JAIST Repository
https://dspace.jaist.ac.jp/

Title	Safety-optimized Strategy for Grasp Detection in High-clutter Scenarios
Author(s)	Li, Chenghao; Zhou, Peiwen; Chong, Nak Young
Citation	2024 21st International Conference on Ubiquitous Robots (UR): 192-197
Issue Date	2024-07-26
Type	Conference Paper
Text version	author
URL	http://hdl.handle.net/10119/19336
Rights	This is the author's version of the work. Copyright (C) 2024 IEEE. 2024 21st International Conference on Ubiquitous Robots (UR), New York, NY, USA, 192-197. DOI: https://doi.org/10.1109/UR61395.2024.10597484. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Description	2024 21st International Conference on Ubiquitous Robots (UR), New York, NY, USA, June 24-27, 2024

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

Safety-optimized Strategy for Grasp Detection in High-clutter Scenarios

Chenghao Li, Peiwen Zhou, and Nak Young Chong

***Abstract*— The detection accuracy and speed of grasp detection models on benchmarks are the focal points of concern in the robotic grasping community. Especially in a collaborative robot setting, the safety of the model is an essential aspect that cannot be overlooked. In this paper, we explore how to enhance the safety of grasp detection models in autonomous vision-guided grasping. Specifically, we propose a simple yet practical Safety-optimized Strategy, which consists of two parts. The first part involves depth prioritization, optimizing the grasp sequence from top to bottom based on the order of depth values, which can mitigate the issue of grasp collisions that may arise when the depth value of the object with the highest grasp quality is significantly higher than that of other objects in high-clutter scenarios. The second part is false-positive protection, where we introduce the robust ArUco marker as the lowest grasp priority. The marker is fixed at certain positions within the camera's field of view, enabling the robot to halt its movement, thereby restraining the robot from grasping objects that should not be grasped. Once the marker disappears, the robot can resume its operations. We validate our method through real grasping experiments with a parallel-jaw gripper and an industrial robotic arm, demonstrating its effectiveness in high-clutter scenarios.**

I. INTRODUCTION

The integration of vision-guided robot grasping technology is crucial for enabling robots to effectively interact with the real world. Traditional visual grasping methods, reliant on manually extracting object features, often face challenges in adapting to dynamically changing unstructured scenarios due to the simplicity of the extracted features. In recent years, the emergence of deep learning, particularly represented by Convolutional Neural Networks (CNNs), has played a pivotal role in advancing computer vision. Unlike artificial features, CNNs leverage a multi-level structure to learn features from extensive data, capturing varying levels of relationships from simple to complex. This characteristic allows for superior feature expression. Consequently, researchers have increasingly applied CNNs to visual grasping, exemplified by methods such as planar grasping representation based grasp detection [1], [2], [3]. A typical planar grasp representation encompasses parameters such as grasp point, angle, and width. Saxena *et al.* [4] successfully employed supervised learning to predict grasp points from images, extending their approach effectively to novel objects. Le *et al.* [5] proposed a representation using a pair of points to depict grasping, while Jiang *et al.* [6] streamlined the 7-dimensional gripper configuration in a real environment to a 5-dimensional rectangle grasping representation (x, y, w, h, θ), which is widely used in later research. Here, (x, y) denotes the grasp center, and w, h, and θ represent the width, height, and angle relative to the horizontal direction, respectively.

Although these grasp detection methods have achieved good detection speed and accuracy on benchmarks such as the Cornell dataset, they overlook some safety issues of grasp detection models in practical grasping applications. In this work, we mainly explore two safety issues. The first safety issue is grasping collisions caused by significant differences in depth values of adjacent objects in high-clutter scenarios. Grasp detection models typically choose to grasp the highest-quality object. When the depth value of the object is much greater than that of adjacent objects, and the model predicts a graspable width that spans adjacent objects, the robot may collide with adjacent objects when grasping the chosen object. The second safety issue is false-positive detection in grasp detection, as grasp detection models can generalize to unknown objects, leading them to detect objects that should not be grasped. For example, in situations where human assistance is required, a person's hand may appear in the robot's grasping field of view. If the grasp detection model mistakenly detects the human hand, it can result in a situation that threatens human safety. To address these two issues, we propose a simple yet practical Safety-optimized Strategy, which enhances the safety of the grasp detection model by using depth prioritization and false-positive protection. Our contributions are as follows:

1) We reveal two safety issues of the grasp detection model: false-positive detection and collisions in high-clutter scenarios.
2) We propose a simple yet practical Safety-optimized Strategy that enhances the safety of the grasp detection model during grasping operations. And can be serve as a baseline for future research.
3) We validate the effectiveness of our proposed method through real grasping experiments in high-clutter scenarios.

II. RELATED WORK

Grasp detection can be classified into two categories based on the diverse modalities of input visual information. The first category involves the use of unimodal data for grasp detection. Johns *et al.* [7] employed simulated depth images to predict grasps, refining the optimal grasp through a grasp uncertainty function. Morrison *et al.* [1] proposed a generative grasp CNN architecture that pixel-wise generates grasps from depth images, effectively addressing challenges related to discrete sampling and computational complexity.

This work was supported by JSPS KAKENHI Grant Number JP23K03756.

Authors are with the School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa 923-1292, Japan {chenghao.li, s2310071, nakyoung}@jaist.ac.jp.

Another recent approach [8] exclusively relied on RGB data, presenting a grasp detection model based on the CSP-ResNet architecture, incorporating multiple residual structures with skip connections. The second category encompasses grasp detection using multimodal data. Wang *et al.* [9] introduced a novel robot grasp detection system, mapping pairs of RGB-D images of novel objects to the optimal grasping pose of a robotic gripper. Jiang *et al.* [6] utilized RGB-D images for grasp inference through a two-step learning process. Lenz *et al.* [10] adopted a two-step approach with a deep learning architecture, encountering challenges in predicting optimal grasps for diverse object types. Yan *et al.* [11] employed a point cloud prediction network for grasp generation, involving initial data preprocessing to obtain color, depth, and masked images. Subsequently, a 3D point cloud of the target object was generated and fed into a pivotal network to predict a grasp. Chu *et al.* [12] proposed a novel architecture capable of simultaneously predicting multiple grasps for multiple objects. Ogas *et al.* [13] discussed a robotic grasping method combining ConvNet for object recognition and a grasping method for objects with known parameters. Kumra *et al.* [14] introduced a Deep CNN architecture using residual layers for predicting robust grasps, highlighting the advantages of a deeper network with residual layers. Asif *et al.* [15] presented EnsembleNet, a consolidated framework generating four grasp representations and synthesizing them to produce grasp scores, ultimately selecting the highest-scoring grasp. Kumra *et al.* [16] also proposed GR-ConvNet, a Generative Residual Convolutional Neural Network model for real-time generation of robust antipodal grasps from n-channel input.

Although these methods have achieved superior detection accuracy and speed through information from different modalities and diverse network structures, they all overlooked how to enhance the algorithmic architecture for safety-critical real-world applications.

III. Proposed Method

In this section, we elaborate on our Safety-optimized Strategy. In the first part, we discussed how to prevent grasping collisions in high-clutter scenes by implementing depth prioritization. The second part focuses on False-positive Protection, where we explain how to reduce false-positive detections in the grasping detection model by introducing ArUco markers and altering the perceptual priorities of the visual system. The final part outlines the overall implementation process of the Strategy.

A. Depth Prioritization

Grasp detection models [14], [15], [16] typically choose to grasp the object with the highest graspable quality based on the acquired images (such as RGB-D, RGB, or Depth images). As shown in Fig. 1. (a), which represents the predicted grasping box by the grasp detection model in a single-object scene. The coordinates (x, y) denote the optimal grasp point for the dinosaur model in the image, and width and θ represent the graspable width and height, typically considered as $w/2$ (the model does not learn this parameter). However, these parameters in image space are not sufficient for real grasping and often require multiple transformations. The parameters of width and θ in image space can directly be mapped to the gripper's width and rotation angle relative to the horizontal direction. Usually, only the position information (x, y) needs to be transformed. Specifically, it is necessary to first convert (x, y) and depth information to the depth camera coordinate system (with the help of camera intrinsic parameters), then transform from the depth coordinate system to the robotic arm's base coordinate system (using the transformation matrix obtained from hand-eye calibration), and finally, through inverse kinematics and forward kinematics, convert the grasp parameters in the robotic arm's base coordinate system into the rotational degrees of the individual joints to achieve grasping.

However, in more complex scenarios, such as high-clutter scenarios where various objects are densely piled up, the depth value corresponding to the optimal graspable object may be significantly greater than the depth values of adjacent objects. Additionally, if the predicted graspable width corresponding to the optimal grasp point is too wide and extends into adjacent objects, it can lead to collisions with the neighboring objects during grasping. As shown in Fig. 1. (b), the dinosaur model is surrounded by many volcano models higher than the dinosaur, and the predicted graspable width extends onto the volcano. This depth difference causes the gripper to collide with the surrounding volcano models during grasping. To mitigate the impact of this issue, we employ a simple yet practical depth prioritization approach. Specifically, we extract the top 10 optimal grasp points predicted by the grasp detection model and rearrange their grasp priority based on the depth values. Here, we set the priority to be based on the ascending order of depth values (*i.e.*, top-down grasping logic). In addition, considering issues such as object reflection and background interference in real scenes, as well as errors caused by the depth camera itself (such as depth holes), the depth information being excessively large or small, leading to robot collisions, we optimize depth prioritization by defining a safe depth range to filter out grasp points with depth values not meeting the requirements, thereby controlling the robot's operation within a certain height range.

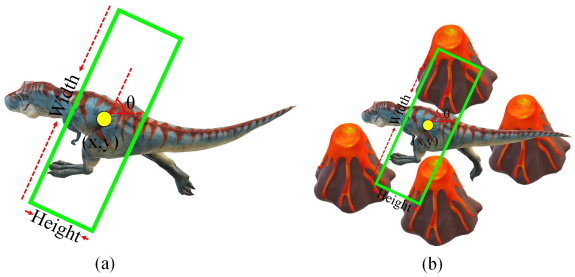


Fig. 1. Grasp detection results in different scenarios. (a) indicates a single object scenario, (b) indicates a high-clutter scenario. In the (b) scenario, the grasp box easily spans over adjacent objects, leading to collisions with other objects when attempting to grasp the target object.

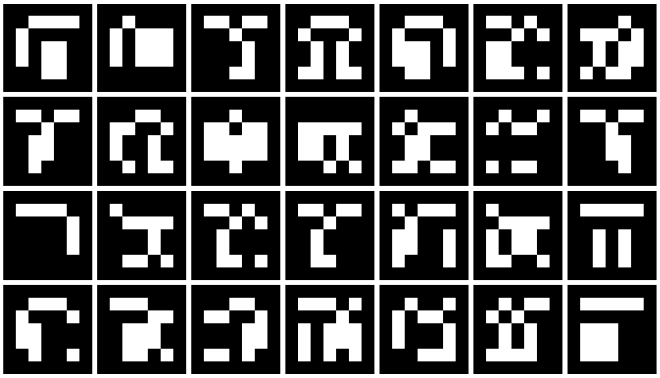
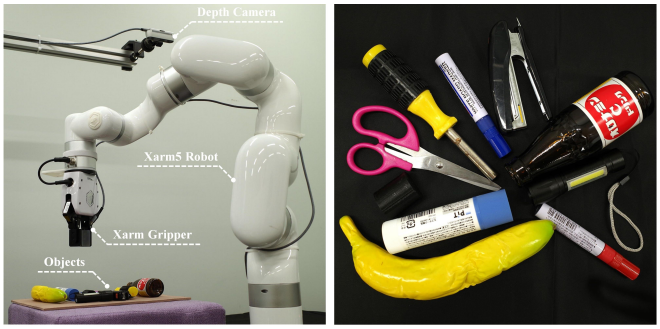


Fig. 2. Some ArUco markers with different appearances.



(a) (b)

Fig. 3. Robot grasping experimental platform (a) and grasping objects (b).

B. False-positive Protection

The grasp detection model can generalize well to unknown objects, however this generality is uncontrollable, which is arbitrary. If the model generalized to an unknown object that should not be grasped (such as a human hand), then this generality becomes a safety concern in practical applications. We conducted human hand detection tests with the GR-ConvNet [16] grasp detection model in different locations and scenarios (detailed in the experiments part). Since human hands in the camera's field of view tend to be at a higher height when manual assistance to the robot in some special situations, they are usually not obstructed by other objects and are fully exposed in the camera's field of view. We followed this characteristic during testing. From Fig. 5, it can be observed that GR-ConvNet [16] has a high probability of detecting human hands in different positions and scenarios (*i.e.*, the optimal grasping point is on the human hand). Assuming that a human hand stays in a certain position for some time, the robot will immediately grasp the recognized human hand after completing the previous grasp, which may causes physical harm to human.

ArUco markers [17] (as shown in Fig. 2) have strong robustness and stability and are widely used in robot hand-eye calibration, UAV landing assistance, path planning, etc. In this paper, we also introduce ArUco markers into real grasping to reduce the false-positive detection of the grasp detection model. Specifically, we propose False-positive Protection, embedding the recognition algorithm of ArUco markers into the entire grasping visual system and restricting the recognition of ArUco markers by the grasping visual system as the highest priority (higher than the grasp detection model) and the grasping of ArUco markers as the lowest priority. In other words, during the grasping process, if an ArUco marker or an object with an ArUco marker appears, the robot will stop moving, thus protecting human hands during manual assistance. When the ArUco marker disappears, the robot can resume its operations, that is, the running of the code will not be interrupted.

Algorithm 1 Safety-optimized Strategy

Input: original points $\{P_1, P_2, P_3, ..., P_n\}$
Output: optimized points $\{P_t, P_{t+1}, P_{t+2}, ..., P_k\}$
for $i = 1, 2, 3, \ldots, n$ **do**
  if $ArUco\ is\ True$ **then**
    return None
  else if $D_i \notin [D_{min}, D_{max}]$ **then**
    filter P_i, $k \leftarrow n - 1$
  else if $D_i \in [D_{min}, D_{max}]$ **then**
    save P_i
  end if
end for
for $j = 1, 2, 3, \ldots, k$ **do**
  $P_c \leftarrow P_j$, $D_c \leftarrow D_j$, $t \leftarrow j - 1$
  while $t \geq 0$ and $D_t > D_c$ **do**
    $D_{t+1} \leftarrow D_t$, $P_{t+1} \leftarrow P_t$, $t \leftarrow t - 1$
  end while
  $D_{t+1} \leftarrow D_c$, $P_{t+1} \leftarrow D_c$
end for
return $\{P_t, P_{t+1}, P_{t+2}, ..., P_k\}$

C. The overall process of the Safety-optimized Strategy

The overall process of the Safety-optimized Strategy is shown in Algorithm 1. $\{P_1, P_2, P_3, ..., P_n\}$ represents the detected top-n graspable points from the grasp detection model in one frame. Firstly, determine whether the ArUco marker is recognized. If so, return a null value, meaning the robot will stop (the code will not stop), until the ArUco marker disappears, and then the robot will resume task execution. If the ArUco marker is not recognized, determine whether the graspable point is within the safe depth range $[D_{min}, D_{max}]$. If so, save this point, otherwise, filter this point. Finally, the filtered graspable points are sorted according to the descending order of the depth value, that is, $\{P_t, P_{t+1}, P_{t+2}, ..., P_k\}$. Due to our focus on our optimization methods, we do not elaborate extensively on the existing grasp detection models and the recognition algorithms for ArUco markers. Detailed information and open-source code can be referred to in [1], [16], [17].

IV. Experiments

In this section, we validated the effectiveness of our proposed method through experiments. Firstly, we introduced

the experimental setup including the experimental equipment and scenarios. Then, we validated the effectiveness of Depth Prioritization in the Safety-optimized Strategy under high-clutter scenarios. Finally, we validated the effectiveness of False-positive Protection in the Safety-optimized Strategy under high-clutter scenarios.

A. Experiments Settings

We used a combination of the Cornell Grasp Dataset [10] and the OCID Grasp Dataset [18] as our dataset and opted for GR-ConvNet (RGB-D version) [16] as the grasp detection model. The training took place on a single NVIDIA RTX 4070Ti GPU with 12 GB of memory. The computer system was running Ubuntu 22.04, and we utilized PyTorch 2.1.2 with CUDA 12.1 as the deep learning framework. Following the training parameters outlined in GR-ConvNet [16], we randomly shuffled the entire dataset, allocating 90% for training and 10% for testing before model training. During the training process, the data was uniformly cropped to 224×224, and data augmentation techniques such as random zoom and random rotation were applied. Since our primary focus was on real grasping scenarios, we directly used the trained GR-ConvNet with the highest detection accuracy to evaluate the effectiveness of our Safety-optimized Strategy in grasping.

Our overall grasping system is illustrated in Fig. 3. (a), primarily consisting of one Intel RealSense D415, one Xarm5 industrial robot, and one Xarm parallel-jaw gripper. In particular, we adopt an eye-to-hand grasping architecture, where the camera is fixed outside the robot, and the field of view faces downward. Fig. 3. (b) illustrates the objects utilized in our grasping experiments, comprising a total of 10 different types. In these experiments, we combine these 10 objects to create 10 distinct high-clutter scenarios.

B. Effectiveness of Depth Prioritization

To demonstrate the effectiveness of Depth Prioritization, we initially tested the Grasping Collision Rate (GC-R) and Grasping Success Accuracy (G-Acc) of the GR-ConvNet across various scenarios. Subsequently, we evaluated the Grasping Collision Rate after integrating Depth Prioritization (DPGC-R) and the Grasping Success Accuracy with Depth Prioritization (DPG-Acc). The Grasping Success Rate here indicates how many of the 10 objects were successfully grasped in each scene, while the Grasping Collision Rate signifies whether objects not successfully grasped experienced collisions.

The experimental results are shown in Table I. The Grasping Collision Rate (GC-R) and Grasping Success Accuracy (G-Acc) of the GR-ConvNet reached 36% and 58%, respectively. After adding Depth Prioritization, the Grasping Collision Rate (GC-R) dropped to 24%, and the Grasping Success Accuracy (G-Acc) rose to 78%, which validated the effectiveness of Depth Prioritization. Furthermore, in Fig. 4, we demonstrate the occurrence of grasp collisions (GC) in high-clutter scenarios using the GR-ConvNet (first row) and the grasping performance of our method (second row) in high-clutter scenarios (DPG). As depicted in the figures, our method successfully alleviates the issue of grasp collisions that occur when the depth value of the object with the highest grasp quality is substantially higher than that of other objects. In other words, it enables the system to prioritize grasping nearby objects with lower-depth values first.

C. Effectiveness of False-positive Protection

In validating the effectiveness of False-positive Protection, we primarily examined the False-positive Detection Rate (FPD-R) of the GR-ConvNet and False-positive Protection Rate (FPP-R) of our method in various scenarios. The False-positive Detection Rate indicates the rate of detecting a human hand, whereas the False-positive Protection Rate signifies the rate at which the system fails to detect a human hand. During the experiments, the human hand moved to different positions and maintained various poses in each scenario, with 10 tests conducted for each scenario.

The experimental results are presented in Table II. The False-positive Detection Rate (FPD-R) of the GR-ConvNet and the False-positive Protection Rate (FPP-R) of our method reached 84% and 83%, respectively. This indicates that although the GR-ConvNet easily detects human hands, the integration of our method helps alleviate this issue to some extent. We also visualize our experimental results in this section, as shown in Fig. 5. The first row represents the False-positive detection (FPD) results of the GR-ConvNet, and the second row represents the False-positive protection (FPP) results of our method. It can be observed from the figure that when ArUco markers appear in the scene, the detection of human hands by the grasp detection model can be suppressed. Video is presented at: https://www.youtube.com/channel/UC-nJesbpK2jbigNBmZTIwjg.

V. Conclusion

This paper proposes a Safety-optimized Strategy from the perspective of safety in grasping. It is divided into two parts. The first part is depth prioritization, which optimizes the grasp sequence from top to bottom based on the order of depth values. This approach helps mitigate the issue of grasp collisions that may occur when the depth value of the object with the highest grasp quality is significantly higher than that of other objects in high-clutter scenarios. The second part is false-positive protection, where robust ArUco markers are introduced as the lowest grasp priority and can help limit human-safety threatening grasping caused by false-positive detection. Finally, we validate the effectiveness of depth prioritization and false-positive protection through experiments conducted in real high-clutter grasping scenarios.

Limitation and Future Works: Since this work is still in its early stages, many parts need improvement. For instance, depth prioritization mainly focuses on the post-processing stage of the grasp detection model, and internal structural defects of the grasp detection model also need to be ad-

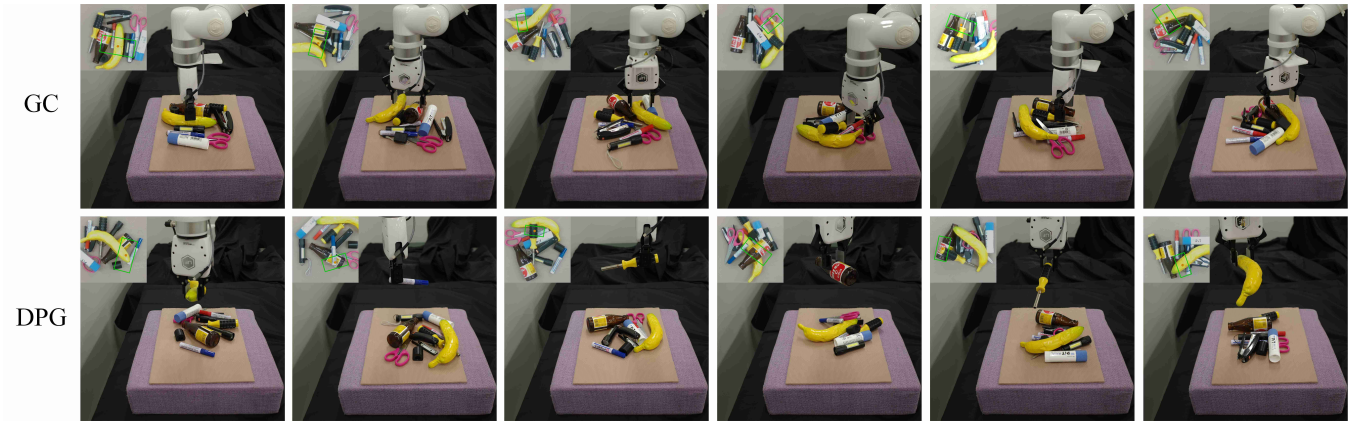


Fig. 4. Visualization of real grasping for GC (first row) and DPG (second row) in different scenarios. The anti-collision performance of grasping through DPG has been significantly improved compared to the original GC grasping method, and effectively reduce the damage during the robot grasping process.

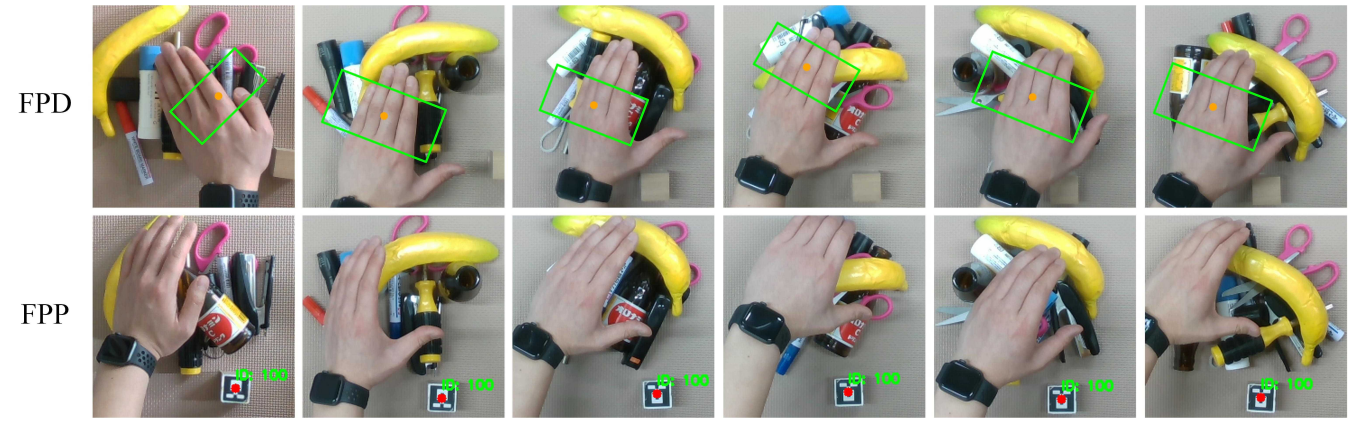


Fig. 5. Visualization of FPD (first row) and FPP (second row) in different scenarios. It can be observed that regardless of the changes in the scenario, when both human hands and ArUco markers appear simultaneously, the ArUco markers are capable of effectively inhibiting the operation of the model.

TABLE I

RESULTS OF DEPTH PRIORITIZATION

Scenarios	1	2	3	4	5	6	7	8	9	10	Overall (%)
GC-R	5/10	3/10	5/10	4/10	3/10	4/10	4/10	3/10	3/10	2/10	36.0
G-Acc	4/10	6/10	5/10	5/10	7/10	5/10	6/10	6/10	7/10	7/10	58.0
DPGC-R	2/10	4/10	3/10	2/10	1/10	2/10	3/10	3/10	1/10	3/10	24.0
DPG-Acc	10/10	8/10	6/10	8/10	9/10	8/10	6/10	7/10	9/10	7/10	78.0

TABLE II

RESULTS OF FALSE-POSITIVE PROTECTION

Scenarios	1	2	3	4	5	6	7	8	9	10	Overall (%)
FPD-R	8/10	9/10	7/10	8/10	8/10	10/10	10/10	8/10	9/10	7/10	84.0
FPP-Acc	9/10	8/10	7/10	9/10	9/10	8/10	8/10	10/10	7/10	8/10	83.0

dressed. Therefore, in our future research, we will pay more attention to the design of the safe grasp detection model.

REFERENCES

[1] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 183–201, 2020.

[2] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping Scenarios," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4768-4775.

[3] D. Park, Y. Seo, D. Shin, J. Choi, and S.Y. Chun, "A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7300-7306.

[4] A. Saxena, J. Driemeyer, and A.Y. Ng, "Robotic grasping of novel objects using vision," *The International Journal of Robotics Research*, vol. 27, no. 2, pp. 157–173, 2008.

[5] Q.V. Le, D. Kamm, A.F. Kara, and A.Y. Ng, "Learning to grasp objects

with multiple contact points," in *2010 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2010, pp. 5062-5069.

[6] Y. Jiang, S. Moseson, and A. Saxena, (2011, May). "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *2011 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 3304-3311.

[7] E. Johns, S. Leutenegger, and A.J. Davison, "Deep learning a grasp function for grasping under gripper pose uncertainty," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 4461-4468.

[8] M. Shan, J. Zhang, H. Zhu, C. Li, and F. Tian, "Grasp Detection Algorithm Based on CSP-ResNet,". in *2022 IEEE International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*. IEEE, 2022, pp. 501-506.

[9] Z. Wang, Z. Li, b. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Advances in Mechanical Engineering*, vol. 8, no. 9, 8(9), p. 1687814016668077, 2016.

[10] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[11] X. Yan, M. Khansari, J. Hsu, Y. Gong, Y. Bai, S. Pirk, and H. Lee, "Data efficient learning for sim-to-real robotic grasping using deep point cloud prediction networks," *arXiv preprint arXiv:1906.08989*, 2019.

[12] F.J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.

[13] E. Ogas, L. Avila, G. Larregay, and D. Moran, "A robotic grasping method using convnets," in *2019 Argentine Conference on Electronics (CAE)*. IEEE, 2019, pp. 21–26.

[14] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 769–776.

[15] U. Asif, J. Tang, and S. Harrer, "Ensemblenet: Improving grasp detection using an ensemble of convolutional neural networks," in *2018 British Machine Vision Conference (BMVC)*, 2018.

[16] S. Kumra, s. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9626–9633.

[17] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez, "Automatic generation and detection of highly reliable fiducial markers under occlusion," *Pattern Recognition*, vol. 47, no. 6, pp. 2280-2292, 2014.

[18] 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)*. IEEE, 2021, pp. 13452-13458.