UNIVERSITY OF LEEDS

This is a repository copy of *A Task-Oriented Grasping Framework Guided by Visual Semantics for Mobile Manipulators*.

White Rose Research Online URL for this paper:
https://eprints.whiterose.ac.uk/212204/

Version: Accepted Version

Article:
Zhang, G. orcid.org/0009-0009-7811-065X, Wang, S. orcid.org/0000-0001-5620-9151, Xie, Y. orcid.org/0000-0003-1158-1587 et al. (3 more authors) (2024) A Task-Oriented Grasping Framework Guided by Visual Semantics for Mobile Manipulators. IEEE Transactions on Instrumentation and Measurement, 73. 7504213. ISSN 0018-9456

https://doi.org/10.1109/tim.2024.3381662

© 2024 IEEE. 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.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

A Task-Oriented Grasping Framework Guided by Visual Semantics for Mobile Manipulators

Guangzheng Zhang, Shuting Wang, Yuanlong Xie, *Senior Member, IEEE*,
Sheng Quan Xie, *Fellow, IEEE*, Yiming Hu, and Tifan Xiong

***Abstract*—The densely cluttered operational environment and the absence of object information hinder mobile manipulators from achieving specific grasping tasks. To address this issue, this paper proposes a task-oriented grasping framework guided by visual semantics for mobile manipulators. With multiple attention mechanisms, we first present a modified DeepLabV3+ model by replacing the backbone networks with Mobilenetv2 and incorporating a novel attention feature fusion module to build a preprocessing module, thus producing semantic images efficiently and accurately. A semantic-guided viewpoint adjustment strategy is designed in which the semantic images are used to calculate the optimal viewpoint that enables the eye-in-hand installed camera to self-adjust until it encompasses all the objects within the task-related area. Based on the improved DeepLabV3+ model and the generative residual convolutional neural network, a task-oriented grasp detection structure is developed to generate a more precise grasp representation for the specific object in densely cluttered scenarios. The effectiveness of the proposed framework is validated through the dataset comparison tests and multiple sets of practical grasping experiments. The results demonstrate that our proposed method achieves competitive results versus the state-of-art methods, which attains an accuracy of 98.3% on the Cornell grasping dataset and achieves a grasping success rate of 91% in densely cluttered scenes.**

***Index Terms*—Task-oriented robotic grasping, visual semantics, absence of object information, deep learning, mobile manipulator.**

I. Introduction

MOBILE manipulators are extensively employed in manufacturing and service fields owing to their flexibility and controllability [1], [2], [3]. Object Grasping is a prevalent and fundamental task for mobile manipulators and is considered a challenging technology in the field of mobile manipulation [4], [5], [6]. In practical applications, mobile manipulators generally have specific task attributes, such as the grasping of specific objects, requiring them to determine which object to grasp and how to grasp. This kind of grasping is usually called task-oriented grasping, which necessitates the mobile manipulators to possess both object recognition and object grasping capabilities simultaneously [7], [8], [9]. Moreover, the operating environment for mobile manipulators is typically complex and changeable. Challenges arise due to factors such as densely cluttered objects and the absence of object information, leading to failures in object grasping [10], [11]. Up to now, achieving task-oriented grasping for mobile manipulators under such conditions remains a challenge.

To achieve task-oriented grasping for mobile manipulators, it is crucial to investigate the grasp detection methods. With the advancements in machine vision and deep learning, robotic grasp detection technologies have shifted from basic physical geometric grasping to deep neural network-based grasping pose prediction [12], [13]. Recent studies in robotic grasp detection have focused on improving the structure of neural networks to attain superior performance. For example, Morrison *et al*. [14] proposed a generative grasping convolutional neural network (GGCNN) that can generate grasp representations based on the depth image of the object. Kumra *et al*. [15] used a similar idea and introduced a residual structure to propose a generative residual convolutional neural network (GRCNN), which improves the detection accuracy while ensuring detection speed. Although these methods provide some new ideas for robotic grasping, there remain at least three main challenges in task-oriented grasping within densely cluttered scenarios.

1) *Accurate and efficient perception for objects in densely cluttered scenarios*: Accurate object recognition in the scene is the initial step for task-oriented grasping. Traditional grasp detection methods [14], [15], [16], [17] lack the ability to recognize specific objects, rendering them insufficient for task-oriented grasping. Recently, object detection methods have been extensively utilized for object recognition in grasping tasks [5], [18]. However, these methods often lead to detection box stacking in densely cluttered scenarios, which greatly affects the subsequent grasp detection and causes a decline in its performance. Semantic segmentation methods have recently been employed for object segmentation in grasping tasks due to their pixel-level classification capabilities [19]. Although these methods achieve high accuracy and partially address the limitations of object detection methods in densely cluttered scenarios, the presence of a large scale of network parameters poses efficiency issues, thereby limiting their real-time applicability. Thus, how to balance the accuracy and efficiency of semantic segmentation methods to achieve high-

The work was supported by National Natural Science Foundation of China under Grant 52275488, Key Research and Development Program of Hubei Province, China under Grant 2022BAA064 and Key Research and Development Program of Henan Province, China under Grant 221111220800. *(Corresponding author: Yuanlong Xie; Tifan Xiong.)*

Guangzheng Zhang, Shuting Wang, Yuanlong Xie, Yiming Hu and Tifan Xiong are with the school of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China (email: hust.gzz@gmail.com; wangst@hust.edu.cn; yuanlongxie@hust.edu.cn; D202280317@hust.edu.cn; xiongtf@hust.edu.cn).

Sheng Quan Xie is with the School of Electronic and Electrical Engineering, University of Leeds, Leeds LS2 9JT, U.K. (e-mail: s.q.xie@leeds.ac.uk).

Related materials are available at https://github.com/HustZgz/Materials.

performance perception still needs to be investigated.

2) *Effective solution for the absence of object information in visual field:* Object occlusion and camera viewpoint limitations are common causes of the absence of object information during the robotic grasping process [8], [20]. Extensive research has been conducted in recent years to address the absence issue of object information caused by occlusion[21], [22]. These studies focused on model training, viewpoint variations, and active search. However, the dependence on prior models limits the application of the model training method. The viewpoint variations method utilizes the motion of the mobile platform to change the camera viewpoint to acquire complete information about objects. Nevertheless, the motion of the mobile platform can also be affected by unpredictable factors, which introduces additional complexity. The active search method employs the motion of the robotic manipulators to change the scenarios to acquire complete information about objects. But such search method often lacks information guidance, resulting in time-consuming [23]. Moreover, for the absence of object information caused by camera viewpoint limitations, the active perception methods are predominantly employed by most studies to solve this issue [24], [25]. These methods are based on scene information and utilize the motion of mobile robots to change the camera viewpoint, typically employed for the perception and reconstruction of large-scale scenes. For grasping tasks of mobile manipulators, environmental disturbances cause some objects within the task-related area to be positioned outside the camera's field of view, resulting in the absence of object information. However, previous studies [26], [27] did not consider this factor, which hinders the mobile manipulators from accurately recognizing the target object during grasping tasks. This greatly reduces the grasping success rate of mobile manipulators in complex environments, thus limiting their application in such scenarios.

3) *Accurate grasp detection for the specific object in densely cluttered scenarios:* Grasp detection is crucial for task-oriented grasping. Recent studies [19], [28] have investigated the integration of semantic segmentation and grasp detection, proposing multistage networks that enable grasp detection in cluttered scenes. However, these methods are task-agnostic, which cannot achieve task-oriented grasp detection. To address this problem, previous studies [29], [30] utilized object detection and semantic segmentation methods to generate detection boxes and masks. By separating the object from the background, these methods achieve task-oriented grasp detection. Although these studies have provided some ideas and methods for task-oriented grasp detection, they only have primarily focused on the grasp detection of a single object or the object in simple and discrete scenes. The performance limitations of the perception and grasp detection modules make it difficult to apply these methods to task-oriented grasping in densely cluttered scenes.

In this paper, a task-oriented grasping framework guided by visual semantics is proposed for mobile manipulators. Compared with existing methods, this paper makes the following main contributions.

1) We propose a task-oriented grasping framework guided by visual semantics for mobile manipulators, which can achieve task-oriented grasping in densely cluttered scenarios with the absence of object information. Different from the existing methods [26], [27] that simply integrate object and grasp detection, the proposed framework leverages visual semantic feedback to guide the adjustment of camera viewpoint and grasp detection to handle more complex grasping scenarios. Experimental results show that this framework enhances the task-oriented grasping success rate of mobile manipulators in complex scenarios.
2) The DeepLabV3+ model is improved by introducing a lightweight network MobileNetv2 and multiple attention mechanisms. Meanwhile, an attention feature fusion module (AFFM) is also proposed to replace the original feature fusion method. Compared to the baseline model, the improved DeepLabV3+ model demonstrates enhanced efficiency and accuracy, achieving real-time semantic segmentation of densely cluttered scenarios.
3) A semantic-guided viewpoint adjustment strategy is developed to address the absence issue of object information in the visual field. Unlike previous strategies that lack information guidance, the proposed strategy improves efficiency by utilizing the average pixel values of object classes in the semantic image to guide the adjustment of the camera viewpoint.
4) A grasp detection structure based on the improved DeepLabV3+ model and GRCNN is proposed, which can achieve task-oriented grasp detection in densely cluttered scenarios. Compared to the previous work [26], [27], [29], [30] that solely focused on task-oriented grasp detection in discrete scenes, the proposed task-oriented grasp detection structure exhibits superior performance in densely cluttered scenarios by incorporating the improved DeepLabV3+ model.

The remainder of this paper is organized as follows. Section II provides an overview of the preliminaries and presents the problem statement. Section III details the proposed task-oriented grasping framework guided by visual semantics. Section IV provides experimental validation and performance evaluation of the proposed framework. Finally, Section V concludes the paper with a summary of the findings.

II. Preliminaries and Problem Statement

A. DeepLabV3+

The accurate perception of objects in the scene is the prerequisite for achieving task-oriented grasping. DeepLabV3+ is a highly accurate semantic segmentation model commonly applied in the optical and remote sensing domains [31]. Deeplabv3+ is composed of two main components: an encoder and a decoder. The encoder extracts features from input images using the Xception backbone network and the Atrous Spatial Pyramid Pooling (ASPP). The Xception network generates

two outputs: a low-level feature and a high-level feature. The former is fused in the decoder, while the latter is utilized in the ASPP for extracting multi-scale features. The ASPP incorporates various components, including a 1 × 1 convolution layer, three dilated convolution layers with expansion rates of 6, 12, and 18 respectively, and a global average pooling (GAP) operation. The outputs of these parts are concatenated to obtain features that encompass multi-scale information [32]. After a 1 × 1 convolution layer, the features are passed to the decoder for further feature fusion.

The decoder integrates the low-level feature with the high-level feature. The low-level feature contributes semantic information and undergoes dimension adjustment through a 1 × 1 convolution layer before fusion. The high-level feature, which provides detailed information, is upsampled using 4 × bilinear interpolation to adjust the feature map size before fusion with the low-level feature. The fused features are further processed through a 3 × 3 convolution layer and upsampled using 4 × bilinear interpolation to generate segmentation results with the same size as the original image [33].

B. GRCNN

After the target object is segmented, a high-performance grasp detection network is required to generate a grasp representation for it. The generative residual convolutional neural network (GRCNN) [15] is a lightweight and accurate grasp detection network. GRCNN generates pixel-wise grasp representations of objects, consisting of three convolution layers, three transposed convolution layers, and five residual layers. In GRCNN, the features of the input image are initially extracted by the three convolution layers, reducing the image size from 224 × 224 to 56 × 56. Subsequently, the features traverse through the five residual layers to prevent gradient vanishing and dimensionality errors. Since the residual layer does not alter the feature size, the convolution transpose operation is employed to up-sample the image, which ensures that the dimensions of the output image are in line with those of the input image. Finally, the highest-quality grasp representation is obtained, which includes the center of grasp, grasp angle, required end width for grasping, and grasp quality score.

C. Grasp Representation

The objective of the grasp detection task in robotic manipulators is to infer the optimal grasp pose for objects. The 5-D grasp representation proposed in [34] is commonly adopted to describe a grasp pose, which can be expressed as

$$G = \{x, y, w, h, \theta\} \tag{1}$$

where (x, y), w, h, and θ denote the center coordinate, width, height, and rotation angle of the grasp rectangle, respectively.

However, for multiple grasp rectangles, this representation method cannot determine the quality of each one. Therefore, an improved 5-D grasp representation proposed in [14] is chosen in this paper, which represents a grasp as follows

$$G_i = \{x, y, \theta_i, w, Q\} \tag{2}$$

where θ_i denotes the rotation angle of grasp rectangle, $\theta_i \in [-\pi/2, \pi/2]$ [35], w denotes the grasp width, $Q \in [0, 1]$ denotes the grasp quality score, which reflects the success rate of the grasp. This grasp representation predicts the quality score of each grasp rectangle, the best grasping candidate can be expressed as $G^* = \max_Q G_i$.

Using the generated grasp representation, the grasp pose is obtained by transforming the 5-D grasp pose from the image coordinate system to the coordinate system of the robotic manipulator end. This transformation allows the robotic manipulator to accurately execute the grasping task.

Motivation: Efficient and accurate perception of objects in the scene is a crucial prerequisite for achieving successful task-oriented grasping with mobile manipulators. While the DeepLabV3+ model offers high segmentation accuracy, its efficiency is not sufficient for real-time semantic segmentation [36]. Therefore, it is necessary to balance the efficiency and accuracy of DeepLabV3+. Furthermore, an efficient and accurate grasp detection network is also required. Although GRCNN demonstrates faster detection speed and higher accuracy, it is susceptible to environmental interference and lacks the ability to generate grasp representations for specific objects [37]. Hence, a grasp detection structure based on DeepLabV3+ and GRCNN is considered, which can address the performance instability of GRCNN caused by environmental interference and generate grasp representations for specific objects. Meanwhile, it is also crucial to consider scene factors in grasping tasks. The absence of object information caused by camera viewpoint limitations often occurs during the grasping process of mobile manipulators, which can lead to grasp failure. Therefore, designing an adjustment strategy to tackle this situation is essential. The main motivation of our work includes three aspects:

1) Improve the DeepLabV3+ model to balance its' efficiency and accuracy.
2) Propose an information-guided viewpoint adjustment strategy to address the problem of the absence of object information.
3) Develop a task-oriented grasping detection structure based on the DeepLabV3+ and CRCNN to generate accurate grasp representation for specific objects in densely cluttered scenes.

III. Proposed Task-Oriented Grasping Framework Guided by Visual Semantics

To address the issues of grasping failure caused by the densely cluttered grasp scenarios and the absence of object information, this paper proposes a novel task-oriented grasping framework guided by visual semantics. Unlike most existing frameworks, the proposed framework aims to improve grasping performance by incorporating visual semantics. It comprises three components: real-time semantic segmentation of scenes, semantic-guided adjustment of camera viewpoint, and task-oriented grasp detection, as shown in Fig. 1. In the framework, the visual semantic information is utilized to guide both the adjustment of camera viewpoint and robotic

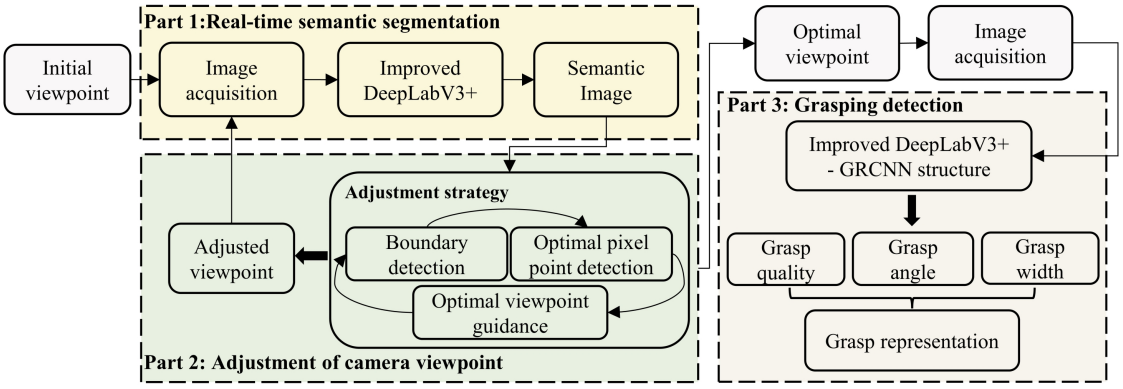


Fig. 1. Task-oriented grasping framework guided by visual semantics.

grasp detection. To elaborate, the first component involves collecting real-time scene images and generating corresponding semantic images through semantic segmentation. The second component focuses on adjusting the camera's viewpoint to the optimal viewpoint based on the obtained semantic image. Lastly, the third component utilizes the images captured from the optimal viewpoint to generate a grasp representation of the specified object in a densely cluttered scenario. The following subsections provide detailed descriptions of each component.

A. Real-time Semantic Segmentation by Improved DeepLabV3+ Model

To meet the requirement of real-time semantic segmentation of objects in the scene, we propose an improved DeepLabV3+ model. Firstly, the lightweight network MobileNetv2 is utilized as a substitute for the original backbone of the DeepLabV3+ to minimize accuracy loss while significantly improving the detection efficiency. Then, the convolutional block attention module (CBAM) is introduced to reduce the loss of image detail information during feature propagation. Furthermore, an attention feature fusion module (AFFM) is proposed to replace the original concatenation method, thereby obtaining superior feature representations. The structure of the improved DeepLabV3+ model is illustrated in Fig. 2. This model first extracts the features from the input image using MobileNetv2, resulting in two distinct features: low-level features and high-level features. The high-level features then pass through the CBAM and ASSP modules to generate multi-scale fusion features. These features are subjected to 1×1 convolution and 4× bilinear interpolation upsampling before being input into the AFFM for feature fusion. Simultaneously, the low-level features pass through CBAM and a 1×1 convolution, which are also input into the AFFM for feature fusion. The output of the AFFM is a fused feature that reflects both the semantic and detailed information of the image. This fused feature then undergoes a 3x3 convolution and 4× bilinear interpolation upsampling to generate a semantic image of the same size as the input image. The modifications to the DeepLabV3+ model are described in the subsequent steps.

In the encoder of the DeepLabV3+ model, the low-level feature generated by the backbone network is directly transmitted

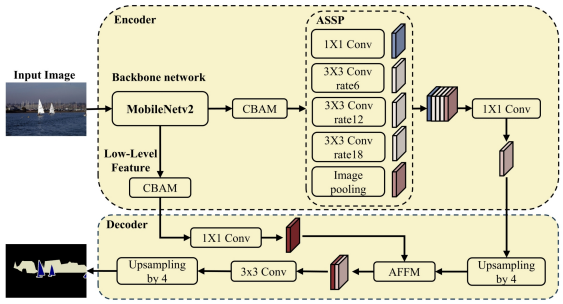


Fig. 2. Structure of improved DeepLabV3+ model.

to the decoder for feature fusion. Simultaneously, the high-level feature generated by the backbone network is directly fed to the ASPP for further processing. To improve the feature extraction efficiency of the model, we replace the original backbone network with MobileNetv2. It has fewer parameters, which can reduce the computational complexity of the model and improve inference speed without sacrificing performance. Moreover, CBAM is integrated into the propagation process of these two features to minimize the loss of intricate features and enhance the segmentation accuracy. CBAM comprises a channel attention module (CAM) and a spatial attention module (SAM) [38], as illustrated in Fig. 3. In CBAM, the initial feature $\mathbf{F}$ is first input into the CAM. Within the CAM, the global maximum pooling and global average pooling operations are applied to feature $\mathbf{F}$, resulting in two outputs. These outputs are then fed into a shared multi-layer perceptron, which produces intermediate outputs. The intermediate outputs are element-wise summed and activated using the sigmoid function, resulting in the channel attention feature $M_{\mathrm{c}}(\mathbf{F})$. The weighted output $\mathbf{F}_1$ is obtained by element-wise multiplication of $M_{\mathrm{c}}(\mathbf{F})$ with $\mathbf{F}$. Subsequently, $\mathbf{F}_1$ is input into the SAM. Within the SAM, the global maximum pooling and global average pooling operation are applied to the feature $\mathbf{F}_1$ respectively, generating two matrices with the same dimension. These matrices are then concatenated along the channel dimension, followed by convolution and sigmoid activation functions, resulting in the spatial attention feature $M_{\mathrm{s}}(\mathbf{F}_1)$. Finally, the refined feature map $\mathbf{F}_{\mathrm{refined}}$ is obtained

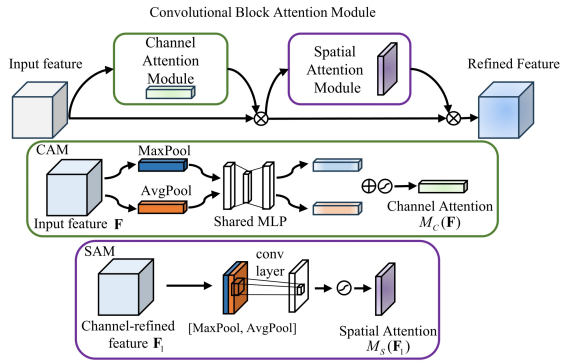


Fig. 3. Structure of CBAM.

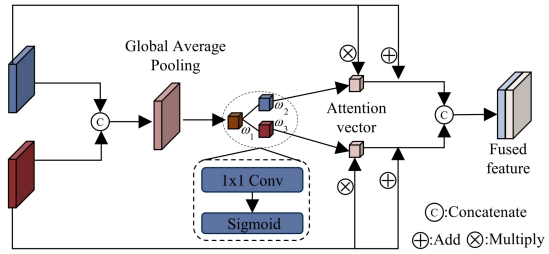


Fig. 4. Structure of AFFM.

by element-wise multiplication of $M_{\mathrm{s}}(\mathbf{F}_1)$ with $\mathbf{F}_1$. This entire process can be represented as

$$\begin{aligned} \mathbf{F}_1 &= M_{\mathrm{c}}(\mathbf{F}) \otimes \mathbf{F} \\ \mathbf{F}_{\mathrm{refine}} &= M_{\mathrm{s}}(\mathbf{F}_1) \otimes \mathbf{F}_1 \end{aligned} \tag{3}$$

where $\mathbf{F}$ denotes the initial feature map, $\otimes$ denotes element-wise multiplication, $\mathbf{F}_1$ is a process variable that represents the input of SAM. $M_{\mathrm{c}}(\mathbf{F})$ and $M_{\mathrm{s}}(\mathbf{F}_1)$ are the output of CAM and SAM, respectively, which can be expressed as

$$M_{\mathrm{c}}(\mathbf{F}) = \sigma(MLP(AvgPool(\mathbf{F})) + MLP(MaxPool(\mathbf{F}))) \tag{4}$$

$$M_{\mathrm{s}}(\mathbf{F_1}) = \sigma(f^{7\times7}([\mathbf{F_1}^{\mathrm{s}}_{\mathrm{avg}}; \mathbf{F_1}^{\mathrm{s}}_{\mathrm{max}}])). \tag{5}$$

It is evident that CBAM enables the model to learn the relationships between channels and the attention weights assigned to different regions, thus obtaining better feature representations and enhancing the performance of semantic segmentation. Additionally, the lightweight structure of CBAM does not significantly compromise the efficiency of semantic segmentation.

In the decoder of the DeepLabV3+ model, the features are fused in a direct concatenation method. However, since different channels contain distinct feature information that contributes differently to image segmentation, it is crucial to highlight the channels with greater contribution. To achieve this goal, a channel-level attention feature fusion module named AFFM is proposed, as illustrated in Fig. 4.

In AFFM, the input features $F_{\mathrm{low}} \in \mathbb{R}^{H\times W\times B}$ and $F_{\mathrm{muti}} \in \mathbb{R}^{H\times W\times C}$ are firstly concatenated to obtain feature $F_{\mathrm{con}} \in \mathbb{R}^{H\times W\times (B+C)}$. Then, the GAP operation is applied to F_{con} to obtain the feature $F_{\mathrm{GAP}} \in \mathbb{R}^{1\times1\times(B+C)}$, which is composed of the mean value of feature map in each channel. Afterwards, F_{GAP} passed through the weight modules, where each module is composed of a 1×1 convolution and a sigmoid function. These modules serve for channel-level compression and decompression, resulting in the generation of channel-level attention features. The channel-level attention features are multiplied element-wise with F_{low} and F_{muti} respectively to realize the weighted distribution of different channel features. The obtained features are added element by element with F_{low} and F_{muti} respectively to obtain the low-level feature with channel attention $F_{\mathrm{ca\_low}}$ and multi-scale feature with channel attention $F_{\mathrm{ca\_muti}}$, which are expressed as

$$F_{\mathrm{ca\_low}} = \omega_2\left(\omega_1\left(F_{GAP}\left(F_{\mathrm{low}} \odot F_{\mathrm{muti}}\right)\right)\right) \otimes F_{\mathrm{low}} \oplus F_{\mathrm{low}} \tag{6}$$

$$F_{\mathrm{ca\_muti}} = \omega_3\left(\omega_1\left(F_{GAP}\left(F_{\mathrm{low}} \odot F_{\mathrm{muti}}\right)\right)\right) \otimes F_{\mathrm{muti}} \oplus F_{\mathrm{muti}} \tag{7}$$

where F_{low} denotes the input low-level feature, F_{muti} denotes the input multi-scale feature, $\odot$ denotes concatenation, $\otimes$ denotes element-by-element multiplication, $\oplus$ denotes element-by-element addition.

Finally, the fusion feature with channel attention can be obtained by concatenating $F_{\mathrm{ca\_low}}$ and $F_{\mathrm{ca\_muti}}$, i.e.,

$$F_{\mathrm{fusion}} = F_{\mathrm{ca\_low}} \odot F_{\mathrm{ca\_muti}}. \tag{8}$$

Compared to the original direct concatenation fusion, AFFM embeds channel-level attention during the feature fusion process, avoiding the loss of important information during the fusion process, thereby improving the semantic segmentation performance of the model.

The introduction of these modules in the DeepLabV3+ model enhances its efficiency without losing significant segmentation accuracy. Based on this model, real-time semantic information about the scene can be acquired, ensuring the successful execution of subsequent viewpoint adjustments.

B. Semantic-Guided Viewpoint Adjustment Strategy

Based on the semantic information of the scene, a semantic-guided viewpoint adjustment strategy is proposed. This strategy enables the eye-in-hand installed camera to self-adjust to the optimal viewpoint that completely covers the area of the grasping task. The strategy is based on the following assumptions:

Assumption1: It is assumed that the longest distance between objects in the scene is within the camera's field of view, and the camera is installed on the robotic manipulator using an eye-in-hand method.

Assumption2: The camera has completely captured some objects, but the others are not fully captured.

Furthermore, the optimal viewpoint mentioned earlier refers to the specific camera viewpoint that captures an image encompassing all target objects. In the optimal viewpoint, the center of the image aligns with the optimal pixel point. The optimal pixel point is defined as the pixel point with the smallest sum of distances from the pixel points belonging to the object classes. Mathematically, it can be expressed as

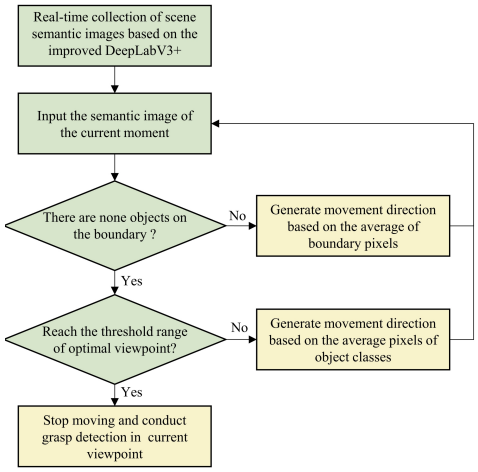


Fig. 5. Flowchart of the semantic-guided viewpoint adjustment strategy.

$$P_{\text{best}}(u,v) = \min \sum_{i=1}^{n} \left[(u-u_i)^2 + (v-v_i)^2 \right]^{\frac{1}{2}}. \quad (9)$$

The flowchart of the semantic-guided viewpoint adjustment strategy is presented in Fig. 5. The improved DeepLabV3+ is utilized to extract the semantic information of the scene. The current semantic image serves as the input to the viewpoint adjustment strategy. In this strategy, the first step is to check if there are objects along the image boundaries. If objects are found, the camera's motion direction is determined based on the average of the object pixels along the boundary. Then, a new semantic image obtained from the adjusted viewpoint is input to make the same determination. This process is repeated until no objects are found along the boundaries. When the objects are not found, the strategy continues by determining if the current viewpoint falls within the range of the optimal viewpoint. If the current viewpoint is within the optimal range, the subsequent grasping is performed directly from the current viewpoint. If the current viewpoint is not within the optimal range, the camera's motion direction is calculated based on the average of the object class pixels in the image. By inputting the semantic image obtained from the adjusted viewpoint, the above process is looped until the camera reaches the optimal viewpoint.

Under the above assumptions and definitions, the camera movement strategy is derived. Firstly, the semantic image obtained from the improved DeepLabV3+ is binarized, with the RGB value of pixels belonging to the background class set to 0, and the RGB value of pixels belonging to the object classes set to 1. Next, boundary detection is performed. Denote the image size as $h \times w$. Firstly, the top, left, bottom, and right boundaries of the semantic image are concatenated to obtain a $2 \times (h+w)$ size one-dimensional vector P_1, consisting only of 0 and 1. Then, we create a $2 \times (h+w)$ size one-dimensional vector where the value of each element within the vector is equal to its index. This vector is element-wise multiplied with P_1 to obtain the vector P_{position}. The elements in P_{position} reflect the position information of the object class along the boundary. Finally, the position of the target point on the boundary can be calculated by

$$p_{\text{goal}} = \sum_{i=0}^{2(h+w)} p_i/2 \quad (10)$$

where $p_i \in [0, 2(h+w)]$ is the element of P_{position}, h and w denote the height and width of the image, respectively, and p_{goal} is position value of the target point on a boundary. By adding and subtracting p_{goal} with h and w, the coordinates of the target point on a certain boundary can be obtained. Assuming that the target point is on the right boundary, the coordinates of the target point can be expressed as $(w, p_{\text{goal}} - w)$. The center of the image is connected to the target point to create a directed straight line. The direction of this line determines the camera's direction that needs to be adjusted at this moment.

Remark 1: The process described above utilizes the pixel-level information in semantic images to guide the camera's motion. Unlike previous methods, the proposed strategy employs two different methods to calculate the camera's motion direction in real time based on the distribution of pixels belonging to the object classes in the semantic image, thereby achieving efficient viewpoint adjustments.

By implementing the above movement strategy, the camera can gradually capture all the objects within the scene. When the camera's field of view encompasses all objects within the scene, it is directed towards the optimal viewpoint based on the optimal pixel point indicated by (9). Since the camera is eye-in-hand installed, hand-eye calibration is needed to establish a coordinate transformation relationship between the robotic manipulator end-effector and the camera. During the calibration process of the eye-in-hand system, the following transformation relationship can be established

$$T_{base}^{cal} = T_{base}^{tool} T_{tool}^{cam} T_{cam}^{cal} \quad (11)$$

where T_{base}^{tool} denotes the transformation matrix from the robotic manipulator end to the robotic manipulator base, which is readable by a demonstrator of the robotic manipulator. T_{tool}^{cam} denotes the transformation matrix from the camera to the robotic manipulator end, which is the requested transformation relationship. T_{cam}^{cal} denotes the transformation matrix from the calibration plate to the camera, which is an external parameter of the camera.

Since the relative positions of the calibration plate and robotic manipulator base T_{base}^{cal} remain fixed during the moving process, the equation can be formulated as

$$T_{base(1)}^{tool} T_{tool}^{cam} T_{cam(1)}^{cal} = T_{base(2)}^{tool} T_{tool}^{cam} T_{cam(2)}^{cal}. \quad (12)$$

Performing matrix multiplication, we get

$$T_{base(2)}^{tool-1} T_{base(1)}^{tool} T_{tool}^{cam} = T_{tool}^{cam} T_{cam(2)}^{cal} T_{cam(1)}^{cal-1}. \quad (13)$$

By marking $T_{base(2)}^{tool-1} T_{base(1)}^{tool}$ as *A*, $T_{cam(2)}^{cal} T_{cam(1)}^{cal-1}$ as *B*, and T_{cam}^{tool} as *X*, it can be seen that solving T_{cam}^{tool} is equal to solve [39]

$$AX = XB. \quad (14)$$

By solving (13), the transformation matrix from the camera to the robotic manipulator end can be obtained. The pose in the coordinate system of the robotic manipulator end can be described as

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = T_{tool}^{cam} \begin{bmatrix} x_{cam} \\ y_{cam} \\ z_{cam} \end{bmatrix} \tag{15}$$

where T_{tool}^{cam} denotes the transformation matrix from the camera to the robotic manipulator end, and $(x_{cam}, y_{cam}, z_{cam})^T$ denotes the pose in the camera coordinate system.

By following the above process, the motion of the camera can be translated into the corresponding motion of the robotic manipulator end-effector. This enables the camera to reach the optimal viewpoint, where the subsequent task-oriented grasp detection can be effectively performed.

C. Grasp Detection Structure Based on the Improved DeepLabV3+ and GRCNN

In Section II, it is mentioned that GRCNN can generate pixel-wise grasp representations of objects and then select the highest quality grasp representation as the final grasp representation. However, GRCNN is a task-agnostic grasp detection method, meaning it lacks the capability to generate grasp representations for specific objects in real scenes, particularly when the desired object is surrounded by other objects within the camera's field of view.

It is notable that our proposed improved DeepLabV3+ model accurately segments specific objects along their contours in densely cluttered scenes, avoiding the problem of detection box stacking in object detection algorithms. Therefore, a task-oriented grasp detection structure is proposed by combining improved DeepLabV3+ with GRCNN to generate a grasp representation for the specific object in such scenes, as depicted in Fig. 6. This structure utilizes the improved DeepLabV3+ to precisely segment various object categories and backgrounds within the images. The resulting mask of the specific object is then utilized to eliminate irrelevant regions and generate an image containing only the specific object. Finally, this image is input into GRCNN to generate the highest quality grasping pose for the specific object. The integration of improved DeepLabV3+ and GRCNN achieves precise target segmentation while mitigating the performance degradation issue of GRCNN caused by environmental interference. This greatly improves the quality of grasp detection results, thereby improving the grasp success rate.

Remark 2: Differing from existing methods [26], [27],[29], [30], the proposed improved DeepLabV3+-GRCNN grasp detection structure effectively addresses the shortcomings of GRCNN in specific scenarios by leveraging the strengths of improved DeepLabV3+. This innovative structure enables high-performance task-oriented grasp detection in densely cluttered scenes.

IV. Experiments Results

A. Experimental Setup

To validate the proposed task-oriented grasping framework guided by visual semantics, experiments were conducted using

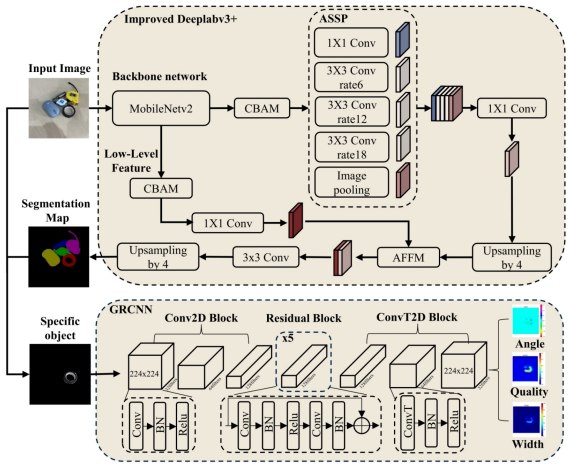


Fig. 6. Improved DeepLabV3+-GRCNN grasp detection structure.

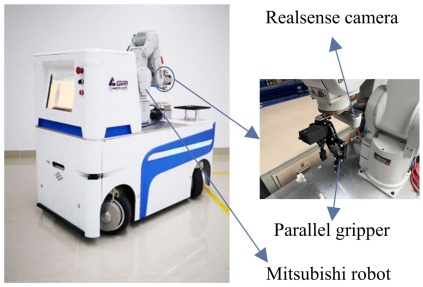


Fig. 7. Experimental platform.

a mobile manipulator developed in our laboratory, as depicted in Fig. 7. The mobile chassis adopts a four-wheel independent drive method and is equipped with a lidar for navigation to achieve movement in the workspace. Furthermore, the mobile platform is equipped with a Mitsubishi robotic manipulator featuring a parallel gripper at the end for object grasping. A Realsense D435i camera is also installed on the robotic manipulator end to capture RGB-D data. In addition, the proposed framework is implemented based on Pytorch with CUDA 11.0. The hardware system comprises an AMD Ryzen 7 5700X CPU and a NVIDIA GeForce RTX 3060 GPU. On this basis, the effectiveness of each module in our proposed framework is verified.

B. Verification of Improved DeepLabV3+ model

The performance of the improved DeepLabV3+ model is evaluated using three specific metrics: mean intersection over union (MIoU), mean pixel accuracy (MPA), and frames per second (FPS) [32]. Among them, MIoU and MPA are primarily employed to evaluate the performance of the model on the test set, while FPS indicates the speed at which the model processes images. MioU and MPA can be expressed as

$$\text{MIoU} = \frac{1}{n+1} \sum_{i=0}^{n} \frac{p_{ii}}{\sum_{j=0}^{n} p_{ij} + \sum_{j=0}^{n} p_{ji} - p_{ii}} \times 100 \tag{16}$$

TABLE I
PERFORMANCE COMPARISON BETWEEN THE PROPOSED MODEL AND THE BASELINE MODEL

Dataset	Model	MIoU%	MPA%	FPS(f/s)
VOC2012	Improved DeepLabV3+	82.41	89.09	30.32
Ours		94.13	97.08	38.13
VOC2012	DeepLabV3+	81.14	89.12	12.90
Ours		93.55	96.43	19.10

$$\mathrm{MPA} = \frac{1}{n+1}\sum_{i=0}^{n}\frac{p_{ii}}{\sum_{j=0}^{n}p_{ij}} \times 100 \tag{17}$$

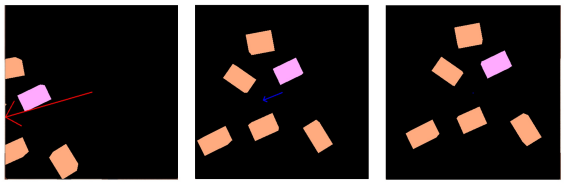
where n is the number of categories, p_{ij} is the number of pixels predicted to belong to category j but actually belong to category i, p_{ji} represents the number of pixels predicted to belong to category i but actually belong to category j, p_{ii} is the number of pixels successfully predicted.

Furthermore, the hyperparameters of the improved DeepLabV3+ model are set by referring to the parameter settings of DeepLabV3+ and incorporating some empirical values of network training hyperparameters. The training process involves a total of 200 epochs, with the initial 50 epochs dedicated to frozen training and a batch size of 6. This is followed by 150 epochs of unfrozen training, with a batch size of 8. Regarding the optimizer and learning rate, we have selected the stochastic gradient descent optimizer and set the initial learning rate to 0.007. To further enhance the training process, we have employed the Cosine learning rate update technique.

Based on the above parameter settings, the performance of the improved DeepLabV3+ model was then tested on two datasets: the VOC2012 dataset and a custom dataset created by our team. Our custom dataset includes various everyday objects arranged in different combinations and perspectives, covering scenarios with both dense and discrete objects. To augment the dataset, we utilized data augmentation methods, such as image rotation and noise addition, to increase the number of samples. Moreover, the images are annotated using the LabelMe annotation software and stored in the PASCAL VOC data format. Table I presents the performance comparison between the proposed model and the baseline model. It is evident that the introduction of MobileNetV2 as the backbone greatly enhances the processing speed of the proposed model. Moreover, the inclusion of CBAM and AFFM further ensures that the segmentation accuracy of the model remains at or slightly surpasses the baseline model. These improvements make the proposed model meet the demands for real-time semantic segmentation.

C. Verification of Semantic-Guided Viewpoint Adjustment Strategy

The effectiveness of the semantic-guided viewpoint adjustment strategy is initially verified through joint simulation conducted on the Webots and PyCharm platforms. The simulation results are presented in Fig. 8, where (a) to (c) illustrate the changes in the objects within the camera's field of view as it adjusts from the initial viewpoint to the optimal viewpoint. In Fig. 8 (a), the red arrow specifies the movement direction of the camera when objects are present on the boundary. In Fig. 8 (b), the blue arrow indicates the movement direction when no objects are present on the boundary. Fig. 8 (c) indicates that the camera has successfully reached the optimal viewpoint. These simulation results demonstrate the feasibility and effectiveness of the semantic-guided viewpoint adjustment strategy on the simulation platform.



(a) (b) (c)

Fig. 8. Adjustment process of camera viewpoint in simulation environment.

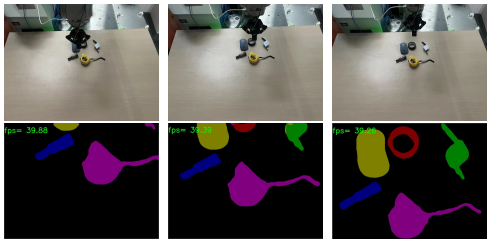


Fig. 9. Adjustment process of camera viewpoint in real environment.

Furthermore, the semantic-guided viewpoint adjustment strategy is implemented on the mobile manipulator presented in Fig. 7. The changes in the objects within the camera's field of view during the viewpoint adjustment process are partially captured in Fig. 9. It is evident from that the improved DeepLabv3+ maintains an average FPS of approximately 39 during the real machine adjustment process, providing real-time semantic information of the scene. With this visual semantic feedback, the proposed strategy can be successfully implemented, thus allowing the camera to adjust from an initial viewpoint where only a portion of the objects is captured to the optimal viewpoint.

D. Verification of Improved DeepLabV3+-GRCNN Structure

The performance of the improved DeepLabV3+-GRCNN structure is initially evaluated on the Cornell grasping dataset. We adopted a commonly used metric in related research [40] to determine the correctness of the prediction results, i.e., a predicted grasp is deemed correct when it simultaneously satisfies the following two conditions:

1) The difference between predicted grasp angle and the ground truth is less than 30°.
2) The intersection over union (IoU) fraction between the ground truth and predicted grasp rectangles should be greater than 25%. The IoU can be calculated by

$$\mathrm{IoU} = \frac{g_p \cap g_t}{g_p \cup g_t} \tag{18}$$

where g_p denotes the predicted grasp rectangle, g_t denotes the ground truth.

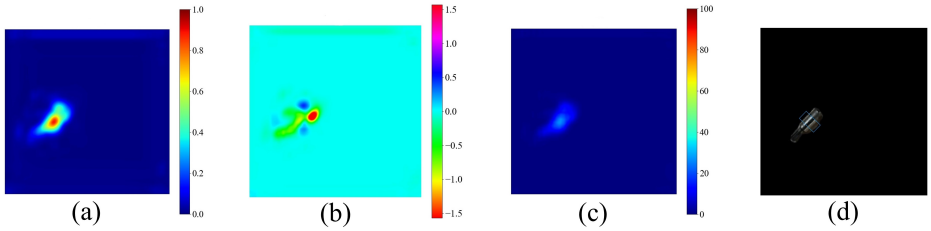


Fig. 10. Grasp detection results of improved DeepLabV3+-GRCNN. (a) Grasp quality. (b) Grasp angle. (c) Grasp width. (d) Grasp representation.

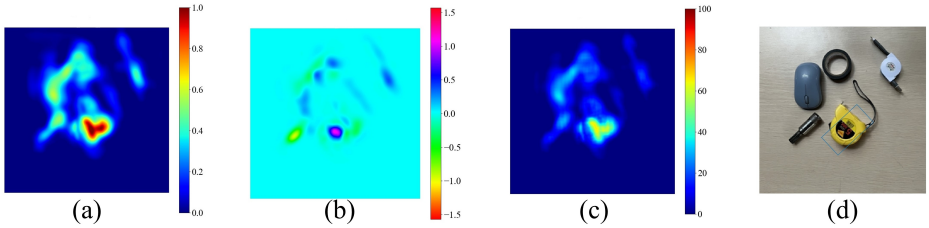


Fig. 11. Grasp detection results of GRCNN. (a) Grasp quality. (b) Grasp angle. (c) Grasp width. (d) Grasp representation.

Meanwhile, two data segmentation methods are used :

1) Image-wise split (IW): randomly split the dataset to verify the model's ability to generalize to different poses of objects.
2) Object-wise split (OW): randomly split the dataset by object category to verify the model's ability to generalize to new objects.

The performance of improved DeepLabV3+-GRCNN structure on the Cornell grasping dataset is shown in Table II, along with comparisons to other methods proposed in related research. It is important to note that, for all comparison methods, we have utilized the data reported in their original papers. As shown in Table II, our proposed DeepLabV3+-GRCNN structure achieves the highest accuracy of 98.3% and 97.5%, surpassing other methods and demonstrating its competitiveness with the state-of-the-art (SOTA) method[44]. While our performance on the grasping dataset is comparable to the SOTA method, it is worth mentioning that its transformer-based architecture presents significant challenges in terms of training and deployment. In contrast, our CNN-based method offers the advantage of being simpler to train and deploy. Furthermore, compared to the baseline model GR-CNN[15], the proposed structure exhibits a slight improvement in performance, indicating the positive impact of separating objects from the background on grasp detection.

To further verify the effectiveness of the improved DeepLabV3+-GRCNN structure on task-oriented grasp detection, a multi-object scenario is selected for testing. In this experiment, the robot system is provided with task information in the form of textual input specifying the category of objects to be grasped, specifically targeting a metal workpiece in this case. Based on this information, the improved DeepLabV3+ retains pixels corresponding to the specified category in the output results, generating an image that only contains the target object. This image is then utilized as input for the GRCNN to achieve task-oriented grasp detection. The detection results

TABLE II
PERFORMANCE EVALUATION ON THE CORNELL GRASPING DATASET

Method	Accuracy%	
	IW	OW
Liu et al.[19]	95.2	-
Cheng et al. [16]	95.4	-
Dong et al. [29]	96.4	96.5
Xu et al. [41]	96.9	95.7
Kumra et al. [15]	97.7	96.6
Dong et al. [42]	98.1	-
Wang et al. [43]	97.9	96.7
Zhang et al. [44]	98.3	96.9
Ours	98.3	97.5

are compared with those obtained only using GRCNN, as shown in Fig. 10 and Fig. 11, respectively. In these figures, (a)-(d) represent the grasp quality, grasp angle, grasp width, and the generated grasp representation, respectively. It is evident the proposed improved DeepLabV3+-GRCNN structure accurately segments the target object and generates an effective grasp representation for it. However, when relying solely on GRCNN, the maximum value of grasping quality is found on other interfering objects, hindering the generation of a grasp representation for the specific object. Through this comparison, it can be observed that the improved DeepLabV3+-GRCNN structure effectively mitigates the limitations of GR-CNN in task-oriented grasp detection.

The effectiveness of the proposed improved DeepLabV3+-GRCNN structure on task-oriented grasp detection is also verified in a densely cluttered scenario depicted in Fig. 12 (a).In this scenario, the ruler is chosen as the target object for grasp detection. Meanwhile, a comparison is made between the improved DeepLabV3+-GRCNN structure and the SSD[45]-GRCNN structure. The results of the SSD and the improved DeepLabV3+ are shown in Fig. 12 (b) and Fig. 12

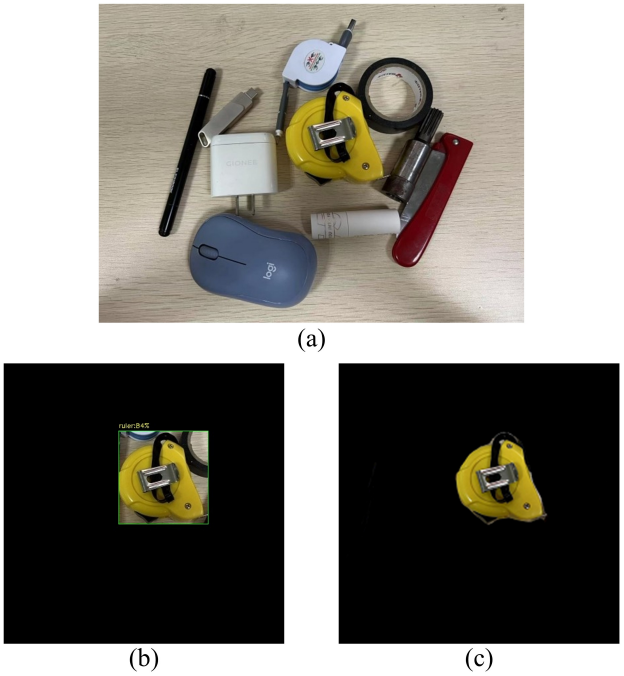


Fig. 12. (a) Densely cluttered scenario. (b) Detection results of SSD. (c) Segmentation results of improved DeepLabV3+ model.

(c), respectively. It is apparent from Fig. 12 that the target image extracted from the detection box generated by SSD contains information from other objects, introducing interference. In contrast, the target image generated by the improved DeepLabV3+ remains free from such interference caused by other objects. On this basis, both images are individually input into GRCNN, and the corresponding grasp detection results are shown in Fig. 13 and Fig. 14, respectively. One can see that the SSD-GRCNN structure generates an incorrect grasp representation for the target object due to interference from other objects. Conversely, the proposed improved DeepLabV3+-GRCNN structure generates an accurate grasp representation for the target object. This demonstrates the effectiveness of the proposed method for task-oriented grasp detection in densely cluttered scenarios.

E. -World Scenarios Grasping

The performance of our proposed structure in real-world scenarios is evaluated through physical grasp experiments. In these experiments, the tested objects were from our custom dataset consisting of 15 categories of daily necessities, covering both regular and irregular objects. The selection of scenes was based on relevant research, focusing on densely cluttered scenes with multiple objects. Furthermore, after each grasping task, we examined object contact within the scene.If no contact was detected between objects, we rearranged the scene and introduced objects that the robot had not previously grasped. This approach allowed us to thoroughly evaluate the performance of the proposed method across various scenarios. To make the results easier to compare and more general, we referred to the experimental settings in relevant research. We set

TABLE III
COMPARISON OF GRASPING METHODS

Method	Success rate(%)	Detection speed(s)
Kuleck et al. [27]	78.0	0.14
SSD-GRCNN	82.0	0.14
Dong et al. [30]	82.0	-
Li et al. [26]	86.0	0.11
Ours	91.0	0.13

the standard for successful grasping as stable grasping of the target object without falling and set the total number of grasping times to 100. Record the detection time of each algorithm, and take the average of these 100 times as the detection speed of the method. The task information in the experiment was transmitted in textual form from a host computer to the robot system. Upon receiving the task information, the robot system employed our proposed method. This method parsed the task information and then generated a grasp representation with the highest quality score for the target object. Subsequently, a series of coordinate transformations are made to this grasp representation to obtain the pose that the robotic manipulator needs to reach. Finally, this pose information was provided to the robotic manipulator controller to execute the grasping of the target object.

Based on the above experimental configuration and methods, the experiments are successfully implemented, and the partial capture results are shown in Fig. 15. It can be seen that our proposed method can accurately grasp the target object in densely cluttered scenes, and stably grasp it to a certain height, which to some extent demonstrates the effectiveness of the proposed method. Besides, the success rates of our proposed method and other task-oriented grasp methods on real robot grasping are presented in Table III. It can be observed that our proposed method performs well in densely cluttered scenes. Compared to the state-of-the-art method, our method exceeded its success rate by 5%, reaching 91%, while falling behind in detection speed by only 0.02 seconds, at 0.13 seconds. The significant improvement in the success rate of grasping is primarily due to the pixel segmentation feature of our proposed method. This feature enables accurate segmentation of the target object and ensures that the grasp detection algorithm is not influenced by other environmental factors during the grasping detection process, thereby maximizing its performance. In terms of detection speed, although we have optimized the detection speed of DeepLabV3+, we are still at a disadvantage compared to lightweight object detection algorithms such as YOLO, which is also one of the directions we need to optimize in the future. These comparisons further demonstrate the effectiveness and practicality of the method proposed in this paper.

Finally, we conducted additional tests in the scenarios with stacked objects to investigate the influence of object stacking on our proposed method. The arrangement of some stacked scenes are shown in Fig. 16. During the experiment, we found that the semantic segmentation module can accurately segment objects according to their contours in the scene with stacked object. However, occlusion between objects lead to incomplete

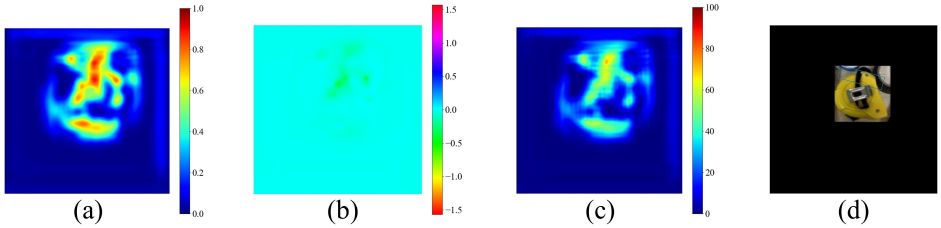


Fig. 13. Grasp detection results of SSD-GRCNN. (a) Grasp quality. (b) Grasp angle. (c) Grasp width. (d) Grasp representation.

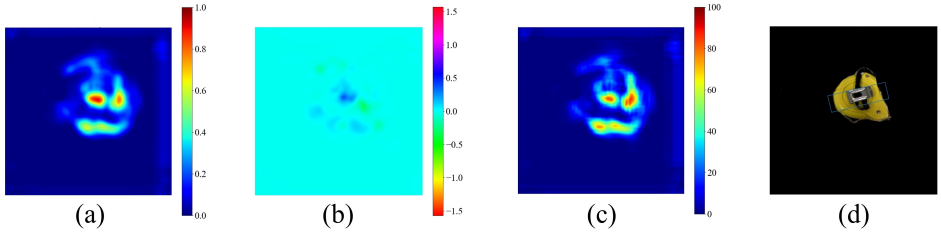


Fig. 14. Grasp detection results of the improved DeepLabV3+-GRCNN. (a) Grasp quality. (b) Grasp angle. (c) Grasp width. (d) Grasp representation.

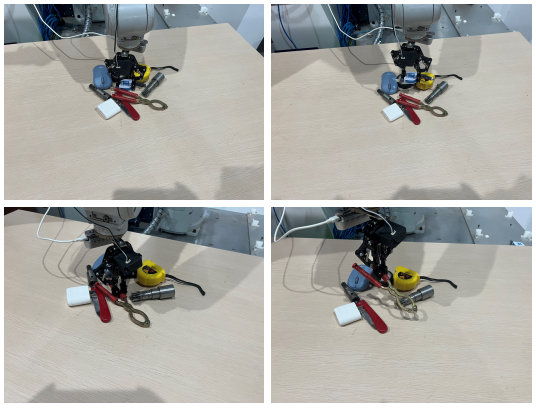


Fig. 15. Grasping effect in the densely cluttered scenario.

masks for certain object categories, potentially affecting the subsequent grasp detection. For the grasp detection module, it can successfully generate grasp representations for objects that are partially occluded. However, when objects were heavily occluded, the module generally struggled to generate suitable grasp representations. We recorded the results of 100 grasp attempts, achieving a success rate of 80%. This outcome indicates that the proposed method demonstrates a certain level of effectiveness in some stacked scenes. Additionally, we also accomplished task-oriented grasping in severely stacked scenes, where objects were stacked from top to bottom, by employing human strategies during the experiment. However, it is important to note that the decision-making process regarding the grasp order in such scenes is not within the scope of this research. Therefore, we will not delve into further discussion on this aspect in this paper.

V. Conclusion

In this paper, we proposed a task-oriented grasping framework guided by visual semantics to achieve task-oriented grasping in densely cluttered scenarios with the absence of

Fig. 16. Scenarios with stacked objects in the experiment.

object information. The DeepLabV3+ model was first modified by introducing the Mobilenetv2, CBAM, and AFFM. These improvements significantly enhance the speed of the DeepLabV3+ model while maintaining its original segmentation accuracy, enabling the model to acquire real-time semantic information about the scene. On this basis, a semantic-guided camera viewpoint adjustment strategy is proposed. This strategy enables the camera to self-adjust to the optimal viewpoint, effectively resolving the issue of the absence of object information in the grasping task area. Finally, an improved DeepLabV3+-GRCNN structure is proposed. In this structure, the object image free from external interference is provided for the grasp detection network by segmenting objects along their contours, thus improving the success rate of task-oriented grasp detection in densely cluttered scenarios. The experimental results validate the effectiveness of our proposed task-oriented grasping framework guided by visual semantics. Compared to the method with the highest success rate of 86% reported in relevant research, our proposed framework achieved a grasp success rate of 91% in densely cluttered scenarios, demonstrating certain advantages.

While our method has achieved significant results in the aforementioned scenarios, we did not consider the objects stacking in the process, which is also another primary cause of information absence. Therefore, our future research direction is to design an object manipulation relationship detection

network that can provide the appropriate grasping order based on the arrangement of objects in the scene, particularly in stacking scenarios. By integrating this network into the framework proposed in this paper, it is expected to achieve more accurate and efficient task-oriented grasping in object stacking scenarios.

REFERENCES

[1] Y. L. Xie, X. Zhang, S. Zheng, C. K. Ahn, and S. Wang, "Asynchronous H_∞ continuous stabilization of mode-dependent switched mobile robot," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 52, no. 11, pp. 6906-6920, Nov. 2022.

[2] J. Meng *et al.*, "Efficient and reliable lidar-based global localization of mobile robots using multiscale/resolution maps," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1-15, July 2021.

[3] L. Q. Jiang *et al.*, "Decoupled fractional supertwisting stabilization of interconnected mobile robot under harsh terrain conditions," *IEEE Trans. Ind. Electron.*, vol. 69, no. 8, pp. 8178-8189, Aug. 2022.

[4] Y. H. Li, Y. Liu, Z. Q. Ma, and P. F. Huang, "A novel generative convolutional neural network for robot grasp detection on gaussian guidance," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1-10, Aug. 2022.

[5] G. B. Wu, W. S. Chen, H. Cheng, W. M. Zuo, D. Zhang, and J. You, "Multi-object grasping detection with hierarchical feature fusion," *IEEE Access*, vol. 7, pp. 43884-43894, May 2019.

[6] F. Peng, Q. Y. Xu, Y. F. Li, M. X. Zheng, and H. Su, "Improved kernel correlation filter based moving target tracking for robot grasping," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1-12, Aug. 2022.

[7] Q. C. Zheng, Z. Peng, P. H. Zhu, Y. Y. Zhao, R. Zhai, and W. P. Ma, "An object recognition grasping approach using proximal policy optimization with YOLOv5," *IEEE Access*, vol. 11, pp. 87330-87343, Aug. 2023.

[8] H. K. Tian, K. C. Song, S. Li, S. Ma, J. Xu, and Y. H. Yan, "Data-driven robotic visual grasp detection for unknown objects: A problem-oriented review," *Expert Syst. Appl.*, vol. 211, pp. 118624, Aug. 2023.

[9] B. Dai, Y. Q. Zhang, and D. H. Lin, "Detecting visual relationships with deep relational networks," *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3076-3086, Apr. 2017.

[10] Y. X. Yang, Z. H. Ni, M. Y. Gao, J. Zhang, and D. C. Tao, "Collaborative pushing and grasping of tightly stacked objects via deep reinforcement learning," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 1, pp. 135-145, Jan. 2022.

[11] E. B. Li, H. B. Feng, S. Y. Zhang, and Y. L. Fu, "Learning target-oriented push-grasping synergy in clutter with action space decoupling," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 11966-11973, Oct. 2022.

[12] S. Q. Duan, G. H. Tian, Z. L. Wang, S. P. Liu, and C. R. Feng, "A semantic robotic grasping framework based on multi-task learning in stacking scenes," *Eng. Appl. Artif. Intel.*, vol. 121, pp. 106059, Feb. 2023.

[13] Y. J. Laili, Z. L. Chen, L. Ren, X. K. Wang, and M. J. Deen, "Custom grasping: A region-based robotic grasping detection method in industrial cyber-physical systems," *IEEE Trans. Autom. Sci. Eng.*, vol. 20, no. 1, pp. 88-100, Jan. 2023.

[14] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robot. Res.*, vol. 39, nos. 2-3, pp. 183-201, June 2020.

[15] S. Kumra, S. Joshi and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," *IEEE Int. Conf. Intell. Robots Syst. (IROS)*, pp. 9626-9633, Feb. 2021.

[16] H. Cheng, Y. Y. Wang, and M. Q. H. Meng, "A robot grasping system with single-stage anchor-free deep grasp detector," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1-12, Apr. 2022.

[17] H. K. Tian, K. C. Song, S. Li, S. Ma, and Y. H. Yan, "Lightweight pixel-wise generative robot grasping detection based on RGB-D dense fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1-12, Aug. 2022.

[18] Y. Zhang, L. H. Xie, Y. H. Li, and Y. Li, "A neural learning approach for simultaneous object detection and grasp detection in cluttered scenes," *Front. Comput. Neurosci.*, vol. 17, pp. 1110899, Oct. 2023.

[19] D. Liu, X. T. Tao, L. H. Yuan, Y. Du, and M. Cong, "Robotic objects detection and grasping in clutter based on cascaded deep convolutional neural network," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1-10, Nov. 2021.

[20] Y. Sun, X. X. Wang, Q. X. Lin, J. H. Shan, S. L. Jia, and W. W. Ye, "A high-accuracy positioning method for mobile robotic grasping with monocular vision and long-distance deviation," *Measurement*, vol. 215, pp. 112829, Apr. 2023.

[21] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," 2017, *arXiv:1711.00199*.

[22] M. Danielczuk *et al.*, "Mechanical search: Multi-step retrieval of a target object occluded by clutter," *Int. Conf. Robot. Autom. (ICRA)*, pp. 1614-1621, May 2019.

[23] Y. Y. Yu, Z. Q. Cao, S. Liang, W. J. Geng, and J. Z. Yu, "A novel vision-based grasping method under occlusion for manipulating robotic system," *IEEE Sens. J.*, vol. 20, no. 18, pp. 10996-11006, May 2020.

[24] H. Ren and A. H. Qureshi, "Robot active neural sensing and planning in unknown cluttered environments," *IEEE Trans. Robot.*, vol. 39, no. 4, pp. 2738-2750, Aug. 2023.

[25] Rasouli, A., Lanillos, P., Cheng, G., and J. K. Tsotsos, "Attention-based active visual search for mobile robots," *Auton. Robot.*, vol. 44, pp. 131–146, July, 2020.

[26] Z. Li *et al.*, "A YOLO-GGCNN based grasping framework for mobile robots in unknown environments," *Expert Syst. Appl.*, vol. 225, pp. 119993, Apr. 2023.

[27] B. Kuleck, K. Młodzikowski, R. Staszak, and D. Belter, "Practical aspects of detection and grasping objects by a mobile manipulating robot," *Ind. Rob.*, vol. 48, no. 5, pp. 688-699, Mar. 2021.

[28] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB," *IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 13452-13458, Oct. 2021.

[29] M. S. Dong, S. M. Wei, X. L. Yu, and J. Q. Yin, "MASK-GD segmentation based robotic grasp detection," *Comput. Commun.*, vol. 178, pp. 124-130, Jan. 2021.

[30] M. S. Dong, Y. X. Bai, S. M. Wei, and X. L. Yu, "Real-World Semantic Grasp Detection Using Ontology Features: Learning to Concentrate on Object Features," *Neural Process. Lett.*, vol. 55, pp. 8419–8439, June 2023.

[31] Q. P. Li and Y. Y. Kong, "An improved SAR image semantic segmentation DeepLabV3+ model network based on the feature post-processing module," *Remote Sens.*, vol. 15, no. 8, pp. 2153, Apr. 2023.

[32] J. Y. He *et al.*, "Method for segmentation of banana crown based on improved DeepLabV3+ model," *Agronomy*, vol. 13, no. 7, pp. 1838, July 2023.

[33] L. C. Chen, Y. K. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," *Eur. Conf. Comput. Vis.(ECCV)*, pp. 801-818, Oct. 2018.

[34] Redmon J and Angelova A, "Real-time grasp detection using convolutional neural networks," *IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 1316-1322, May 2015.

[35] K. Hara, R. Vemulapalli, and R. ChellappaHe, "Designing deep convolutional neural networks for continuous object orientation estimation," 2017, *arXiv:1702.01499*.

[36] L. F. Mo, Y. S. Fan, G. Y. Wang, X. M. Yi, X. M. Wu, and P. Wu, "DeepMDSCBA: An improved semantic segmentation model based on DeepLabV3+ model for apple images," *Foods*, vol. 11, no. 24, pp. 3999, Dec. 2022.

[37] S. Yu, D. H. Zhai, Y. Q. Xia, H. R. Wu, and J. Liao, "SE-ResUNet: A novel robotic grasp detection method," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 5238-5245, Apr. 2022.

[38] S. Woo, J. Park, J.Y. Lee, and I.S. Kweon, "CBAM: Convolutional block attention module," *Eur. Conf. Comput. Vis. (ECCV)*, pp. 3-19, Oct. 2018.

[39] R. Horaud and F. Dornaika, "Hand-eye calibration," *Int. J. Robot. Res.*, vol. 14, no. 3, pp. 195-210, June 1995.

[40] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," *IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 3304-3311, Aug. 2011.

[41] R. N. Xu, F-J. Chu, and P. A. Vela, "GKNet: Grasp keypoint network for grasp candidates detection," *Int. J. Rob. Res.*, vol. 41, no. 4, pp. 361-389, Feb. 2022.

[42] M. S. Dong, Y. X. Bai, S. M. Wei, and X. L. Yu, "Robotic Grasp Detection Based on Transformer," *Int. Conf. Intel. Robot. Appl. (ICIRA)*, pp. 323-328, Aug. 2022.

[43] S. Wang, Z. Zhou and Z. Kan, "When Transformer Meets Robotic Grasping: Exploits Context for Efficient Grasp Detection," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 8170-8177, July 2022.

[44] Q. Zhang, J. W. Zhu, X. Y. Sun, and M. M. Liu, " HTC-Grasp: A Hybrid Transformer-CNN Architecture for Robotic Grasp Detection," *Electronics*, vol. 12, pp. 1505, Mar. 2023.

[45] W. Liu *et al.*, "SSD: Single shot multibox detector," *Eur. Conf. Comput. Vis.(ECCV)*, pp. 21-37, Sep. 2016.

Guangzheng Zhang received the M. S. degree in mechanical engineering from Guangxi University (GXU), Nanning, China, in 2022. He is currently working toward the Ph.D. degree in mechanical engineering with the School of Mechanical Science and Engineering of Huazhong University of Science and Technology (HUST). His research interests include deep learning, robot dynamics, and robot visual grasping.

Shuting Wang received the B.S. and M.S. degrees from the school of energy and power engineering Wuhan University of Technology, Wuhan, China, in 1991 and 1994, respectively, and the Ph.D. degree from the School of Mechanical Science and Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China, 2002.

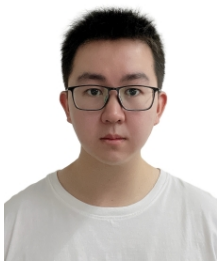
He is currently a full professor and Vice-president with the School of Mechanical Science and Engineering of HUST. He has published more than 100 articles. His research interests include mobile robot, mechanical design, and numerical control.

Yuanlong Xie (Senior Member, IEEE) received his B.S.degree in electrical engineering and Ph.D.degree in mechanical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2014 and 2018, respectively.

He was an Academic Visitor with the School of Electronic and Electrical Engineering, University of Leeds, Leeds, U.K., from 2017 to 2018. In Nov. 2018, he joined the HUST as a Postdoctoral Fellow and became an associate professor in Jan. 2024. He has published more than 120 academic journal and conference papers, and holds more than 60 patents. He serves as guest editor of some leading international journals, including Sensors, IET Collaborative Intelligent Manufacturing. His research interests include mobile robot, robot control and servo control.

Sheng Quan Xie (Fellow, IEEE) received the Ph.D. degrees in mechatronics engineering from Huazhong University of Science and Technology, Wuhan, China and in mechanical engineering from the University of Canterbury, Christchurch, New Zealand, in 1998 and 2002, respectively.

In 2003, he joined the University of Auckland and became a Chair Professor of (Bio) Mechatronics in 2011. Since 2017, he has been the Chair Professor of Robotics and Autonomous Systems with the University of Leeds, Leeds, U.K. He has authored or coauthored 8 books, 15 book chapters, and more than 400 international journal and conference papers. His research interests include medical and rehabilitation robots and advanced robot control. Prof. Xie is an elected Fellow of Engineers New Zealand. He has also served as a Technical Editor for IEEE/ASME TRANSACTIONS ON MECHATRONICS.



Yiming Hu received the B. S. degree in mechanical engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2017. He is currently pursuing the Ph.D. degree in mechanical engineering and automation. His research mainly focuses on solving mapless mobile robot navigation problems via deep reinforcement learning.

Tifan Xiong received the Ph.D. degree from the School of Mechanical Science and Engineering, Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007. He is currently a lecturer with the School of Mechanical Science and Engineering of HUST. He has published more than 10 articles. His research interests include mobile robot, robot control, robot visual grasping.