Detection-driven 3D Masking for Efficient Object Grasping

LULU LI (lulu.li@utt.fr)
Universite de Technologie de Troyes

Abel CHEROUAT

Hichem SNOUSSI

Ronghua HU

Tian WANG

Research Article

Keywords: Object Detection , Priority Grasping , Grasping Prediction , GraspNet

Posted Date: April 26th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2781544/v1

Version of Record: A version of this preprint was published at The International Journal of Advanced Manufacturing Technology on November 9th, 2023. See the published version at https://doi.org/10.1007/s00170-023-12574-9.

Abstract

Robotic arms are currently in the spotlight of the industry of future, but their efficiency faces huge challenges. The efficient grasping of the robotic arm, replacing human work, requires visual support. In this paper, we first propose to augment end-to-end deep learning gasping with a object detection model in order to improve the efficiency of grasp pose prediction. The accurate positon of the object is difficult to obtain in the depth image due to the absent of the label in point cloud in an open environment. In our work, the detection information is fused with the depth image to obtain accurate 3D mask of the point cloud, guiding the classical GraspNet to generate more accurate grippers. The detection-driven 3D mask method allows also to design a priority scheme increasing the adaptability of grasping scenarios. The proposed grasping method is validated on multiple benchmark datasets achieving state-of-the-art performances.

1 Introduction

Grasping prediction refers to the prediction of robot's movements and postures when grasping objects, which aims to enable the robot to grasp objects more accurately and efficiently. It is highly relevant in a variety of fields such as automated logistics, smart homes, healthcare, ser- vices, and defense applications [1, 2, 3], where efficient grasp prediction is needed to improve the robot's ability to complete transport tasks. Indeed, some results have been achieved in computer vision for grasp prediction. In an early work, Miller et al. [4] have used heuristic rules to generate and evaluate three-fingered grasps, modelling objects as a set of shape primitives for grasp prediction based on simple shapes such as spheres, cones and cylin- ders. Yun Jiang et al. [5] have used a rectangular represen- tation approach for learning target grasps on RGB-D images [6]. Other methods focus on using edges and contours to determine shapes and force closures to grasp 2D planar objects. Piater [7] uses K-means clustering to estimate 2D grasping directions for simple objects, in particular squares, triangles and circular 2D 'blocks' [5]. In real-world grasping, however, these shape-based methods are inaccurate to classify the objects into com- plete 3D shapes, even less to define the grasping prim-itives. In recent years, grasp detection methods based on 6D pose estimation [8–14] have been proposed for gras-ping regression. Other methods aim at obtaining grasp data directly from the sensors of the robot arm without estimating the object's pose [15, 16, 17]. Regression-based 6D pose estimation requires a large amount of high-quality training data and a complex training pro-cess. Similarly, the acquisition, calibration and labelling of robotic arm sensor data is a very complex process and therefore not widely available. The basic idea is to deal with grasp perception similarly to object detection in computer vision. Although these methods generalize well to the acquisition of knowledge of new objects, they have not yet proven to be efficient and reliable eno-ugh [15]. Furthermore, it is not possible to effectively recognize the best grasping pose for all objects in a multi-object scene.

To avoid these problems and to balance object pose estimation and grasp effectiveness, we propose a detec-tion-driven 3D masking method in order to enhance gras-ping efficiency. In summary, we make following contri-butions:

1) We propose to generate the grasp pose problem as a multi-model perception problem. Taking the RGB and depth image as the input for the robot arms system. The object detection method is adopted for fast target locali-zation.

2) Contribute a benchmark dataset. Our dataset cont-ains 890 RGB-D images with hundreds of scenes. Whi-ch is able to evaluate the effectiveness of the method.

3) We use the target detection YOLO deep learning method for fast target localization and assessment of grasp ease. Using target position and priority informa-tion are then used for accurate grasp prediction. Instead of the regression-based pose estimation method in the second step, we use the GraspNet network in order to predict and score grasps based on the masked point cloud. This detection-driven 3D mask method is validat-ed on multiple benchmark datasets.

The proposed method reduces the computational effort of 3D point cloud segmentation and matching, and im-proves computational efficiency and prediction accu-racy. Furthermore, the proposed priority grasping met-hod enhances scene adaptation.

2 Related Works

2.1 Object detection

Object detection consists of identifying and locating objects in a visual scene. By detecting and recognizing objects in images or videos, the robot can determine the object position, the category of the object well as esti-mating the grasping pose from the detection informa-tion [18, 19, 20, 21]. Examples related to grasping are PoseCNN [8], Faster R-CNN Inception-V2 [22] and partial depth estimation [23] all of which have achieved good results and are widely applied. The YOLO family [24, 25] is one of the most efficient real-time object detectors. Considering that YOLOv7 version has the highest accuracy, i.e. 56.8% *AP*, and the size of the parameters is 36.9 *M*. The processing time is 161 *FPS* for the 640×640 image, it is suitable for the real time control of the robot. It will be considered in our work in order to improve the efficiency of the grasping prediction.

2.2 Grasp prediction

There have been many approaches for solving the gra- sping problem. Morrison et al. [26] proposed a genera- tive grasping convolutional neural network (GG-CNN) to predict the grasp quality and pose of each pixel. This method is a real-time, object-independent grasp synth- esis method that can be used for closed-loop grasping. AnyGrasp, the visual grasp perception system generat- es spatially dense and temporally smooth grasp poses [27]. Mahler et al. [28] performed grasp evaluation usin-g a convolutional neural network (GQ-CNN) model wit-h parallel plate graspers in single-target scenes. In Dex-Net 3.0, the authors added support for suction-based end-effectors, obtaining higher accuracy in grasp evalu-ation than before [29]. In Dex-Net 4.0, it was proved that objects can be selectively grasped by the arm's suction cups or grippers [16]. Our work is based on GraspNet (an end-to-end dense grasping pose estima-tion network) [30, 31] for effective grasping, which is composed of an approach network, an operation network and a tolerance network. The input to GraspNet is the complete point cloud and the output is the top dense grasping predictions. The disadvantage is that the best grasp pose cannot be estimated for all objects in a multi-object scene. Therefore, we use the detection-driven 3D masking method based on GraspNet to en-hance the effectiveness of the grasp by estimating the best grasp pose for all targets in a multi-object scene. The representation of grasping pose is shown in Fig. 1.

2.3 Grasp prediction metrics

In order to demonstrate the efficiency of the method, the correct prediction of grasping must satisfy [5] the following constraints:

1. The gripper is above the object close to the center of the object.
2. No collision.
3. The predicted gripper width *W* is greater than the object to fit the holder.

Quality estimates of grasping poses are often based on point metrics; rectangular metrics [2, 5]; force closure analysis [30, 31, 32, 33]; Grasp success rate (GSRs) [34] or the geometry of a section cup model through height map space [35] etc. To better assess our detection-driven 3D masking grasp method, here we use improved point metric. As schematically shown in Figure 1. We predict a gripper [30, 31], which includes the approach vector *V*, the approach distance *D* from the gripper to the origin of the gripper, the in-plane rotation around the approach axis *R* and the gripper width *W*.

Comparing the gripper center point to the grasping point distance from the predicted object surface, we consider it as a correct grasping prediction if it is within a certain threshold. It is worth noting that the Jiang Y indicated point metric consists of comparing the center of the rectangle to the ground-truth distance, and is not always reliable, as the prediction does not take the orientation into account. In contrast, our predicted grasping pose includes the direction, so there is no such problem in our approach.

Table 1. The experimental environment configuration and camera type, parameters are described, where the factor depth is the scale of depth converted to meters, Fx Fy Cx Cy are the camera focal length and the optical center internal parameters.

Item	Value or name
Training environment	5 cartes GPU Tesla T4 (16Go) Ubuntu 20.04, CUDA 11.3, Pytorch 1.10, python 3.6
Test and evaluation environment	Ubuntu 20.04, CUDA 11.3, Pytorch 1.10, python 3.7
Camera	RealSense Depth Camera D435i Factor depth 1000 Fx Fy 321.0473327636719 Cx 323.1285400390625 Cy 174.82257080078125

Table 2. Experimental parameters.

Item	Value
Number of training sets	450
Grasping predicted images(unseen)	200
Grasping predicted images(novel)	120
Epoch	300
Initial learning rate	0.01
Final one cycle learning rate	0.1
Batch size	16

Table 3
Evaluation results of the correct pose prediction rate for different grasp priorities. Where Others Round Right represents the order of grasping, i.e. Others, Round-sided objects, Right-sided objects. The rest of the grasp sequences work in a similar way.

Grasping sequence	**Novel**			**Composite**		
	T1(%)	**T2(%)**	**Average(%)**	**T1(%)**	**T2(%)**	**Average(%)**
Others Round Right	90.88	75.81	83.35	93.77	86.85	90.31
Others Right Round	92.98	79.62	86.30	95.63	89.88	92.76
Round Others Right	81.43	70.42	75.93	92.69	86.41	89.55
Round Right Others	79.75	72.75	76.25	94.22	89.16	91.69
Right Others Round	90.10	83.13	**86.62**	97.17	93.29	**95.23**
Right Round Others	86.33	81.61	83.97	96.84	93.01	94.93

3 Proposed Approach

This section is dedicated to the proposed method, including the acquisition and labelling of the dataset the systematic structure of the detection-driven 3D m-asking method and the experimental setup.

3.1 Dataset

We have taken 450 RGB images of 3D printed objects in multi-object scenes for Transfer Learning [36]. Vari-ous shapes of objects such as orange rectangles, blue trapezoids, large blue cylinders, yellow cones, pink co-lumns, irregular objects, etc. have been used. The training dataset consists of 12 different objects, 10 targets per scene. Moreover, 440 images divided into 3 categ-ories with 10 objects per scene were tested on the ima-ges. Of these, 200 are similar scenes and 120 are com-pletely new scenes. Another 120 images (composite dataset) consisting of training objects and unknown targets were used for the ablation study. In the prior training stage, each image is labelled with:

1 The object class;

2 The bounding value of the mask;

3 The grasp priority is defined.

The object classes are simply defined as right-sided objects, round-sided objects and others. The grasping priority is defined as the order of grasping objects. The depth image aligned to the RGB image is added at the grasp prediction stage and used to generate the full po-int cloud. The RGB-D images for all datasets are from the camera RealSense D435i.

3.2 Architecture

Grasp prediction consists of two stages. The first stage is to use YOLOv7 as a priori algorithm to detect the target to be grasped. The second stage sequentially per- forms grasp prediction on the cropped point cloud based on the priority of the grasped object. Here, we introduce priority grasping predictions for more effi-cient grasping. Initially, the priority information of the grasped object is pre-defined in the first stage accord-ing to the object ease of grasping. For example, the grasping order is defined as round-edged objects first, followed by right-edged objects and finally irregular objects. Then priority is calculated based on all object classes and grasping ease scores obtained by YOLOv7 recognition in the image working area. Lastly, grasping prediction is performed based on the grasping priority of the target.

To illustrate the detection-driven 3D mask approach described in this paper, Fig. 2 gives a diagram of the system structure using YOLOv7 and GraspNet. The in- puts are RGB images and predefined object grasping difficulty values. After processing the inputs through YOLOv7, the outputs obtained in the first stage are normalized object classes, bounding boxes, object mas- ks and priority scores. The inputs of the second stage are the RGB-D images and the outputs of the first stage. The point cloud obtained by RGB-D is first cropped into classified grasping regions using a priori informa- tion output by the first stage network.

The grasps are then sequentially predicted by Grasp- Net following the estimated priority and the correspon- ding object masks. All feasible grasp poses are estimat- ed avoiding invalid grasp predictions. The grasp repre- sentation is the same as the GraspNet output.

3.3 Experimental setup

The camera Realsense D435i is used in the experiments

Table 4
The prediction correct results of the grasping of different categories of objects, and their average values. The grasping order in experiment is right-sided, others, round-sided. The dataset of scene seen is selected 110 images. Unseen and Novel scenes select all. *T1 T2* are different thresholds.

Object	Seen		Unseen		Novel	
	T1	T2	T1	T2	T1	T2
Right	99.84	88.70	98.83	90.95	89.16	88.13
Others	99.47	83.04	98.40	81.63	93.30	83.08
Round	98.16	79.66	96.64	77.68	87.85	78.17
Average	99.15	83.80	97.96	83.42	90.10	83.13

which can acquire RGB-D images in real time at resol- utions of 640×360 and frame rates of 30 fps. Table 1 shows the details of the experimental environment. Training is performed using the server Tesla T4 Ubuntu 20.04.5 LTS. Python 3.6, Pytorch 1.10 and CUDA 11.3 are used to train the model of object recognition. Test- ing and evaluation used Python 3.7, Pytorch 1.10, CU-DA 11.3. The camera parameters show the camera dep-th scaling and camera intrinsic parameters. The depth scaling is 1000. Table 2 shows the training parameters. The training dataset is composed of 450 images, but a better model can be obtained using a Transfer Learning method.

The model trained with these parameters is used to obtain efficient grasping poses. As shown in Fig. 3 (1–10), objects with different priorities are in turn succ- essfully predicted with the exact appropriate grasping pose.

4 Results

This section introduces the predicted results and the ablation study.

4.1 Prediction results

The experiments have compared the correct prediction rate of the grasping with different priorities. From Table 3, we can notice that the test set performs dif-ferent results when grasping targets with different priority order. Novel and Composite represent different datasets, with Novel being the dataset with completely unknown objects and scenes for the training model, while scenes in Composite are composites of known objects and completely unknown targets. The values *T1* and *T2* are the different thresholds in the point metric assessment method. Let the distance from the center of the grip point to the grasp point on the object surface be *D*. Let the minimum value in the length and width

Table 5
Comparison of methods.

Methods	Graspnet with YOLO	Graspnet without YOLO
Predict the grasping pose of all objects in the scene	√	×
Object location	√	×
Priority grasping	√	×
Object classification (distinguish categories of objects)	√	×
Binary image segmentation (without classification)	×	√

of the object bounding box be *M*. Then, *T1* is equal to the difference between *D* and one-third *M*. *T2* is equal to the difference between *D* and a quarter of *M*.

The results obtained from different scenarios of the dataset show that the best priority grasping order is right, others, round. Later ablation studies were also performed on this grasping priority order. The results of one priority grasping prediction are given in Fig. 3. The results show that the sequence of object grasping is exactly the order of target priority. Even with similar classes of objects in different scenes the results are the same. This demonstrates the consistency of the detec-tion-driven 3D mask priority grasping method.

To illustrate the efficiency of the method described in this paper, an example of the grasping results is given in Fig. 4. The results give the same number of grippers as the targets. It can be seen from the results that the method described in this paper predicts the best grasping poses for all objects, whether in simple scenes with no occlusion or stacked objects, or in complex scenes with occlusions or stacked objects. The classical grasping method without YOLOv7 cannot predict the best grasping poses for all objects. This shows the ef-fectttiveness of the method described in the paper.

Table 4 gives the average correct prediction rate of objects for each category. The results are based on the grasping priority order: right, others, round. As we can see, the results show that right-sided objects are the best at correct pose prediction and the easiest to grasp while round-sided objects are not as easy to grasp, illu- strating the necessity of the priority grasping method.

4.2 Ablation studies

To assess the efficiency of YOLOv7 as priori algorithm for the detection-driven 3D masking method, we conducted ablation studies. Here, we present the results of grasping poses with and without the prior algorithm YOLOv7 in Fig. 5. The results show dense predicti-

ons of the grasping poses with the 50 highest scoring grippers. In order to demonstrate the efficiency of our method, the top 1 gripper with the highest grasp target score is also given in Fig. 5 (4). As it is clearly

Gras-pNet without YOLOv7 cannot predict the grasping poses for all objects, while our method can accurately predict the best grasps for all objects. Then we compa-red the different methods and the results are reported in Table 5. Our method not only predicts the best grasp pose for all objects in the scene, but also locates object positions, identifies categories of objects, and priori-tises grasps.

5 Conclusion

In this paper, we first propose to integrate object detec-tion as priori for detecting the object to be grasped and for obtaining the mask of the grasp target, the grasp working zone, and the priority information. Then, effi-cient grasping prediction is performed based on the target grasping priority and grasping area.

We also propose to use the priority grasping method to grasp different categories of objects sequentially and efficiently, making it possible for the robotic arm to grasp intelligently and selectively. It also enables all objects in a multi-object scene to be grasped effecti-vely. The experiments demonstrate the effectiveness of priority grasping and the efficiency of the detection-driven 3D mask method.

Declarations

Funding Information This work is partially funded by the China Scholarship Fund.

Author contribution All authors equally contributed to the content of this article.

Ethics The authors have no conflicts of interest in the development and publication of current research.

References

1. Papacharalampopoulos A, Makris S, Bitzios A, et al (2016) Prediction of cabling shape during robotic manipulation. The International Journal of Advanced Manufacturing Technology 82: 123-132. https://doi.org/10.1007/s00170-015-7318-5
2. Le M T, Lien J J J (2022) Robot arm grasping using learning-based template matching and self-rotation learning network. The International Journal of Advanced. Manufact-turing Technology 121(3-4): 1915-1926. https://doi.org/1 0.1007/s00170-022-09374-y
3. Dang A T, Hsu Q C, Jhou Y S (2022) Development of hum- an–robot cooperation for assembly using image processing techniques. The International Journal of Advanced Manufa- cturing Technology 120(5-6): 3135-3154. https://doi.org/1 0.1007/s00170-022-08968-w
4. Miller A T, Knoop S, Christensen H I, et al (2003) Automatic grasp planning using shape primitives. In: 2003 IEEE International Conference on Robotics and Automation 1824-1829. https://doi.org/10.1109/ROBOT.2003.1241860
5. Jiang Y, Moseson S, Saxena A (2011) Efficient grasping from rgbd images: Learning using a new rectangle representation. In: 2011 IEEE International conference on robotics and automation pp. 3304-3311. https://doi.org/10.1 109/ICRA.2011.5980145

6. Redmon J, Angelova A (2015) Real-time grasp detection using convolutional neural networks. In: 2015 IEEE international conference on robotics and automation (ICRA) pp. 1316-1322. https://doi.org/10.48550/arXiv.14 12.3128
7. Piater J H (2002) Learning visual features to predict hand orientations.
8. Xiang Y, Schmidt T, Narayanan V, et al (2017) Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. arXiv preprint arXiv:1711.00199. https:/ /doi.org/10.48550/arXiv.1711.00199
9. X. Yan et al (2018) Learning 6-DOF Grasping Interaction via Deep Geometry-Aware 3D Representations. In: 2018 IEEE International Conference on Robotics and Automation pp. 3766-3773. https://doi.org/10.48550/arXiv.1708.07303
10. Le T T, Le T S, Chen Y R, et al (2021) 6D pose estimation with combined deep learning and 3D vision techniques for a fast and accurate object grasping. Robotics and Autonomous Systems 141: 103775. https://doi.org/10.1016/j.robot.2021. 103775
11. Wolnitza M, Kaya O, Kulvicius T, et al (2022) 3D object reconstruction and 6D-pose estimation from 2D shape for robotic grasping of objects. arXiv preprint arXiv:2203.01 051. https://doi.org/10.48550/arXiv.2203.01051
12. Gupta H, Thalhammer S, Leitner M, et al (2022) Grasping the Inconspicuous. arXiv preprint arXiv:2211.08182. https:/ /doi.org/10.48550/arXiv.2211.08182
13. Jin M, Li J, Zhang L (2022) DOPE++: 6D pose estimation algorithm for weakly textured objects based on deep neural networks. PloS one 17(6): e0269175. https://doi.org/10.137 1/journal.pone.0269175
14. Huang R, Mu F, Li W, et al (2022) Estimating 6D Object Poses with Temporal Motion Reasoning for Robot Grasping in Cluttered Scenes. IEEE Robotics and Automation Letters. https://doi.org/10.1109/LRA.2022.3147334
15. ten Pas A, Gualtieri M, Saenko K, et al (2017) Grasp pose detection in point clouds. The International Journal of Rob- otics Research 36(13-14): 1455-1473. https://doi.org/10.485 50/arXiv.1706.09911
16. Mahler J, Matl M, Satish V, et al (2019) Learning ambidext- rous robot grasping policies. Science Robotics 4(26): eaau4984. https://doi.org/10.1126/scirobotics.aau4 984
17. Metzner M, Albrecht F, Fiegert M, et al (2022) Virtual train- ing and commissioning of industrial bin picking systems using synthetic sensor data and simulation. International Journal of Computer Integrated Manufacturing 1-10. http://dx.doi.org/10.1080/0951192X.2021.2004618
18. Mallick A, del Pobil A P, Cervera E (2018) Deep learning based object recognition for robot picking task. Proceedings of the 12th international conference on ubiquitous informati- on management and communication pp. 1-9. https://doi.org/ 10.1145/3164541.3164628
19. Zeng A, Song S, Yu K T, et al (2022) Robotic pick-and-place of novel objects in clutter with multi-affordance grasp- ing and cross-domain image matching. The International Journal of Robotics Research 41(7): 690-705. https://doi. org/10.48550/arXiv.1710.01330
20. Zhang Z, Zheng C (2022) Simulation of Robotic Arm Grasping Control Based on Proximal Policy Optimization Algorithm. Journal of Physics: Conference Series. IOP Publishing 2203(1): 012065.

https://ma.x-mol.com/paperRe direct/1499629009831354368

21. Wang T, Chen Y, Qiao M, et al (2018) A fast and robust convolutional neural network-based defect detection model in product quality control. The International Journal of Advanced Manufacturing Technology 94:3465-3471. https: //doi.org/10.1007/s00170-017-0882-0
22. LI Zhengming, Zhang Jinlong (2020) Detection and Posi-tioning of Grab Target Based on Deep Learning. Informa-tion and control 49(2): 147-153.
23. Kato H, Nagata F, Murakami Y, et al (2022) Partial Depth Estimation with Single Image Using YOLO and CNN for Robot Arm Control. In: 2022 IEEE International Conference on Mechatronics and Automation (ICMA) pp. 1727-1731. https://doi.org/10.1109/ICMA54519.2022.9856055
24. Wang C Y, Bochkovskiy A, Liao H Y M (2022) YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.026 96. https://doi.org/10.48550/arXiv.2207.02696
25. Glenn Jocher, et al (2021) ultralytics/yolov5: v5.0 – YOLO v5-P6 1280 models, AWS, Supervise.ly and YouTube integ- rations. https://doi.org/10.5281/ZENODO.4679653
26. Morrison D, Corke P, Leitner J (2018) Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. arXiv preprint arXiv:1804.05172. https://doi.org/ 10.48550/arXiv.1804.05172
27. Fang H S, Wang C, Fang H, et al (2022) AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains. arXiv preprint arXiv:2212.08333. https://doi.org/ 10.48550/arXiv.2212.08333
28. Mahler J, Liang J, Niyaz S, et al (2017) Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. arXiv preprint arXiv:1703.09312. https://doi.org/10.48550/arXiv.1703.09312
29. Mahler J, Matl M, Liu X, et al (2018) Dex-net 3.0: Compu-ting robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. In: 2018 IEEE International Conference on robotics and automation (ICRA) pp. 5620-5627. https://doi.org/10.48550/arXiv.170 9.06670
30. Fang, H.S., Wang, C., Gou, M. and Lu, C. (2019) GraspNet: A Large-Scale Clustered and Densely Annotated Dataset for Object Grasping. arXiv preprint arXiv:1912.13470. https:// doi.org/10.48550/arXiv.1912.13470
31. Fang, H.S., Wang, C., Gou, M. and Lu, C. (2020) Graspnet-1billion: A large-scale benchmark for general object grasp- ing. In Proceedings of the IEEE/CVF conference on comput- er vision and pattern recognition pp. 11444-11453.
32. Bottarel F, Vezzani G, Pattacini U, et al (2020) GRASPA 1.0: GRASPA is a robot arm grasping performance benchm- ark. IEEE Robotics and Automation Letters 5(2): 836-843. https://doi.org/10.48550/arXiv.2002.05017
33. Wang C, Fang H S, Gou M, et al (2021) Graspness discovery in clutters for fast and accurate grasp detection. Proceedings of the IEEE/CVF International Conference on Computer Vision pp. 15964-15973. https://doi.org/10.1109/ICCV4892 2.2021.01566

34. Mehrkish A, Janabi-Sharifi F (2022) Grasp synthesis of continuum robots. Mechanism and Machine Theory 168: 104575. https://doi.org/10.1016/j.mechmachtheory.2021.10 457

35. Tung K, Su J, Cai J, et al (2022) Uncertainty-based Explor-ing Strategy in Densely Cluttered Scenes for Vacuum Cup Grasping. In: 2022 International Conference on Robotics and Automation (ICRA) pp. 3483-3489. https://doi.org/1 0.1109/ICRA46639.2022.9811599

36. Roy D, Panda P, Roy K (2020) Tree-CNN: a hierarchical deep convolutional neural network for incremental learning. Neural Networks 121: 148-160. https://doi.org/10.1016/j.ne unet.2019.09.010

Figures

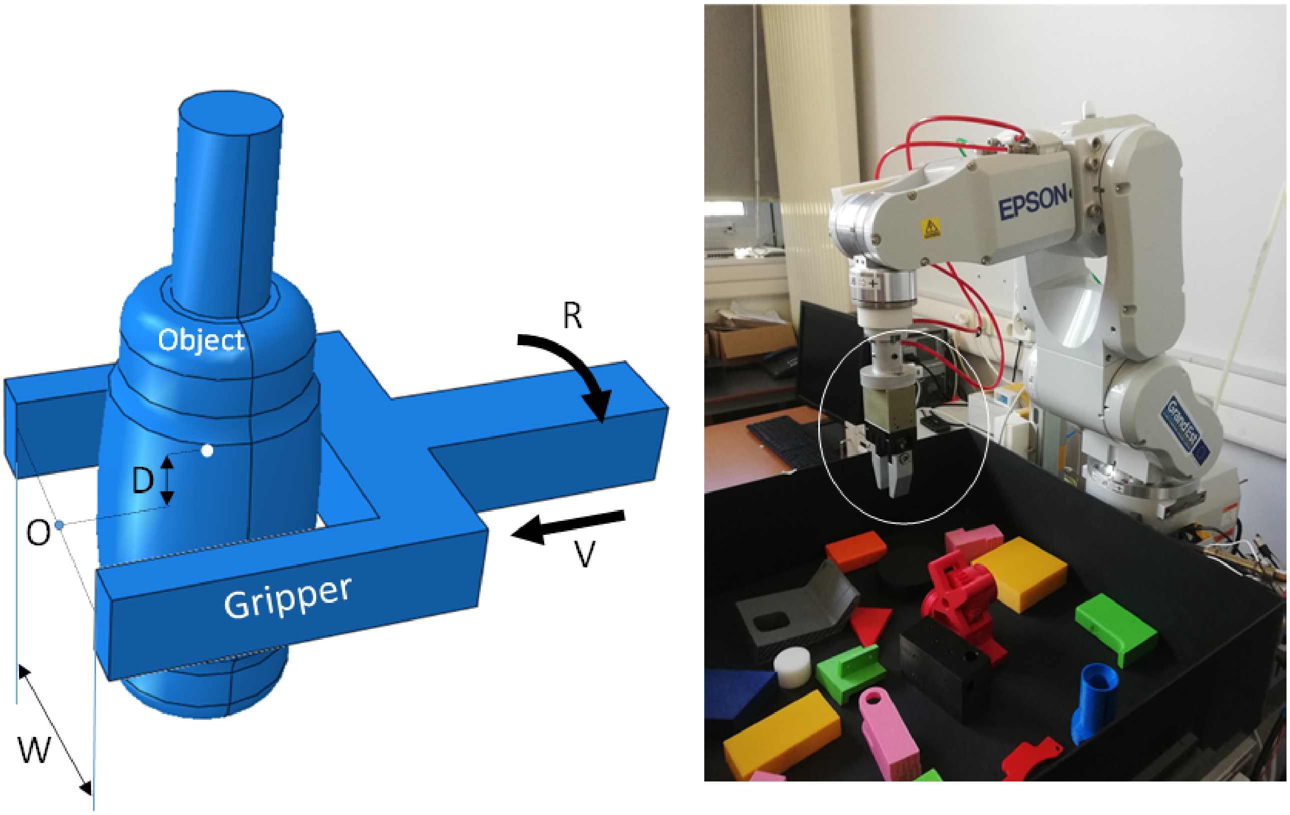

Figure 1

Grasp pose. i.e. predicting the gripping posture of the EPSON robot arm on the right.

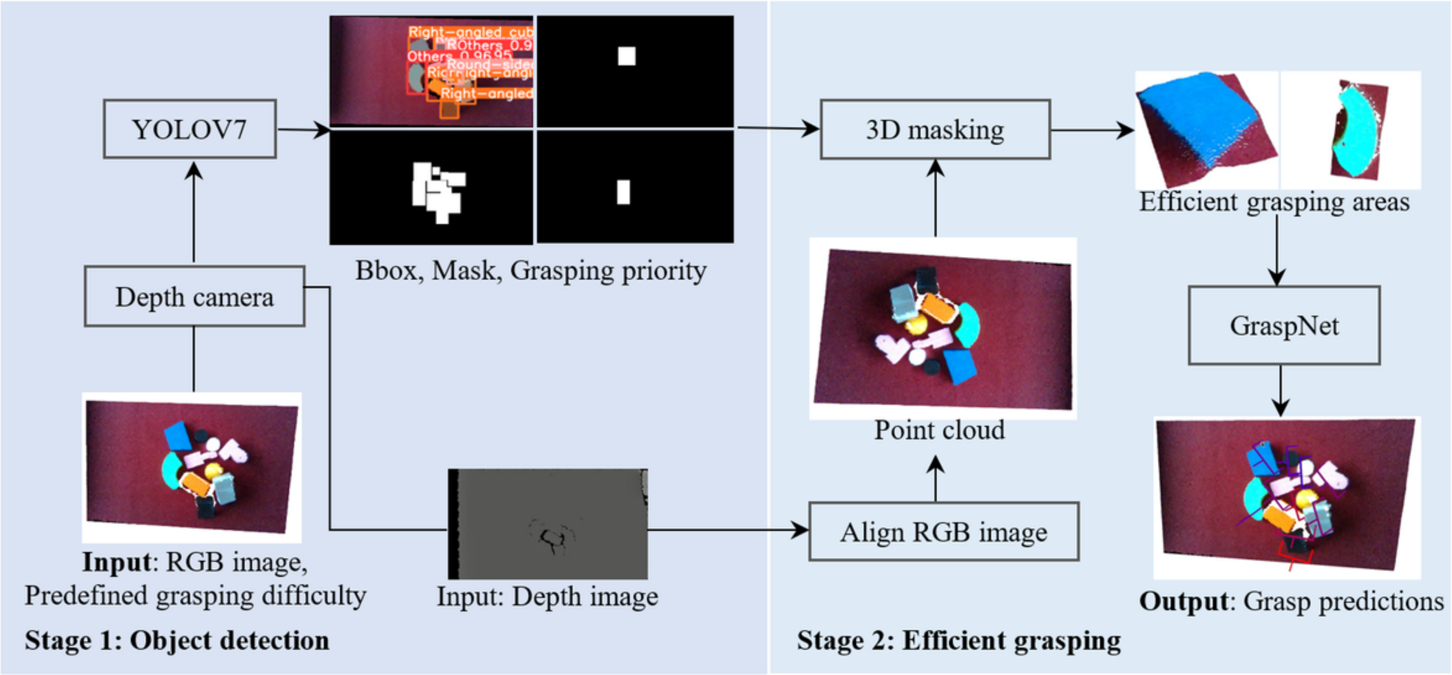
YOLOV7
Bbox, Mask, Grasping priority
Depth camera
Input: RGB image,
Predefined grasping difficulty
Input: Depth image
Stage 1: Object detection
3D masking
Efficient grasping areas
Point cloud
GraspNet
Align RGB image
Output: Grasp predictions
Stage 2: Efficient grasping

Figure 2

Overview of the detection-driven 3D mask method for efficient grasping.

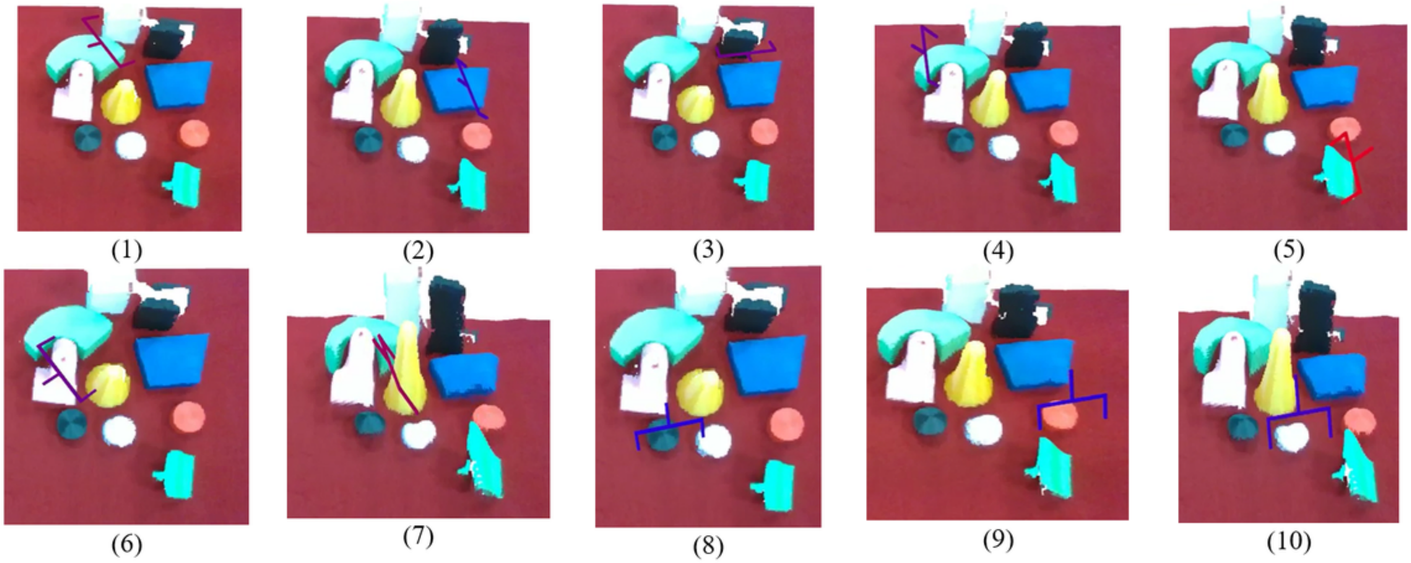

Figure 3

Example of priority grasping.

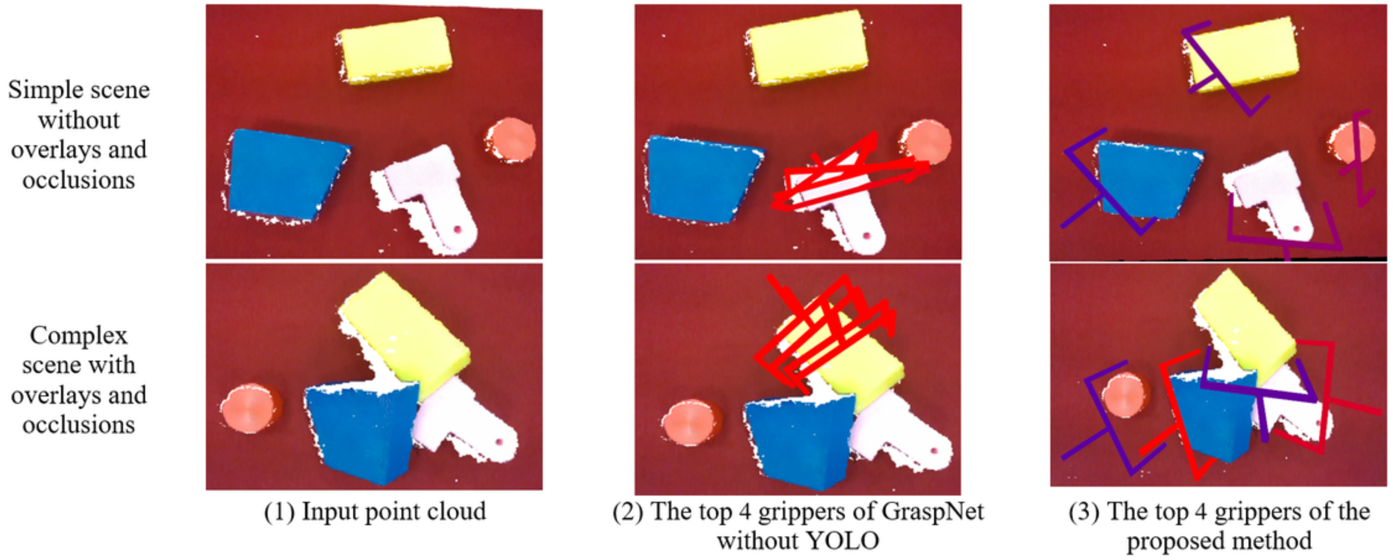
Simple scene without overlays and occlusions
Complex scene with overlays and occlusions
(1) Input point cloud
(2) The top 4 grippers of GraspNet without YOLO
(3) The top 4 grippers of the proposed method

Figure 4

Example of proposed YOLO-based masking and classical GraspNet without YOLO Complex.

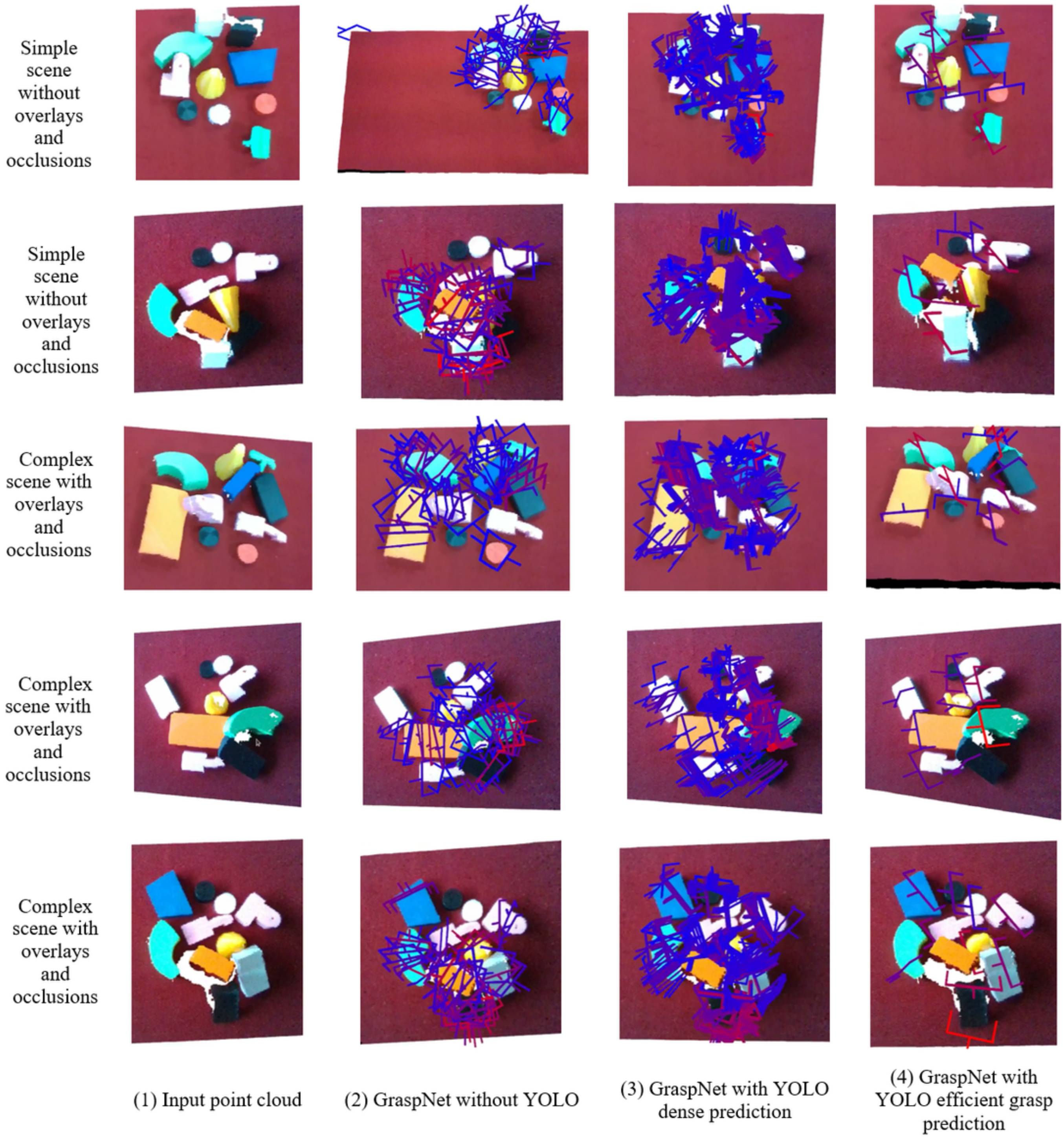
Simple scene without overlays and occlusions
Simple scene without overlays and occlusions
Complex scene with overlays and occlusions
Complex scene with overlays and occlusions
Complex scene with overlays and occlusions
(1) Input point cloud
(2) GraspNet without YOLO
(3) GraspNet with YOLO dense prediction
(4) GraspNet with YOLO efficient grasp prediction

Figure 5

Comparison of proposed YOLO-based masking and GraspNet without YOLO.