, and a high-resolution Photoneo PhoXi industrial sensor (1032x772 with 0.05 mm depth precision) is set 1300mm above the table.

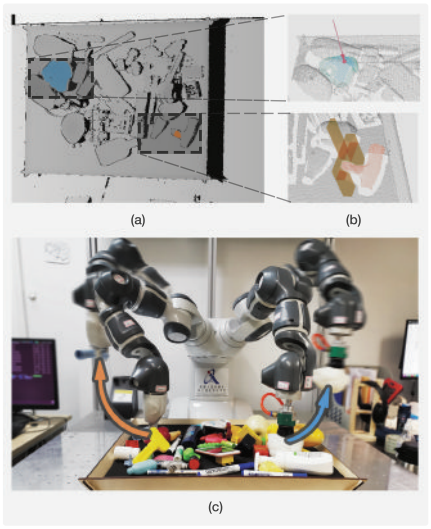

Fig. 6: POIS in clutter. (a) POIS calculates the target masks using depth input. (b) Mask-GPD and Mask-SPD use the corresponding masks to detect picking poses. (c) Following the detected poses, the robot picks and removes the targets from the bin.

As shown in Fig. 6, the experimental results show that POIS has decent compatibility with our ambidextrous robot.

VII. Conclusion

In this paper, we propose a policy-oriented instance segmentation method (POIS) for ambidextrous robot picking, which can output a pair of target masks for paralleled picking. We build a new dataset that contains 6k synthetic scenes and 100 real scenes for policy-oriented instance segmentation. Our method can be trained on synthetic data and achieves decent policy-oriented instance segmentation performance for unseen objects in real cluttered scenes. It may generate inaccurate segmentation result in heavily-occluded scenes, and we plan to make further refinement with RGB images.

References

[1] P. Yang, Z. Zhao, and Z.-J. M. Shen, "A flow picking system for order fulfillment in e-commerce warehouses," *IISE Transactions*, no. just-accepted, pp. 1–23, 2020.

[2] A. Schuster, M. Kupke, and L. Larsen, "Autonomous manufacturing of composite parts by a multi-robot system," *Procedia Manufacturing*, vol. 11, pp. 249–255, 2017.

[3] M. R. Pedersen, L. Nalpantidis, R. S. Andersen, C. Schou, S. Bøgh, V. Krüger, and O. Madsen, "Robot skills for manufacturing: From concept to industrial deployment," *Robotics and Computer-Integrated Manufacturing*, vol. 37, pp. 282–291, 2016.

[4] B. Kumar, L. Sharma, and S.-L. Wu, "Job allocation schemes for mobile service robots in hospitals," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2018, pp. 1323–1326.

[5] S. García, D. Strüber, D. Brugali, A. Di Fava, P. Schillinger, P. Pelliccione, and T. Berger, "Variability modeling of service robots: Experiences and challenges," in *Proceedings of the 13th International Workshop on Variability Modelling of Software-Intensive Systems*, 2019, pp. 1–6.

[6] T. Nakamura, A. Yuguchi, A. Maël, G. A. G. Ricardez, J. Takamatsu, and T. Ogasawara, "Ontology generation using gui and simulation for service robots to operate home appliances," in *2019 Third IEEE International Conference on Robotic Computing (IRC)*. IEEE, 2019, pp. 315–320.

[7] S. Hasegawa, K. Wada, Y. Niitani, K. Okada, and M. Inaba, "A three-fingered hand with a suction gripping system for picking various objects in cluttered narrow space," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1164–1171.

[8] H. Nakamoto, M. Ohtake, K. Komoda, A. Sugahara, and A. Ogawa, "A gripper system for robustly picking various objects placed densely by suction and pinching," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 6093–6098.

[9] Y. Deng, X. Guo, Y. Wei, K. Lu, B. Fang, D. Guo, H. Liu, and F. Sun, "Deep reinforcement learning for robotic pushing and picking in cluttered environment," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 619–626.

[10] H. S. Stuart, S. Wang, and M. R. Cutkosky, "Tunable contact conditions and grasp hydrodynamics using gentle fingertip suction," *IEEE Transactions on Robotics*, vol. 35, no. 2, pp. 295–306, 2018.

[11] G. Ponraj Joseph Vedhagiri, A. V. Prituja, C. Li, G. Zhu, N. V. Thakor, and H. Ren, "Pinch grasp and suction for delicate object manipulations using modular anthropomorphic robotic gripper with soft layer enhancements," *Robotics*, vol. 8, no. 3, p. 67, 2019.

[12] J. Borràs, G. Alenya, and C. Torras, "A grasping-centered analysis for cloth manipulation," *IEEE Transactions on Robotics*, 2020.

[13] D. Liang, J. Song, W. Zhang, Z. Sun, and Q. Chen, "Pasa hand: A novel parallel and self-adaptive underactuated hand with gear-link mechanisms," in *International Conference on Intelligent Robotics and Applications*. Springer, 2016, pp. 134–146.

[14] N. Correll, K. E. Bekris, D. Berenson, O. Brock, A. Causo, K. Hauser, K. Okada, A. Rodriguez, J. M. Romano, and P. R. Wurman, "Analysis and observations from the first amazon picking challenge," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 172–188, 2016.

[15] 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 *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 1386–1383.

[16] D. Morrison, A. W. Tow, M. Mctaggart, R. Smith, N. Kelly-Boxall, S. Wade-Mccue, J. Erskine, R. Grinover, A. Gurman, T. Hunn *et al.*, "Cartman: The low-cost cartesian manipulator that won the amazon robotics challenge," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7757–7764.

[17] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, 2019.

[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," in *Conference on robot learning*. PMLR, 2020, pp. 1369–1378.

[19] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.

[20] C. Mitash, R. Shome, B. Wen, A. Boularias, and K. Bekris, "Task-driven perception and manipulation for constrained placement of unknown objects," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 5605–5612, 2020.

[21] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2901–2910.

[22] 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 *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7283–7290.

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

[24] S. V. Shah, I. Sharf, and A. Misra, "Reactionless path planning strategies for capture of tumbling objects in space using a dual-arm robotic system," in *AIAA Guidance, Navigation, and Control (GNC) Conference*, 2013, p. 4521.

[25] Y. Choi, D. Kim, S. Hwang, H. Kim, N. Kim, and C. Han, "Dual-arm robot motion planning for collision avoidance using b-spline curve," *International journal of precision engineering and manufacturing*, vol. 18, no. 6, pp. 835–843, 2017.

[26] Y. Fei, D. Fuqiang, and Z. Xifang, "Collision-free motion planning of dual-arm reconfigurable robots," *Robotics and Computer-Integrated Manufacturing*, vol. 20, no. 4, pp. 351–357, 2004.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[29] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[30] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis *et al.*, "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 19–34.

[31] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv preprint arXiv:1911.01911*, 2019.

[32] M. Gschwandtner, R. Kwitt, A. Uhl, and W. Pree, "Blensor: Blender sensor simulation toolbox," in *International Symposium on Visual Computing*. Springer, 2011, pp. 199–208.

[33] T. To, J. Tremblay, D. McKay, Y. Yamaguchi, K. Leung, A. Balanon, J. Cheng, W. Hodge, and S. Birchfield, "NDDS: NVIDIA deep learning dataset synthesizer," 2018, https://github.com/NVIDIA/Dataset_Synthesizer.

[34] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 23–30.

[35] K. Da, "A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.