Segmentation-Based Approach for a Heuristic Grasping Procedure in Multi-Object Scenes

Segmentation-based approach for a heuristic grasping procedure in multi-object scenes

Davide Ceschini
DET - Politecnico di Torino
Torino, Italy
ceschinidavide@gmail.com

Riccardo De Cesare
Comau S.p.A.
Grugliasco (TO), Italy
riccardo.decesare@comau.com

Enrico Civitelli
Cognitive Robotics - Comau S.p.A.
Grugliasco (TO), Italy
enrico.civitelli@comau.com

Marina Indri
DET - Politecnico di Torino
Torino, Italy
marina.indri@polito.it

***Abstract*—Object manipulation in unstructured environments is important for many industrial applications where the items vary in shape, size, and material. This paper introduces a two-step pipeline for object picking, which combines instance segmentation with a heuristic based grasp point selection. The grasping points are determined using the 2D segmentation masks and depth images. A voxel-downsampling procedure reduces the depth noise, and the Theil-Sen algorithm ensures a robust linear regression for the grasping attitude determination. Unlike other methods, our approach does not require extensive training, as well as a fine labelled dataset for picking, and hence it is also independent of object shapes. Using SAM's ViT-h version and a binary object detector trained on a large dataset, our method is robust and class agnostic. The experiments, made using a RealSense D435i camera and a Racer 3 manipulator, show that our pipeline has a good success rate in simple and moderately complex scenarios, balancing computational efficiency and accuracy.**

***Index Terms*—Object picking, multi-object scenes, segmentation, grasp point selection.**

I. Introduction

The object manipulation task in unstructured environments has become important in many industrial fields and logistics applications, where objects vary in shapes, dimensions and material, and are often highly cluttered. This problem encompasses several areas, such as object recognition, object localization, grasp determination and motion planning, and it has challenged researchers over the years to develop computationally efficient and robust approaches, nowadays predominantly based on Deep Learning. Among the picking solutions, two categories arise: single-step and multi-step methods. The former directly provide the grasp's pose based on some quality metric, such as Dex-Net 2.0, 3.0, and 4.0 [1]–[3], thus being computationally efficient. However, the source of errors can be hardly identified, because the action is performed in only one step; in addition, they require the datasets to be well and densely annotated, which can be hard to find. Thus, multi-step pipelines split the problem by combining the segmentation of the single instances with various grasping pose determination solutions.

Since these approaches can be seen as the stack of object localization, pose estimation and grasp determination, various methods commonly employed to tackle the three sub-tasks have been proposed in literature. In [4], [5] instance segmentation and grasp determination are combined in an end-to-end fashion, by providing as output the segmentation mask of the first object to be picked together with the grasping pose. This is done by using the predicted grasps as point proposals for the instance segmentation. However, they require the training of both the segmentation and the grasping networks, which can be hard and time consuming. On the contrary, some solutions (e.g., [6]) employ a foundation object detector model based on DINO [7], specifically tuned on the dataset of interest, to be used as input prompt to a foundation class agnostic segmentation model as SAM [8], prior to the grasp determination network. These segmentation models provide only the difference between objects and background, which is not sufficient for some applications. A possible strategy is to exploit semantic segmentation models, which however require to be trained on large scale datasets to become category robust. As a result, other approaches make use of few shot semantic segmentation architectures, which provide both the segmentation masks and the category information, just by training on a few images per category, as in [9]. Another challenging problem concerns the random nature of the objects, which prevents the picking of every item with a single gripper. Some methods (e.g., [10]) have been developed to use a grasping module that accounts for different grippers (vacuum, two finger, magnetic). The module proposes several candidates for each gripper: the one with the highest confidence score is chosen. However, all the cited methods require a trained grasping pose determination model as part of the pipeline, so that it is difficult to extend their application to unseen shapes.

This paper proposes an approach based on a two-step pipeline, taking inspiration from [11]. Splitting the task into segmentation and picking can be beneficial, since the former is a well studied problem with many available datasets. Picking is performed on a heuristic basis that does not require training on a specific dataset, and it is improved with experimental considerations. The segmentation employs SAM, which boosts its robustness, together with a binary object detector trained on a large scale dataset that makes it robust to unseen objects and class agnostic; in addition, the choice of the grasping point accounts for the randomness of the shapes. The developed pipeline is implemented in the case of a manipulator equipped with a suction effect gripper to evaluate its effectiveness in practice. The carried out tests show that the proposed sequence has a good success rate in some simple to moderately complex scenarios, and that it can be a good trade-off between computational complexity and accuracy for some practical applications.

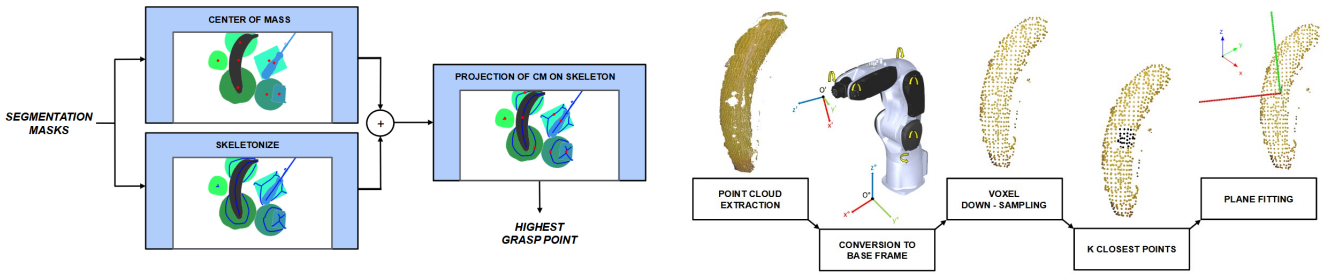

Fig. 1: Sketch of the pipeline. The figure on the left shows that starting from the segmentation masks the grasping points are determined for every object. According to the desired metric, the first object is picked and its point cloud is extracted, as shown in the right figure; then it undergoes the described process to calculate the picking attitude.

II. The Proposed Method

A. Pipeline

The proposed picking method consists of a two-step pipeline: instance segmentation followed by a heuristic choice of the grasping point. The instance segmentation is performed with the Mask R-CNN [12] implementation of the Torchvision library. The segmentation core is further improved with ViT-h version of SAM by using the found bounding boxes as prompts. The grasping point is chosen on the basis of the 2D segmentation masks and the depth image. The former's center of gravity is calculated and is then projected onto its skeleton, computed according to [13]. In fact, while the center of gravity can be properly used to pick up a convex object, since it is certainly located within the object itself, in the case of a non-convex object it may fall outside its contour, and hence its subsequent projection allows to select the closest internal point. The incompleteness of the depth information is accounted for by superimposing it with the masks when calculating the skeleton. To determine the proper gripper orientation for picking, the extracted point cloud is first down-sampled using a voxel-downsampling procedure. The algorithm divides the space into bins of equal dimensions, and considers the center of gravity of all the points inside each bin as its representative point. This way, the noise of the acquired point cloud is reduced, while keeping its meaningful information regarding the object shape; such optimization has been revealed important for the determination of the grasping orientation, especially for planar surfaces. A KNN is then performed to find the K-closest point to the chosen one to subsequently apply the Theil-Sen algorithm, which is an outlier robust estimator for linear regression, thus finding the approaching attitude. The object that is first picked is the one with the highest coordinate in the base frame, since it is thought to be the less occluded, hence the easiest to extract. Further filters are also implemented to account for the real working conditions: pose out of range errors and picking of the bin. Figure 1 shows a sketch of the pipeline.

B. Base Network Training

The dataset used for the training and testing of Mask R-CNN is the combination of Suctionnet [14] and ARMBench [15]. The large variety of objects shapes, colors and materials made the base network robust to unseen items. At first, two separate models were trained on the distinct datasets. These were then evaluated on the reciprocal test's subsets, and the best one was chosen to initialize the combined training. The resulted network performance was first checked on the combined test dataset. Table I provides the results of the models trained on Suctionnet, ARMBench and their combination when inferred on the union of the test of ARMBench and Test Novel of Suctionnet. The latter contains object categories that are very dissimilar from the ones of its training subset.

Dataset	AP_{75}	AP
Suctionnet	0.29	0.27
ARMBench	0.36	0.34
Combined	**0.59**	**0.53**

TABLE I: Average Precision (AP) of Mask R-CNN calculated on the segmentation masks. AP refers to the mean of the AP from thresholds 50 to 95 % with 1% step size. AP_{75} is the AP calculated with threshold 75%.

It is noticeable that the combined training improves the performances of the network; the corresponding model was then chosen for the experimental tests.

III. Experiments

A. Physical Setup

The experimental tests were carried out on a subset of the YCB dataset [16] and relied on a RealSense D435i camera, which captures both the RGB and depth images using stereo vision. The setup includes a COMAU Racer-3 manipulator and a computer equipped with Intel Core i7-1165G7 and NVIDIA RTX-A5000.

B. Results

The experiments have been carried out by manually grouping the objects into four clusters that were strongly influenced by the physical properties of the objects. The four clusters can be classified as easy, medium, hard and transparent/translucent. For each category two scenarios have been investigated: single object and cluttered objects. The former included items separated from each other, and it was used to test the capability of the network to detect and pick the single object. The latter scenes were made up of many objects but in a cluttered context, which is the typical working condition, in order to

determine the success rate of the pipeline. Each scene has been evaluated also with SAM; for a matter of repeatability, the SAM's scenes were made as close as possible to the Mask R-CNN ones, however small differences, that had negligible impact, were present.

1) ***Easy****:* The easy category contains objects of regular convex shape and large sizes (e.g., containers, boxes), hence it was not problematic to detect and pick them. From the detection and segmentation results, no critical aspects arose. Regarding the picking quality outcomes, the grasping points were often valid, except for some of the scenes where the high inclination of the corresponding item was responsible for making the grasping point fall on the object's edge. Figure 2 shows an example of the objects tested and the capability of the pipeline to empty the box.

Fig. 2: Picking timeline with SAM for the Easy scenario. An infinity cycle was performed during which the pipeline was shown to empty the box several times. The red dots represent the picking points.

2) ***Medium****:* This bin was made up of smaller objects with more irregular shapes (e.g., round elements, plastic bags, scissors); as a result, picking was more difficult. In fact, despite a good scene segmentation, some of the objects had to be removed by the user. This was due to either the small surface area for the suction cup (e.g., scissors, cutlery) or to the suction effect that did not work on the material under investigation (plastic, textile). Figure 3 shows the pipeline execution.

3) ***Hard****:* This category included objects of high irregular shapes, making it the one with the highest failure rate. In particular, given the complex shapes and colors of some objects, the over-detection led to parts segmentation problems, hence to multiple grasping choices. Moreover, in some situations certain objects were not detected at all. However, the use of SAM turned out to be a good choice, improving the quality of the segmentation, which was more evident than in the other cases. But the shape of many objects made the heuristic fail in most cases. Figure 4 provides some examples of the errors described.

4) ***Transparent/Translucent****:* For this category it was noticed how the light effects may strongly influence the detection and segmentation part, and whenever the object was detected its depth information was either missing or noisy due to the camera, thus preventing in any case the completion of the picking action. This can be seen in Figure 5.

A qualitative example of improvements provided by SAM on the final picking can be found in Figure 6.

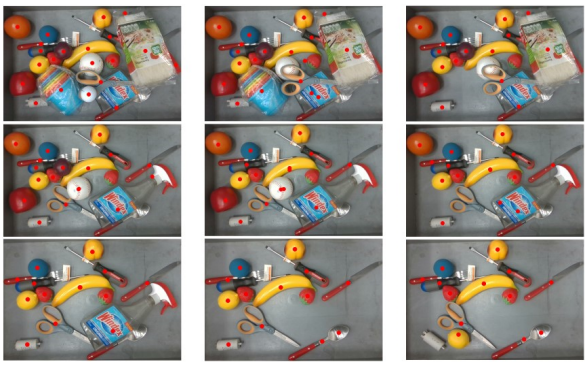

Fig. 3: Picking timeline with SAM for the Medium scenario. The pipeline capability was tested with an infinity cycle, however, with respect to the easy scenario, some of the objects remained untakeable.

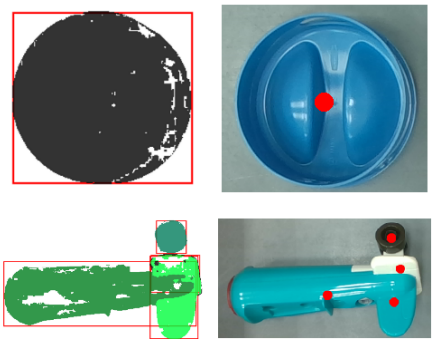

Fig. 4: Errors in the Hard scenario with SAM. The figures above show how the irregular shape caused the choice of the grasping point based on the 2D mask to be erroneous. The figures below display the part segmentation errors due to the multiple colors. The empty areas on the segmentation represent the missing depth.

Tables II and III report quantitative results of the carried out experiments, in terms of success rate for Mask R-CNN and SAM pipeline, respectively. The Hard and Transparent categories are not present due to their high failure rate. The numerical values reported reflect the previous analysis.

It can be seen that the pipeline performs well in the Easy scenario, with minor difficulties in the Medium one that are not caused by the method itself, but mainly due to the gripper, which cannot pick objects with a small surface area or made of certain materials.

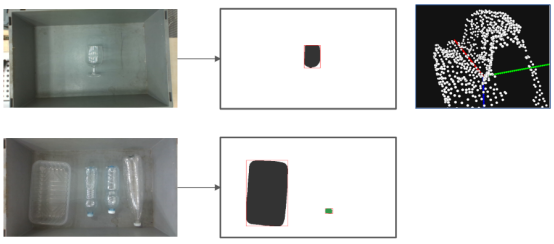

Fig. 5: Scene segmentation for transparent objects.

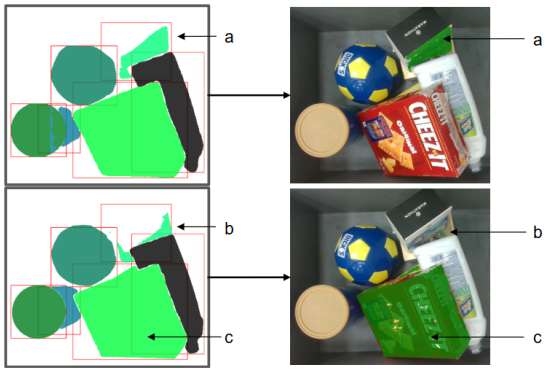

Fig. 6: SAM vs Mask R-CNN. The figure above shows the segmentation masks of Mask R-CNN, while the one below is obtained with SAM by using as input prompts the bounding boxes that Mask R-CNN provides together with its own masks. Mask R-CNN segments the edge of object a) and considers it the first to be picked, thus preventing its picking since the grasping point would fall on the edge. Instead SAM is capable of distinguishing the object below, b), hence leading to the successful picking of c).

Scene	**N°pickings**	**SR (%)**	$t_{CPU}[s]$	$t_{GPU}[s]$
Easy	46	82.6	5.5	0.46
Medium	78	60.3		

TABLE II: Success rate per category with Mask R-CNN prediction and computational times.

Scene	**N°pickings**	**SR (%)**	$t_{CPU}[s]$	$t_{GPU}[s]$
Easy	42	92.8	43.2	0.50
Medium	57	78.9		

TABLE III: Success rate per category with SAM prediction and computational times.

IV. Conclusions

In this work, a two-step picking pipeline is presented. It is based on a heuristic grasping pose determination algorithm, with some experimental improvement, which does not require training, hence heavily annotated grasping datasets. The segmentation core can either use the SAM refinement, with great improvements, or the base Mask R-CNN, which was binary trained on a moderately large dataset, making it robust to unseen shapes and category agnostic. The Mask R-CNN based method has been shown to be a good trade-off between accuracy and computational costs in a simple to medium scenario, however both the object detection and the grasping algorithm showed problems when moving to high complex setups. In fact, the choice of the grasping point based on the 2D segmentation suffered from the irregular shapes of the objects. Furthermore, the multiple colors, light contrasts accompanied with unseen shapes caused segmentation errors, such as missed detection and part segmentation, especially in highly cluttered scenarios. Thus, more robust object detectors as well as grasp determination method should be further explored.

V. Acknowledgement

The authors would like to express their gratitude to Luca Di Ruscio, Simone Panicucci, and the Comau team for their invaluable assistance in setting up this project.

This study was carried out within the FAIR - Future Artificial Intelligence Research and received funding from the European Union Next-GenerationEU (PIANO NAZIONALE DI RIPRESA E RESILIENZA (PNRR) – MISSIONE 4 COMPONENTE 2, INVESTIMENTO 1.3 – D.D. 1555 11/10/2022, PE00000013). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

References

[1] 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, eaau4984, 2019.

[2] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-net 3.0: Computing robust robot suction grasp targets in point clouds using a new analytic model and deep learning," *arXiv preprint arXiv:1709.06670*, 2017.

[3] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[4] S. Ainetter, C. Böhm, R. Dhakate, S. Weiss, and F. Fraundorfer, "Depth-aware object segmentation and grasp detection for robotic picking tasks," *arXiv preprint arXiv:2111.11114*, 2021.

[5] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 452–13 458.

[6] J. Li and D. J. Cappelleri, "Sim-suction: Learning a suction grasp policy for cluttered environments using a synthetic benchmark," *IEEE Transactions on Robotics*, vol. 40, pp. 316–331, 2024.

[7] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang, "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," *arXiv preprint arXiv:2303.05499*, 2023.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[9] X. Liu, Y. Zhang, and D. Shan, "Unseen object few-shot semantic segmentation for robotic grasping," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 320–327, 2022.

[10] S. D'Avella, A. M. Sundaram, W. Friedl, P. Tripicchio, and M. A. Roa, "Multimodal grasp planner for hybrid grippers in cluttered scenes," *IEEE Robotics and Automation Letters*, vol. 8, no. 4, pp. 2030–2037, 2023.

[11] J. Dirr, A. Siepmann, D. Gebauer, and R. Daub, "Evaluation metric for instance segmentation in robotic grasping of deformable linear objects," *Procedia CIRP*, vol. 120, pp. 726–731, 2023.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 2961–2969.

[13] W. Abu-Ain, S. N. H. S. Abdullah, B. Bataineh, T. Abu-Ain, and K. Omar, "Skeletonization algorithm for binary images," *Procedia Technology*, vol. 11, pp. 704–709, 2013.

[14] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "Suctionnet-1billion: A large-scale benchmark for suction grasping," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8718–8725, 2021.

[15] C. Mitash, F. Wang, S. Lu, V. Terhuja, T. Garaas, F. Polido, and M. Nambi, "Armbench: An object-centric benchmark dataset for robotic manipulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9132–9139.

[16] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, p. 027836491770071, 04 2017.