Research Status of Robot Grab Detection Based on Vision

Abstract: With the progress of science and technology and the continuous improvement of ~~people's~~ living standards, robots are more and more widely used in life and production, and robot grasping technology is also constantly improving. In practical application, accurate graspdetection of target objects is an important part of robot grasping tasks. In this paper, the parallel two-fingered gripper is used as the end of the robot arm's grasping, and the research status of grasping detection, which is the key part in the grasping process of the robot arm based on vision, is summarized. The 2D planar grasping and 6-DOF spatial grasping are compared and analyzed in detail. At the same time, it also summarizes the commonly used evaluation indexes of capturing data sets and capturing detection, and points out the challenges faced by vision-based robot capturing and the future direction of solving these challenges.

Keywords: robot;grasp detection; 2D planar grasping; 6-DOF spatial grasping

1 Introduction

In recent years, it has become normal for robots to work instead of people. Among them, the importance of robot grasping is self-evident. Kumra and Kanan[1] think that the robot grasping system consists of three parts: grasping detection system, grasping planning system and control system.Among them, grabbing detection is to process the image information and point cloud information collected by the vision device through relevant algorithms, and generate the pose corresponding to the end effector's ability to grab the target object[2].To provide reliable perception information for the subsequent robot planning and robot control process, to achieve successful grabbing. As shown in Fig. 1, it is a simulation experiment platform composed of an RGB-D camera, an ur manipulator and the object to be grabbed; Fig. 2 is the flow chart of the grab detection system, including target location, target posture estimation, grip posture estimation, and finally grab execution.

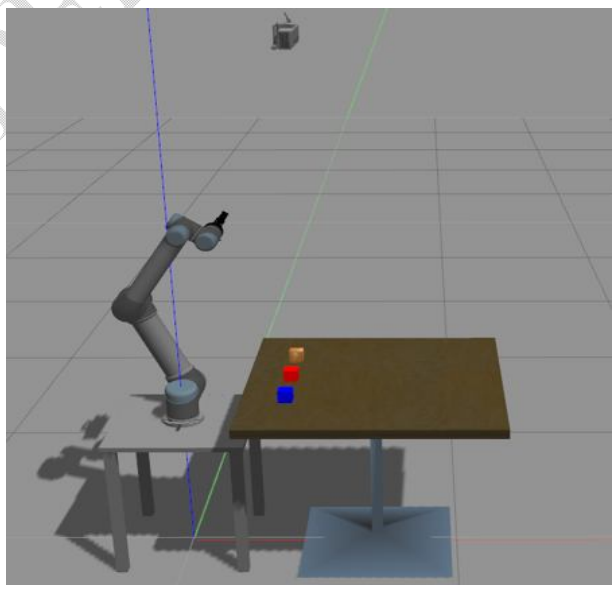

Fig. 1. mechanical arm grasping simulation platform

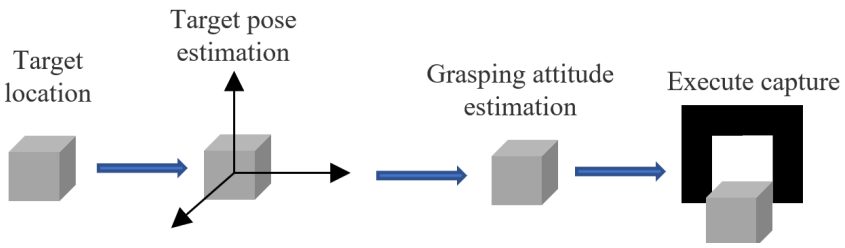

Fig. 2. Flow chart of grab detection system

Many factors need to be considered in the capture task, such as the physical properties of the object, the type of capture terminal, etc. Therefore, different grabbing ends are often used in different task scenarios, and the most typical ones are suction type and grippertype, as shown in Fig. 3. Among them, the grasping detection method using aparallel two-finger gripper has been widely studied because of its simple and dexterous characteristics. The main feature of this device is that the gripper is used as the gripper, and the hands are parallel. Because it has the important advantages of low cost and simple and convenient maintenance, it is widely used.

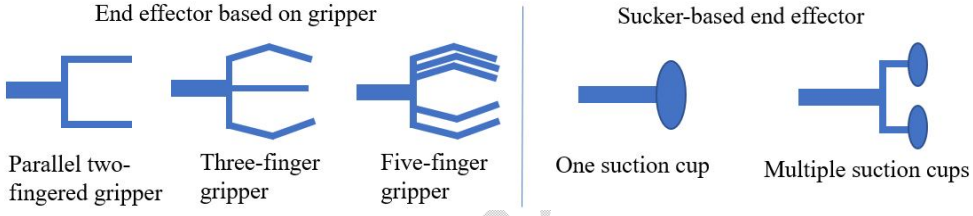

Fig. 3. Different kinds of end actuators. (left) grippertype, (right) suction type

Robot grab detection technology can be divided into two categories[3].One is the analysis method based on physical constraints, also known as thehard coding method[4], and the other is theempirical method based on prior information. The analysis method refers to analyzing the target model according to various parameters of the manipulator, using mathematical and physical models in geometry, motion and dynamics to calculate, and feeding back the appropriate grasping posture for the grasping task. The analysis method has a solid theoretical foundation, but the deficiency is that the model of the robot end effector and the grasping object in the real three-dimensional world is very complex, and it is difficult to realize a high-precision model. In contrast, the empirical method does not depend on the real-world modeling method. It uses the previous successful grasping experience to detect the grasping posture and judge its rationality by different methods. According to the characteristics of the target object, it uses similarity to classify and estimate the position and posture of the target object, to achieve the goal of grasping. With the continuous development of robot grasping technology and the increasing number of grasping methods and applications, grasping detection technology can be divided into 2D plane grasping method and 6-DOF space grasping method. For 2D plane grasping, the grasping direction is constrained to one direction, and the 6D grasping posture can be simplified to 3D representation, including 2D plane position and 1D rotation angle, so the height and rotation along other axes are fixed; For 6-DOF spatial grasping, the gripper can grasp objects from different angles, so the key to grasping is to obtain the accurate 6-day grasping posture of the gripper. This paper introduces the current research in detail from these two directions.

2 2D plane grab

2D grasping means that the target object is laid flat on the workbench, and the robot's end actuator only grasps it from one direction. 2D grasping methods can be divided into grasping

contact point detection and directed rectangle detection. In grasping contact point detection, the grasping position detection of the target object is the detection of its surface grasping point position, and the grasping posture of the end actuator is uniquely determined by the grasping contact point. However, in the grasping detection of the directional rectangle, the grasping posture of the end effector is uniquely determined by the directional rectangle frame.

2.1 Grab the contact point detection method

Generally, the method of grabbing contact points detection is aimed at objects with a certain shape, and the grabbing contact points are selected from candidate samples, and the possibility of successful grabbing of candidate points is evaluated by analytical method or method based on deep learning. Grab representation based on contact points is composed of grab quality, grab center, grab direction and grab width. As shown in fig. 4, where (u, v) represents the center point of the grab, and w represents the width of the gripper and $\emptyset$ represents the grasping direction. In the early research, Paiter[5] used k-means algorithm to detect the grab point position of simple objects; In 2014, Domae et al. [6] represented the grabber model by using two mask images. One image represented the contact area filled by the target object to achieve stable grabbing, and the other image represented the collision area during grabbing. Grapability metrics were calculated by convolving the mask images with the binarized depth map.

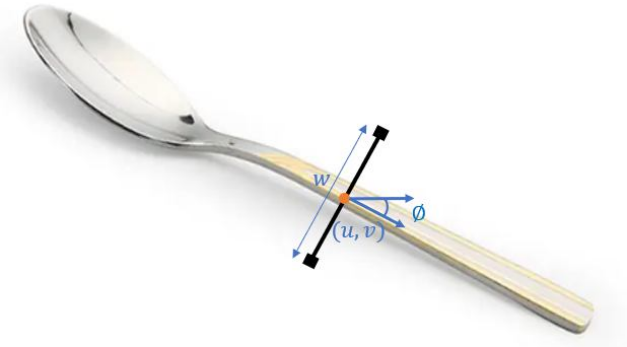

Fig. 4. grab representation based on touch point grab detection

With the development of deep learning, the method based on deep learning can help to evaluate the grasping quality of candidate grasping contact points. Mahler et al. [7] proposed a data set Dex-Net1.0 for 3D object model in 2016, which uses multi-view convolutional neural network (MV-CNNs) to provide similarity measure through 3D object classification. The following year, Mahler et al. [8] used depth images to detect the edge of objects, randomly sampled points on the edge of objects to form Grab candidate frames, rotated and scaled depth image blocks according to Grab candidate frames, and used the depth map combining depth image blocks with grab depth as the input of their proposed grab quality convolutional neural network (GQ-CNN) to predict whether the grab was successful or not, and finally grabbed by sorting. To train the network, they randomly placed thousands of 3D models in Dex-Net 1.0 to generate a Dex-Net 2.0 data set containing 6.7 million samples. In this method, the capturing angle of the image and the features of the image are coupled together, which reduces the learning demand of convolutional neural network for capturing features from different angles, thus reducing the learning difficulty and improving the performance of the network.

The method based on deep learning can also predict the most likely grabbing contact point by estimating the grabbing distance at pixel level. In some research of grab detection, from the point of view of image target position, focus on the information of image target area to predict the position of grab contact point. In 2018, Do et al. [9] proposed the deep learning network

AffordanxeNet, which has two branches. One branch is used to locate and classify objects, and the other branch is used to assign each pixel in an object to its most likely contact point label. The network adopts three key components: deconvolution layer sequence, robust adjustment strategy and multi-task loss function to effectively deal with the problem of grabbing the contact mask. In 2019, chu et al. [10] proposed the AffContext, which has nothing to do with the object category, to predict the grabbing position of the robot operation. This AFF Context is used to extract the example areas of images across categories. Each proposal is evaluated by training the self-attention mechanism of the rich semantic features of this area, and the candidate grabbing objects are generated according to each proposal, which improves the flexibility of robot operation. In 2021, Cao et al. [11] used the lightweight model, which introduced the grab representation of Gaussian kernel to encode the training samples, so as to highlight the maximum grab quality at the center. At the same time, in order to extract multi-scale information and enhance the distinguishability of features, the receptive field module (RFB) is added to enhance the feature extraction ability of the network. Combining pixel and channel attention, the features of the captured object are further highlighted by suppressing noise features. In 2022, Xu et al. [12] proposed Grab Network (GKN) based on key point detection. Each candidate capture point in the network is detected as a pair of key points, and the difficulty of detection is reduced by grouping the key points into pairs. In order to ensure the correspondence between key points, the filtering strategy based on direction prediction eliminates the wrong correspondence and improves the detection performance. Their GKN network achieves a good balance between accuracy and real-time performance on Cornell and Jacquard data sets. These methods are similar to segmentation problems, so the quality of candidate capture strictly depends on the segmentation accuracy of the picture.

In addition, some researchers generate grasping quality pixel by pixel for robot grasping actions, and perform the highest quality grasping contact points. In 2017, Zeng et al. [13] inferred dense pixel-level probability graphs for four different grabbing actions through the full convolution network, and performed the action with the highest score. It performs well in grasping unknown objects in chaotic environment, and has good generalization. In 2019, Cai et al. [14] collected effective grasping samples by correcting the robot's grasping action, and predicted the grasping contact points at pixel level according to the network designed by Cai et al. [13], which used a full convolution residual network similar to Zeng et al. [13] to learn the paw grasping mode; Neither of these two methods can segment the target object, nor predict the pixel-by-pixel forward graph for each pixel, so it is a direct method to estimate the quality of grabbing, and it is not necessary to sample the candidate grabbing objects. Morrisonet al. [15]proposed Generating Grab Convolutional Neural Network (GG-CNN) in 2018, which predicts the grab quality and posture of each pixel, while avoiding discrete sampling of candidate grabs and shortening the calculation time. Then, aiming at the problem of single data when the camera position is fixed or the viewing angle of the target object is fixed, they put forward a multi-view selection (MVP) controller [16], which uses the active sensing method to select the real-time grabbing posture with rich viewing angles. In 2020, Morrison et al. [17] proposed GG-CNN2 on the basis of GG-CNN[15], which changed the size of filters, the number of filters and the size of expansion convolution compared with the benchmark.

2.2 Detection method of directed rectangle

Because the method based on grabbing contact point detection shows a good grabbing effect

when the object is in a certain shape, the detection effect is not ideal when the shape of the object is irregular, and the generalization performance is poor. In 2011, Jiang et al. [18] first proposed a directed rectangle as the grasping representation of the gripper in order to obtain a more comprehensive grasping representation of the robot with more information, as shown in Fig. 5. The directed rectangle is determined by a five-dimensional vector (x, y, w, h, θ), where (x, y) and (h, w)respectively represent the coordinates of the center point of the grab and the width and length of the grab rectangle, and $\theta \in (0°, 180°)$ represents the clockwise rotation angle of the rectangular frame, that is, the grab angle. With the determination of these parameters, the problem of detecting and grabbing the whole object is transformed into the problem of finding five vectors in the image. With the development of deep learning, detection methods based on directed rectangle mainly include three categories. classification-based method,regression-based method and detection-based method.

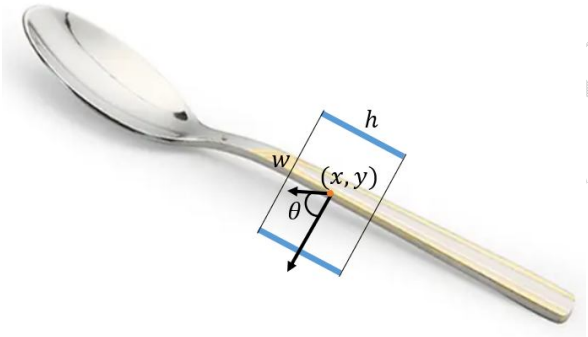

Fig. 5. grab representation based on directed rectangle grab detection

In 2015, Lenz et al. [19] applied the deep learning method to robot grasping for the first time, and proposed a sparse automatic encoder (SAE) to train the network. Multimodal information (RGB image, depth image and surface norm information) was used to sort the candidate captured image blocks, and the sliding window was used for detection. A two-step cascade structure with two depth networks was used, and the detection results of the first network were continuously evaluated by the second network. However, due to the time-consuming sampling process of sliding window, the speed of this method is slow. In 2016, Wang et al. [20] proposed a two-stage grab detection method based on real-time classification, which uses multiple sparse automatic encoders (SAE) for classification. Compared with the work of Lenz et al., Wang et al. generated the candidate capture objects faster, reduced the search range of the candidate objects by using a variety of prior information and preprocessing, and also reduced the parameters of the candidate capture to be estimated. However, this model does not support the end-to-end learning of the candidate capture. In 2017, Asif et al. [21] layered the input image, obtained the object category and grab attitude probability through different levels of calculation, and fused the results to infer the category and grab attitude of unknown objects. This method improves the accuracy of candidate capture, but the network training takes a long time. In 2018, Park and Chun [22] proposed a multilevel spatial transformation network (STN), which uses STN and depth residual block instead of sliding window. Their method allows the observation of intermediate results, such as the grab position and direction of many grab configuration candidates.

These methods belong to directed rectangle detection methods based on classification. The process of grab detection based on classification is shown in Fig. 6. This method uses the classifier to evaluate the grab position and direction of the image target area and generate candidate grab, and selects the candidate with the highest score to perform grab. The method is simple, and the

intermediate process of capturing parameter generation is visible, which is convenient to judge the quality of candidate capturing and has high accuracy. However, this method is time-consuming because the sliding window search time of candidate objects is long.

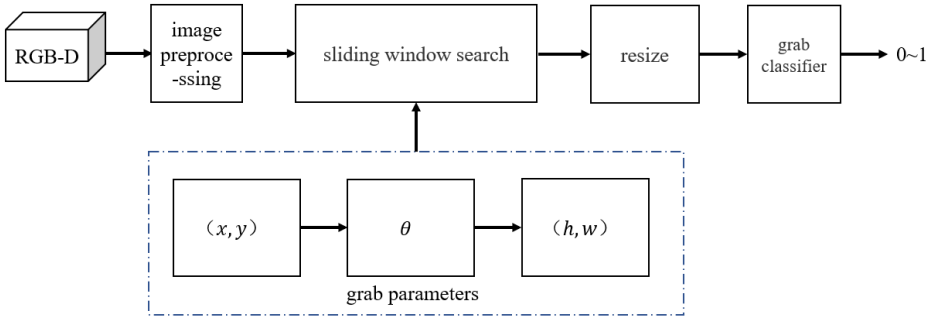

Fig. 6. Flow chart of grab detection based on classification

In 2015, Remdom et al. [23] introduced real-time detection through single-stage regression. This method uses the global features of the image to perform single-stage regression on the directed rectangular box, without using standard sliding window or area proposal techniques that focus on local area information. This method significantly improves the performance of grab detection in accuracy and time. Kumra et al. [24] proposed a deep convolution neural network (DCNN). The network uses RGB-D images as the input of the deep convolution neural network. Two pre-trained ResNet-50 networks are used to extract the features of RGB images and depth images, respectively. The neural network with three fully connected layers fuses these feature vectors, so as to generate capture parameters. In 2019, Zhang et al. [25] combined regression problem with classification problem, and fused RGB images and depth information in CNN model to achieve accurate feature expression. They proposed a new robust loss function Welsch function, which reduced the back propagation of discrete and low-contribution internal points and enhanced the robustness of the training process. Compared with the algorithm proposed by Redmon[23], the experimental results have greatly improved the accuracy and speed of image segmentation and object segmentation. In the same year, Kurmra et al. [26] proposed a convolution neural network (GR-CNN) for generating residuals. The addition of residual blocks not only avoids the over-fitting problem caused by too deep network layers, but also ensures the accuracy of network training. Compared with similar networks with millions of parameters and complex architecture, their model has lower calculation cost and faster speed. In 2022, Kurmra et al. [27] proposed an improved version of GR-CNN, GR-CNNv2. In the improved version, the network also consists of encoder, residual layer and decoder, but they add a new dropout layer after each regularization output to improve the generalization of the network, and use Mish instead of ReLU as a new activation function in the whole network, which greatly improves the stability of the network.

These methods belong to regression-based capture detection methods, and the process of regression-based capture is shown in Fig. 7. This method no longer only pays attention to the local area, but uses the global information in the image to train the model, directly obtains the capture parameters of position and direction, and generates candidate capture. Because it is a real-time one-time evaluation, it is faster; However, the intermediate process of capturing parameter generation is invisible.

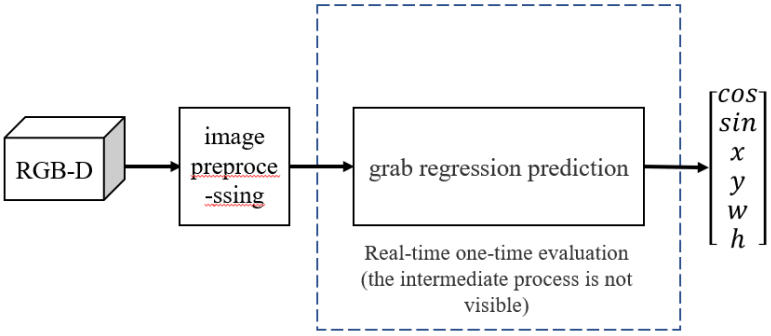

Fig. 7. Flow chart of grab detection based on regression

In 2018, Guo et al. [28] proposed a mixed depth structure that fuses visual and tactile information. This structure is divided into two stages: data collection and capture detection. In the data collection stage, the position information of the target object is extracted and the robot is operated to grab it. Meanwhile, the robot tactile sensing device records tactile information; In the capture detection stage, tactile features and image features are taken as inputs, three proportions of reference rectangles are set, and each position on the feature map is sampled by sliding window. Finally, the capture score and capture parameters of each reference rectangle are output, and the corresponding capture rectangle is regressed. In this method, the reference rectangle is used to enable the network to detect every position in the image, and tactile information is added to enable the network to learn from multiple angles, so as to obtain better capture detection results. Chu et al. [29] transformed the problem of grabbing rectangular regression into a combined problem of region detection and direction classification. Similar to the reference rectangle box introduced by Guo et al. [28], they use horizontal reference rectangle to regress and predict the grab rectangle. However, the reference rectangle in these two methods does not consider the rotation angle and cannot accurately describe the direction of the grab rectangle. Zhou et al. [30] predicted multiple grasping postures of parallel gripper robot by using RGB images input by fully convolution neural network. Different from Guo[28] and Chu's[29] methods, they used a directional anchor frame with multiple directions to indicate the direction of grasping rectangular frame. Removing the direction classification in grasping parameter regression task made the network pay more attention to other parameter features. In 2020, Depierre et al. [31] introduced the fusion of grabbed regression prediction and evaluation quality into the scorer. Scorer evaluates by inputting a set of grabbing parameters in the neighborhood of grabbing position, regression value of directional anchor frame and grabbing score, and generates grabbing and non-grabbing scores. Song et al. [32] proposed a single-stage crawling detection method based on regional suggestion network. Through the anchor frame matching strategy based on the rotation angle and the center position, the distance between the directional anchor frame and the real grab rectangle is minimized and the direction of the anchor frame and the real rectangle frame are matched. Finally, the grab parameters are regressed by the directional anchor frame.

These methods belong to detection-based methods, which refer to some key ideas in the task of target detection. By setting a certain proportion of reference anchor frames to sample the feature map, based on the prior information of these anchor frames, the regression problem of grabbing the directed rectangle is simplified. However, the regression quality of the rectangular frame captured by this method is very dependent on the prior information of the anchor frame, which leads to the heavy anchor frame mechanism and greatly increases the complexity of the network.

2.3 Summary

2D grasping method is suitable for grasping from a single angle in a fixed situation. When training data is generated, the placement of each object in the plane has probability distribution. If it is extended to any angle of view, the data of many angles of view do not exist in the training set, resulting in the single angle of view of the target object. Moreover, for the 2D plane crawling detection task, the network itself can't learn the position suitable for crawling because the crawling angle is limited to one-way crawling on the plane. This method can't grab from any angle, and it has great limitations in the task of grabbing. In addition, at present, there are only two public data sets for plane capture detection: Cornell capture data set and Jacquard data set, and these two data sets are single-target scenes. The data scale in real scenes is small, but the data scale in simulation environments is large. Due to the lack of training data sets, it is difficult for such methods to be fully applied in practice, which restricts the development and application of this field to some extent.

3 6-DOF space capture

6-DOF spatial grasping means that the manipulator can grasp from any angle according to the position of the object in three-dimensional space, as shown in Fig. 8. With the development of depth camera, six-degree-of-freedom grasping based on object grasping posture has gradually become the research hotspot of robot grasping direction. According to whether the captured object is known or not and whether the 3D model is available, it can be classified into partial model-based capture methods and complete model-based capture methods.

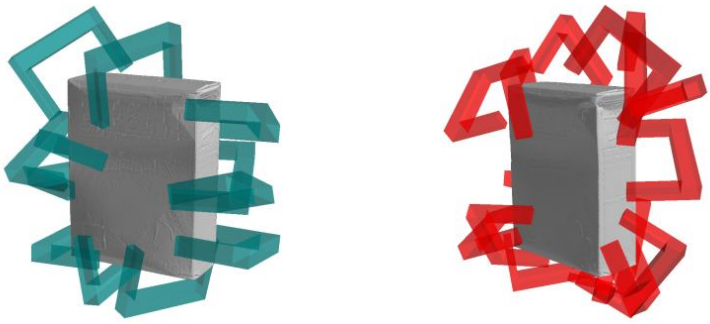

Fig. 8. 6-DOF spatial grab representation

3.1 Grab method based on partial model

Partial model-based capture method usually consists of two independent parts: generating candidate capture and evaluating the quality of candidate capture. In generating candidate grabbing, the geometric information captured by sensors will be used as a heuristic or constraint to build an adaptive grabbing configuration on a given object. The generated grabbing configuration is evaluated by quality metrics, and finally the grabbing configuration with the best performance is selected as the final grabbing gesture.

Zapata-Impata et al. [34] analogizes the way humans grab through the center of mass and perpendicular to the main axis of an object, transforming the problem into finding a plane perpendicular to the approximate main axis of the object and passing through its center of mass, and finding potential grabbing contact points at the opposite edges of the plane. In literature [35], Zapata-Impata et al. improved the ranking metric of point clouds in candidate areas, and used a self-defined function as a metric to calculate the stability of candidate contact points, thus further improving the accuracy of grab detection. However, this method requires extremely high

viewpoint distance and point cloud image quality of depth camera. In 2019, Liang et al. [36] sampled the grabbing candidates based on PointNet network [37] using geometric information as constraints, and the grabbing quality evaluation network selected the candidate grabbing with the highest score. In the same year, Lou et al. [38] proposed a voxel-based 3D depth convolution network to generate an executable six-degree-of-freedom grasping gesture in a cluttered environment. By sampling the whole target object, the candidate grab is obtained and voxelized. Convolution network is used to predict the stability of the candidate grab, and the final grab posture is obtained by evaluating the feasibility of the candidate grab position. Compared with Lou et al.' s method of uniformly sampling the whole image to get the candidate capture, reference [39] also uses voxel-based convolutional neural network, but they get the optimal capture posture by pre-defining capture posture and using convolutional neural network to classify and evaluate 24 pre-defined capture postures, but this method takes longer time. In 2019, Mousavian et al. [40] put forward a model similar to the Generative Confrontation Network (GAN), which provides grabbing suggestions through the variational self-encoder network (VAE) as the generator model, and then the grabbing evaluator evaluates the grabbing quality and gets the final grabbing. The following year, Murali et al. [41] further improved the method in reference [40], and introduced a learning collision detector based on the information of the grabber and the original scene information, which improved the accuracy of the grabbing task in the cluttered scene. These methods are based on classification and follow the principle of "sampling before evaluation": firstly, according to the input three-dimensional data of different forms, the local features of the target object are extracted and sampled to get grabbing suggestions; Then the capture prediction network generates candidate captures according to capture suggestions; Finally, the candidate grabs are classified and evaluated by the quality evaluation network to obtain the final grab.

Qin et al. [42] proposed a single grab suggestion network (S4G) based on Pointnet++[43], and directly generated candidate grabs by regression grab suggestions, instead of using a sliding window-like way. A good collision-free grab is achieved in the chaotic scene. Zhao et al. [44] put forward a single-view point cloud as input in 2020. Their networks include scoring network (SN), grabbing area suggestion network (GRN) and refining network (RN). Firstly, SN network selects the appropriate grab position in the whole scene, and returns to the grab point with high confidence; Then, GRN network builds a capture area centered on the selected high confidence points, and regression capture suggestions; Finally, RN network refines each grabbing suggestion by combining local features to generate accurate grabbing posture. In 2020, Fang[45] et al. proposed a large-scale capture data set with the same evaluation criteria and named it Grassnet-1Billion, aiming at the problems of insufficient training data and lack of evaluation criteria in current research. An end-to-end crawling detection network is proposed for this data set. The encoding and decoding structure is used to extract the global features of the point cloud images, and the second part of the ApproachNet processes the features of the point clouds to group the cylindrical areas to obtain the point clouds in the capture area. Finally, the capture parameters are regressed and the capture stability is evaluated by the capture prediction net and the quality evaluation net. Li Huijun et al. [46] recently proposed a six-degree-of-freedom grabbing detection algorithm based on point cloud features. The algorithm includes three modules: sampling, optimization and evaluation of grasping pose. The sampling module generates a plurality of candidate captures based on the geometric information of surface normals and curvatures of object point clouds; In the optimization module, the force balance method is used to optimize the

candidate grabbing in three directions, so that each sampling point is also the optimal grabbing in the local area. The evaluation module evaluates the samples through CNN model, and selects the capture candidate with the highest score as the final capture. Their method doesn't need to project the point cloud on a 2D plane or convert it into a 3D voxel, which effectively avoids the over-fitting problem of CNN model. These methods are based on regression, taking the whole point cloud image as the input, and combining with the global features, directly regress to capture the parameters of each point cloud, without the time-consuming sampling and searching process of local features.

According to the common characteristics between objects, some researchers classify and compare some models of unknown objects with existing object models, and transfer the grab representation of similar known objects to unknown objects. This method guides the selection of unknown objects by past experience. In 2019, Tian et al. [47] proposed grab transfer based on support vector machine and particle swarm optimization algorithm. Assuming that the new object and the sample object have the same topological structure and similar shape, 3D segmentation is performed on the new object by using geometric and semantic features, the grasping space is calculated for the existing sample object by using active learning, and the original grasping configuration is mapped to the new object by bidirectional contact mapping. Finally, the grasping representation is refined by combining the objective functions defined by contact points, normals, joint angles and grasping quality based on force closure. In 2020, Patten et al. [48] transferred their learning experience from existing object grabbing to unknown object grabbing, and proposed a dense geometric correspondence matching network (DGCM-Net). Firstly, the network uses the dense geometric information correspondence matching between the target object and the database samples, then identifies the sample features closest to the target object through the global geometric coding, and finally uses the local correspondence matching to transfer the relevant capture parameters. When transferring the capture parameters, they reconstruct the correspondence between the depth images by normalizing the coordinate space of the object, and transfer the capture posture from the sample to the target object. Their method marks the grasping posture related to the task, and can give priority to the grasping posture that does not affect the functional use of the object.

3.2 Grab method based on complete model

A complete model-based crawling method usually relies on a pre-built 3D model database, which is marked with multiple sets of feasible crawling and quality indicators provided by auxiliary tools such as Graspit [57]. In the process of execution, it is necessary to associate the sensor input with the objects in the database for capture planning.

Aiming at the problem of grasping detection of some objects, it is often transformed into the problem of 6D attitude estimation. The 6D grasping attitude estimated on a complete 3D model is transformed from object coordinates to camera coordinates by coordinate transformation. Early studies were mainly based on visual and geometric similarity [49]-[52]. With the development of deep learning, 6D object attitude estimation algorithm based on deep learning has been used by many researchers to assist robots in grasping tasks. In 2017, Zeng et al. [53] proposed a data-driven method of self-monitoring in the Amazon Picking Challenge. In this method, the neural network of the whole winder is used to segment and mark multiple views in the scene, and then the scanned 3D object model is fitted to the segmented point cloud, so as to restore the 6-D pose of the object. Billings et al. [54] firstly predicted the middle contour representation of the region of

interest (ROI) in the object and the uncovered part of the object, and then obtained the 6D pose of the object by 3D translation regression and 3D direction regression. Wang et al. [55] used RGB images as input to semantically segment each known object and get the 6D pose of the object. The attitude estimation error of the target object is further corrected by iteration, and the final 6-day attitude of the target object is obtained.

Although the accurate 3D model can help to obtain the grasping posture of the target object. However, when the existing three-dimensional model is different from the target model, the generated six-dimensional pose will have great deviation, which will lead to the failure of the grabbing task. In this case, according to the local area of the target object, some researchers get the complete model by 3D reconstruction, and then register it with the database to get the 6-D pose of the target object. Lundell et al. [56] proposed a robust grabbing method for unknown shape target objects caused by occlusion or lack of scene information. The local view is used as the input of the deep neural network, and the voxel meshes with complete shape are output. The voxel meshes are averaged and candidate capture samples are generated for each group of averaged meshes. Mark et al [57] proposed the network architecture of PointSDF. The network architecture is divided into two parts: 3D reconstruction of target and capture prediction. The 3D reconstruction network receives point cloud information and queries the target points, and returns the shortest distance from each query point to the surface of the reconstructed object. The prediction network uses the capture configuration and the size of the point cloud to predict the capture posture. Aiming at the problem that a single depth camera can't provide the information of the occluded area, some researchers use tactile perception to fill the geometric information of the occluded area of the target to accurately model the object. Varley[58] uses the constantly moving robot hand to contact with the object to obtain tactile information, and combines the local point cloud of the visible part of the target for voxelization, and uses feature fusion as the input of the network to guess the geometric shape of the object. This way of multi-modal perception further promotes the robot to capture the complete spatial model and provides a moresound object model for grasping planning. In 2020, Yang[59] et al. used the 3D geometric information of the object to further improve the candidate grasping posture. They used the RGB-D image with segmentation mask as the input of the grasping suggestion network (GPNet) and the 3D shape reconstruction network (SRNet) respectively, resulting in a six-degree-of-freedom grasping posture and 3D point cloud reconstruction of the object. By projecting the grasping posture to the nearest point in the point cloud, the final grasping output was obtained. This method is more accurate than the single grasping prediction system to obtain the grasping posture.

These methods get the grabbing posture of the target object through the prior knowledge of the known model, and when faced with the problem of 3D data segmentation, they often need complicated calculation process. At the same time, this method requires that the target object has enough correlation with the model to find the corresponding relationship. However, due to the accuracy of sensors and the limitations of existing databases, such methods may perform poorly in novel objects and dense and messy scenes.

3.3 Summary

As for the working scene, 6-DOF spatial capture is not limited to the capture direction and can capture the target from any angle. This advantage makes it more effective than 2D planar capture in dense and cluttered or occluded scenes. In terms of data processing, the advantages of 3D point cloud data over planar 2D images are: (1) it can express the geometric shape information

and spatial position and posture of objects more truly and accurately; (2) it is less affected by the change of illumination intensity, imaging distance and viewpoint; (3) there are no problems such as projection transformation in 2D images; The above advantages of 3D point cloud data make it possible to overcome many shortcomings of 2D planar images in robot recognition and capture. At the same time, no matter which method is used to predict the 6-day grabbing posture, most of the cases need to be segmented. The processing of the three-dimensional data increases the complexity of the network structure and greatly increases the training time of the network, which makes it difficult to meet the real-time requirements. Faced with these problems, with the gradual improvement of 3D data representation, the continuous improvement of computer capabilities and the continuous updating of sensors, the grab detection technology based on this method will become more and more mature.

4. Grab detection dataset

Cornell grab dataset: Cornell grab dataset is the most commonly used plane grab data set, and most of the grab detection papers since 2015 have been evaluated and verified on this data set. Cornell grasp data set captures data through cameras in real scenes, so the target categories and numbers collected in the data set are limited, involving 240 different categories of objects, with a total of 885 RGB images and 885 depth images. Each image corresponds to several grab tags. This data set provides the corresponding point cloud file, which is used to generate the corresponding depth image. Because it is the first large-scale crawling detection data set, this data set is used by many researchers to verify the performance of the algorithm.

Jacquard crawls dataset:Jacquard crawls dataset is a large-scale synthetic data set launched in 2018. This data set is based on a subset of ShapeNet. In the simulation environment, objects in the scene are captured from multiple perspectives and corresponding annotation information is added to each picture. The data set contains a total of 11,619 different categories of objects, including 54,485 RGB images and a corresponding number of depth images, and the number of images is more than 50 times that of Cornell grasp data set. In the real robot grasping experiment, the data set has a better performance than the small sample data set marked by the real scene because of its variety and huge quantity.

YCB-Video dataset:YCB-Video dataset is made based on YCB data set. Twenty-one objects are selected from YCB data set, and 3-9 points are selected from these 21 objects to build a real indoor scene. Then, 92 videos are made by RGB-D camera and saved as picture frames. All the videos in the whole data set contain 133,827 frames, and the six-day pose is marked semi-automatically.

Dex-Net 2.0 dataset: Dex-Net 2.0 dataset is based on 1500 three-dimensional object models selected in Dex-Net 1.0, and the grasping representation of each object model when it is grabbed perpendicular to its surface is obtained on the virtual desktop. These grasping representations can be used as real grasping gestures. In this data set, a virtual depth camera is used to shoot the depth map. For each capture representation, the capture point is the center, and the depth map is rotated until the capture direction is parallel to the horizontal axis of the image. Then, the depth map intercepts 32*32 areas with each grab point as the center as the data set, which contains a total of 6.7 million samples.

GraspNet-1Billion dataset: This dataset was proposed in 2019 and contains 88 everyday objects with high-quality 3D mesh models. These images are collected from 190 chaotic scenes, and each scene takes 512 RGB-D images by kinect camera and realsense camera respectively,

which contains a total of 97,280 images. For each image, the 6D posture and 6-DOF grasping posture of the object are densely marked by force closure. The grabbing gestures of each scene range from 3 million to 9 million, with a total of more than 1.1 billion grabbing gestures. In addition, an online evaluation system is provided, which can uniformly evaluate the current mainstream crawling detection algorithms.

5. Common evaluation indicators

5.1 Plane grab detection:

There are two indicators toevaluate the performance of plane grab detection: point measurement and rectangle measurement.

（1）Point measurement

Point evaluation predicts the distance threshold between the predicted capture center and the actual capture center. If any of these distances is less than a certain threshold, then this capture is regarded as a successful capture.

（2）Rectangular measurement

Rectangle measurement is to evaluate the results according to the rectangle measurement standard on Cornell capture data set. If the predicted capture rectangle G and any regular rectangle label G' can meet the following two conditions at the same time, it is considered that the rectangle represents a reasonable capture position.

$$||G_\theta - G'_\theta| < 30° \quad (1)$$

$$J(G, G') = \frac{|G \cap G'|}{|G \cup G'|} > 25\% \quad (2)$$

The formula (1) indicates that the angle difference between the predicted rectangle and the regular rectangle label is less than 30, and the formula (2) indicates that the Jaccard similarity coefficient between the predicted rectangle and the regular rectangle label is greater than 25%.

5.2 6-DOF spatial grab detection:

(1) Average distance of model points

Attitude mainly includes rotation R and translation T, and the accuracy evaluation standards are mainly ADD (average distance of model points). Given a 3D model M, assuming that the real attitude is R and T, and the predicted attitude is R' and T', the ADD error can be obtained from formula (3).

$$e_{ADD} = avg\|(Rx + T) - (R'x + T')\|, x \in M \quad (3)$$

(2)Chamfer Distance

Chamfer distance is abbreviated as CD, which is used to calculate the average shortest point distance between the generated point cloud and the real point cloud of the object. Used to compare the similarity between the generated point cloud and the real point cloud of the object. The number of point clouds between the point clouds generated in CD and the real point clouds of objects is not necessarily the same. It is assumed that S_1 and S_2represent two groups of three-dimensional point clouds respectively, and the calculation method is as shown in formula (4).

$$CD(S_1, S_2) = \sum_{x \in S_1} \min_{y \in S_2} \| x - y \|_2^2 + \sum_{x \in S_2} \min_{y \in S_1} \| y - x \|_2^2 \quad (4)$$

(3)Earth Mover's Distance

EMD distance is initially used to measure the similarity between two images. In point cloud analysis, EMD distance represents the similarity of two points cloud images. Assuming that P and Q are twopoints sets, and the number of points contained in the two points sets is equal, it is noted

as N. The minimum distance of EMD is obtained by formula (5).

$$minEMD\ (P,Q) = min\sum_{i=1}^{N}\ \sum_{j=1}^{N}\ d_{ij} \quad (5)$$

In formula (5), d_{ij}represents the distance from p_i to q_i.

6. Conclusion

In this paper, the grasping actuator of parallel two-fingered gripper is taken as the research object, and the robot grasping detection technology from 2D plane grasping to 6-DOF space grasping is introduced in detail, and the current mainstream grasping data sets and commonly used evaluation indexes are summarized. Through a large number of investigations and studies on the current research situation, it is not difficult to find that most of the literatures at this stage have the following problems:

(1) The effect of grabbing detection is not good in the case of multi-target and complex scene. It is suggested that researchers gradually change the detection of single object without occlusion into multi-target grabbing detection in complex scene.

(2) Aiming at the tiny parts of the target object that can be used for grabbing, it is often difficult to get a good grabbing posture because of inaccurate positioning or large grabbing frame, so the detection accuracy of small targets should be emphasized.

(3) In some tasks that need real-time grasping, the network takes too long to detect the target, which can't meet the real-time requirement of robot grasping.

(4) At present, most of the researches focus on the grabbing detection of static objects, but few algorithms focus on the grabbing detection of dynamic objects. The task of dynamic target recognition and positioning is more complicated, and because the position of the moving object changes all the time, it is also difficult to predict the grasping attitude.

(5) At present, compared with the objects in daily life, there are fewer kinds and fewer numbers of data sets used for training networks. However, a network with good generalization and enough stability needs huge data sets to support it, so it is also an important work to expand the existing data sets and make a more perfect data set for crawling detection.

For these two methods, whether researchers use 2D plane capture detection or 6-DOF space capture detection, this paper provides a detailed introduction of the two methods, which can provide some help for researchers, promote the development of robot capture technology, speed up the robot's working ability in complex environment and improve people's quality of life.

References

[1]Kumra S, Kanan C. Robotic Grasp Detection using Deep Convolutional Neural Networks[C]. International Conference on Intelligent Robots & Systems. IEEE, 2016.

[2]Sahbani A, El-Khoury S, Bidaud P. An overview of 3D object grasp synthesis algorithms[J]. Robotics and Autonomous Systems. 2012:326-336.

[3]Bohg J, Morales A, Asfour T, et al. Data-driven grasp synthesis— A survey[J]. IEEE Transactions on Robotics, 2014, 30(2): 289-309.

[4]Caldera S, Rassau A, Chai D. Review of deep learning methods in robotic grasp detection[J]. Multimodal Technologies and Interaction, 2018, 2(3): 57.

[5]Piater J H . Learning Visual Features to Predict Hand Orientations[J]. icml workshop on machine learning of spatial knowledge, 2000.

[6]Domae Y, Okuda H, Taguchi Y, et al. Fast graspability evaluation on single depth maps for bin picking with general grippers[C]. International Conference on Robotics & Automation. IEEE, 2014.

[7]Mahler J, Pokorny F T, Hou B, et al.Dex-Net 1.0: A cloud-based network of 3D objects for robust grasp planning using a Multi-Armed Bandit model with correlated rewards[C] . International Conference on Robotics and Automation (ICRA). IEEE, 2016.
[8]Mahler J, Liang J, Niyaz S, et al. Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics[J]. 2017.
[9]Do T T, Nguyen A, Reid I. AffordanceNet: An End-to-End Deep Learning Approach for Object Affordance Detection[C]. International Conference on Robotics and Automation (ICRA). IEEE, 2018.
[10]Chu F J, Xu R, Vela P A. Detecting Robotic Affordances on Novel Objects with Regional Attention and Attributes[J]. 2019.
[11]Cao H, Chen G, Li Z, et al. Lightweight Convolutional Neural Network with Gaussian-based Grasping Representation for Robotic Grasping Detection[J]. 2021.
[12]Xu R, Chu F-J, Vela PA. GKNet: Grasp keypoint network for grasp candidates detection. The International Journal of Robotics Research[J]. 2022; 41(4):361-389.
[13]Zeng A, Song S, Yu K T, et al. Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching[J]. The International Journal of Robotics Research, 2017.
[14]Cai J, Cheng H, Zhang Z, et al. MetaGrasp: Data Efficient Grasping by Affordance Interpreter Network[C]. International Conference on Robotics & Automation. 2019.
[15]Morrison D, Corke P, Leitner J. Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach[J]. 2018.
[16]Morrison D, Corke P, Leitner J. Multi-View Picking: Next-best-view Reaching for Improved Grasping in Clutter[C] . International Conference on Robotics and Automation. Institute of Electrical and Electronics Engineers Inc, 2019.
[17]Morrison D, Corke P, Leitner J. Learning robust, real-time, reactive robotic grasping[J]. The International journal of robotics research, 2020, 39(2/3):183-201.
[18]Jiang Y, Moseson S, Saxena A. Efficient grasping from RGBD images: Learning using a new rectangle representation[C]. IEEE, 2011.
[19]Ian Lenz, Honglak Lee, Ashutosh Saxena. Deep learning for detecting robotic grasps[J]. The International Journal of Robotics Research,2015,34(4-5).
[20]Wang Z, Li Z, Wang B, and H. Liu. Robot grasp detection using multimodal deep convolutional neural networks[J]. 8,9(2016-9-01), 2016, 8(9).
[21]Asif, Bennamoun, Sohel, et al. RGB-D Object Recognition and Grasp Detection Using Hierarchical Cascaded Forests[J]. 2017,33(3): 547-564.
[22]Park D, Chun S Y. Classification based Grasp Detection using Spatial Transformer Network[J]. 2018.
[23]Redmon J, Angelova A . [IEEE 2015 IEEE International Conference on Robotics and Automation (ICRA) - Real-time grasp detection using convolutional neural networks[J]. 2015:1316-1322.
[24]Kumra S, Kanan C. Robotic Grasp Detection using Deep Convolutional Neural Networks[C]. International Conference on Intelligent Robots & Systems. IEEE, 2016.
[25]Qiang Z, Qu D, Fang X, et al. Robust Robot Grasp Detection in Multimodal Fusion[C].2017.
[26]Kumra S, Joshi S, Sahin F. Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network[J]. 2019.

[27]Kumra S, Joshi S. Sahin, F. GR-ConvNet v2: A Real-Time Multi-Grasp Detection Network for Robotic Grasping[J]. 2022.
[28]Guo D, Sun F, Liu H, et al. A hybrid deep architecture for robotic grasp detection[C]. International Conference on Robotics and Automation (ICRA). IEEE, 2017.
[29]Chu F J, Vela P A. Deep Grasp: Detection and Localization of Grasps with Deep Neural Networks[J]. 2018.
[30]Zhou X, Lan X, Zhang H, et al. Fully Convolutional Grasp Detection Network with Oriented Anchor Box[C]. International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.
[31]Depierre A, E Dellandréa, Chen L. Optimizing Correlated Graspability Score and Grasp Regression for Better Grasp Prediction[J]. 2020.
[32]Song Y, Gao L, Li X, et al. A novel robotic grasp detection method based on region proposal networks[J]. Robotics and Computer-Integrated Manufacturing, 2020, 65:101963.
[33]Zapata-Impata B S, Mateo C M, Gil P, et al. Using Geometry to Detect Grasping Points on 3D Unknown Point Cloud[C].14th International Conference on Informatics in Control, Automation and Robotics. 2017.
[34]Zapata-Impata B S, Gil P, Pomares J, et al. Fast Geometry-based Computation of Grasping Points on Three-dimensional Point Clouds[J]. International Journal of Advanced Robotic Systems, 2019, 16(1).
[35]Gualtieri M, Pas A T, Saenko K, et al. High precision grasp pose detection in dense clutter[J]. IEEE, 2016.
[36]Liang H, Ma X, Li S, et al. PointNetGPD: Detecting Grasp Configurations from Point Sets[C]. International Conference on Robotics and Automation. IEEE, 2019.
[37]Qi C R, Su H, Mo K, et al. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation[C]. Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[38]Lou X, Yang Y, Choi C. Learning to Generate 6-DoF Grasp Poses with Reachability Awareness[J]. 2019.
[39]Choi C, Schwarting W, Delpreto J, et al. Learning Object Grasping for Soft Robot Hands[J]. IEEE Robotics & Automation Letters, 2018:1-1.
[40]Mousavian A, Eppner C, Fox D. 6-DOF GraspNet: Variational Grasp Generation for Object Manipulation[C]. International Conference on Computer Vision (ICCV). IEEE, 2019.
[41]Murali A, Mousavian A, Eppner C, et al. 6-DOF Grasping for Target-driven Object Manipulation in Clutter[C]. International Conference on Robotics and Automation. IEEE, 2020.
[42]Qin Y, Chen R, Zhu H, et al. S4G: Amodal Single-view Single-Shot SE(3) Grasp Detection in Cluttered Scenes[J]. 2019.
[43]Qi C R, Li Y, Hao S, et al. PointNet++: Deep Hierarchical Feature Learning on Point Sets in a Metric Space[C]. 2017.
[44]Zhao B, Zhang H, Lan X, et al. REGNet: REgion-based Grasp Network for Single-shot Grasp Detection in Point Clouds[J]. 2020.
[45]Fang H S, Wang C X, Gou M H, et al. Graspnet-1billion: A large-scale benchmark for general object grasping[J]. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,2020.
[46]Li Huijun, Qu Xiaochang, Ye Bin. Six-degree-of-freedom robot grasping based on three-dimensional point cloud characteristics of unknown objects [J]. Control Theory and Application, 2022, 39(6):9.

[47]Tian H, Wang C, Manocha D, et al. Transferring Grasp Configurations using Active Learning and Local Replanning[C]. International Conference on Robotics and Automation (ICRA). 2019.
[48]Patten T, Park K, Vincze M. DGCM-Net: Dense Geometrical Correspondence Matching Network for Incremental Experience-based Robotic Grasping[J]. 2020.
[49]Miller A T, Allen P K. GraspIt! : A versatile simulator for robotic grasping[J]. IEEE Robotics & Automation Magazine, 2005, 11(4):110-122.
[50]Bohg J, Morales A, Asfour T, et al. Data-Driven Grasp Synthesis – A Survey[J]. IEEE Transactions on Robotics, 2014, 30(2):289-309.
[51]Brook P, Ciocarlie M, Hsiao K. Collaborative Grasp Planning with Multiple Object Representations[C]. 2011 IEEE International Conference on Robotics and Automation. IEEE, 2011.
[52]Hinterstoisser S, Holzer S, et al. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. 2011.
[53]Zeng A, Yu K T, Song S, et al. Multi-view self-supervised deep learning for 6D pose estimation in the Amazon Picking Challenge[J]. International Conference on Robotics and Automation (ICRA), 2017.
[54]Billings G, Johnson-Roberson M. SilhoNet: An RGB Method for 6D Object Pose Estimation[J]. 2018.
[55]Wang C, Xu D, Zhu Y, et al. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion[J]. 2019.
[56]Lundell J, Verdoja F, Kyrki V. Robust Grasp Planning Over Uncertain Shape Completions[C]. International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019.
[57]Mark V, Lu Q, Sundaralingam B, et al. Learning Continuous 3D Reconstructions for Geometrically Aware Grasping[J]. 2019.
[58]Watkins-Valls D, Varley J, Allen P. Multi-Modal Geometric Learning for Grasping and Manipulation[J]. 2019.
[59]Yang D, Tosun T, Eisner B, et al. Robotic Grasping through Combined Image-Based Grasp Proposal and 3D Reconstruction[J]. 2020.