Artificial Intelligence Review
https://doi.org/10.1007/s10462-020-09888-5

Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review

Guoguang Du[1] · Kai Wang[1] · Shiguo Lian[1] · Kaiyong Zhao[1]

Abstract

This paper presents a comprehensive survey on vision-based robotic grasping. We conclude three key tasks during vision-based robotic grasping, which are object localization, object pose estimation and grasp estimation. In detail, the object localization task contains object localization without classification, object detection and object instance segmentation. This task provides the regions of the target object in the input data. The object pose estimation task mainly refers to estimating the 6D object pose and includes correspondence-based methods, template-based methods and voting-based methods, which affords the generation of grasp poses for known objects. The grasp estimation task includes 2D planar grasp methods and 6DoF grasp methods, where the former is constrained to grasp from one direction. These three tasks could accomplish the robotic grasping with different combinations. Lots of object pose estimation methods need not object localization, and they conduct object localization and object pose estimation jointly. Lots of grasp estimation methods need not object localization and object pose estimation, and they conduct grasp estimation in an end-to-end manner. Both traditional methods and latest deep learning-based methods based on the RGB-D image inputs are reviewed elaborately in this survey. Related datasets and comparisons between state-of-the-art methods are summarized as well. In addition, challenges about vision-based robotic grasping and future directions in addressing these challenges are also pointed out.

Keywords Robotic grasping · Object localization · Object pose estimation · Grasp estimation

1 Introduction

An intelligent robot is expected to perceive the environment and interact with it. Among the essential abilities, the ability to grasp is fundamental and significant in that it will bring enormous power to the society Sanchez et al. (2018). For example, industrial

✉ Guoguang Du
george.du@cloudminds.com

[1] CloudMinds Technologies Inc., Beijing, China

Published online: 17 August 2020

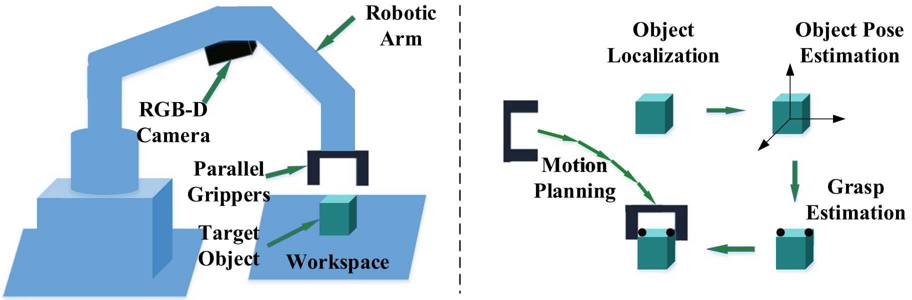

Fig. 1 The grasp detection system. (Left) The robotic arm, equipped with one RGB-D camera and one parallel gripper, is to grasp the target object placed on a planar work surface. (Right) The grasp detection system involves target object localization, object pose estimation, and grasp estimation

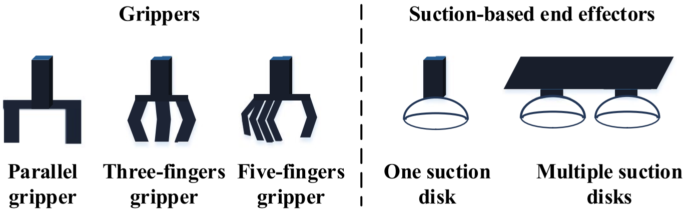

Fig. 2 Different kinds of end effectors. (Left)Grippers. (Right)Suction-based end effectors. In this paper, we mainly consider parallel grippers

robots can accomplish the pick-and-place task which is laborious for human labors, and domestic robots are able to provide assistance to disabled or elder people in their daily grasping tasks. Endowing robots with the ability to perceive has been a long-standing goal in computer vision and robotics discipline.

As much as being highly significant, robotic grasping has long been researched. The robotic grasping system Kumra and Kanan (2017) is considered as being composed of the following sub-systems: the grasp detection system, the grasp planning system and the control system. Among them, the grasp detection system is the key entry point, as illustrated in Fig. 1. The grasp planning system and the control system are more relevant to the motion and automation discipline, and in this survey, we only concentrate on the grasp detection system.

The robotic arm and the end effectors are essential components of the grasp detection system. Various 5-7 DoF robotic arms are produced to ensure enough flexibilities and they are equipped on the base or a human-like robot. Different kinds of end effectors, such as grippers and suction disks, can achieve the object picking task, as shown in Fig. 2. The majority of methods paid attentions on parallel grippers (Mahler et al. 2017), which is a relatively simple situation. With the struggle of academia, dexterous grippers (Liu et al. 2019; Fan and Tomizuka 2019; Akkaya et al. 2019) are researched to accomplish complex grasp tasks. In this paper, we only talk about grippers, since suction-based end effectors are relatively simple and limited in grasping complex objects. In addition, we concentrate on methods using parallel grippers, since this is the most widely researched.

The essential information to grasp the target object is the 6D gripper pose in the camera coordinate, which contains the 3D gripper position and the 3D gripper orientation to execute the grasp. The estimation of 6D gripper poses varies aiming at different grasp manners, which can be divided into the 2D planar grasp and the 6DoF grasp.

2D planar grasp means that the target object lies on a planar workspace and the grasp is constrained from one direction. In this case, the height of the gripper is fixed and the gripper direction is perpendicular to one plane. Therefore, the essential information is simplified from 6D to 3D, which are the 2D in-plane position and 1D rotation angle. In earlier years when the depth information is not easily captured, the 2D planar grasp is mostly researched. The mostly used scenario is to grasp machine components in the factory. The grasping contact points are evaluated whether they can afford the force closure Chen and Burdick (1993). With the development of deep learning, large number of methods treated oriented rectangles as the grasp configuration, which could be beneficial from the mature 2D detection frameworks. Since then, the capabilities of 2D planar grasp are enlarged extremely and the target objects to be grasped are extended from known objects to novel objects. Large amounts of methods by evaluating the oriented rectangles (Jiang et al. 2011; Lenz et al. 2015; Pinto and Gupta 2016; Mahler et al. 2017; Park and Chun 2018; Redmon and Angelova 2015; Zhang et al. 2017; Kumra and Kanan 2017; Chu et al. 2018; Park et al. 2018; Zhou et al. 2018) are proposed. Besides, some deep learning-based methods of evaluating grasp contact points (Zeng et al. 2018; Cai et al. 2019; Morrison et al. 2018) are also proposed in recent years.

6DoF grasp means that the gripper can grasp the object from various angles in the 3D space, and the essential 6D gripper pose could not be simplified. In early years, analytical methods were utilized to analyze the geometric structure of the 3D data, and the points suitable to grasp were found according to force closure. Sahbani et al. (2012) presented an overview of 3D object grasping algorithms, where most of them deal with complete shapes. With the development of sensor devices, such as Microsoft Kinect, Intel RealSense, etc, researchers can obtain the depth information of the target objects easily and modern grasp systems are equipped with RGB-D sensors, as shown in Fig. 3. The depth image can be easily lifted into 3D point cloud with the camera intrinsic parameters and the depth image-based 6DoF grasp becomes the hot research areas. Among 6DoF grasp methods, most of them aim at known objects where the grasps could be precomputed, and the problem is thus transformed into a 6D object pose estimation problem (Wang et al. 2019; Zhu et al. 2020; Yu et al. 2020; He et al. 2020). With the development of deep learning, lots of methods (ten Pas et al. 2017; Liang et al. 2019; Mousavian et al. 2019; Qin et al. 2020; Zhao and Nanning 2020) illustrated powerful capabilities in dealing with novel objects.

Both 2D planar grasp and 6DoF grasp contain common tasks which are object localization, object pose estimation and grasp estimation.

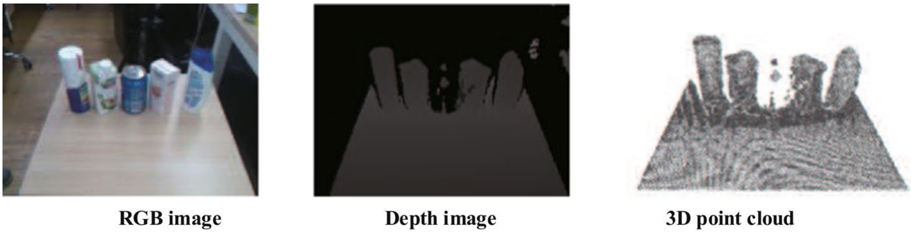

Fig. 3 A RGB-D image. The depth image is transformed into 3D point cloud

In order to compute the 6D gripper pose, the first thing to do is to locate the target object. Aiming at object localization, there exist three different situations, which are object localization without classification, object detection and object instance segmentation. Object localization without classification means obtaining the regions of the target object without classifying its category. There exist cases that the target object could be grasped without knowing its category. Object detection means detecting the regions of the target object and classifying its category. This affords the grasping of specific objects among multiple candidate objects. Object instance segmentation refers to detecting the pixel-level or point-level instance objects of a certain class. This provides delicate information for pose estimation and grasp estimation. Early methods assume that the object to grasp is placed in a clean environment with simple background and thus simplifies the object localization task, while in relatively complex environments their capabilities are quite limited. Traditional object detection methods utilized machine learning methods to train classifiers based on hand-crafted 2D descriptors. However, these classifiers show limited performance since the limitations of hand-crafted descriptors. With the deep learning, the 2D detection and 2D instance segmentation capabilities improves a lot, which affords object detection in more complex environments.

Most of the current robotic grasping methods aim at known objects, and estimating the object pose is the most accurate and simplest way to a successful grasp. There exist various methods in computing the 6D object poses, which varies from 2D inputs to 3D inputs, from traditional methods to deep learning methods, from textured objects to textureless or occluded objects. In this paper, we categorize these methods into correspondence-based methods, template-based methods and voting-based methods, where only feature points, the whole input and each meta unit are involved in computing the 6D object pose. Early methods tackled this problem in 3D domain by conducting partial registration. With the development of deep learning, methods using RGB image only can provide relatively high accurate 6D object poses, which highly improves the grasp capabilities.

Grasp estimation is conducted when we have the localized target object. Aiming at 2D planar grasp, the methods are divided into methods of evaluating the grasp contact points and methods of evaluating the oriented rectangles. Aiming at 6DoF grasp, the methods are categorized into methods based on the partial point cloud and methods based on the complete shape. Methods based on the partial point cloud mean that we do not have the identical 3D model of the target object. In this case, two kinds of methods exist which are methods of estimating grasp qualities of candidate grasps and methods of transferring grasps from existing ones. Methods based on complete shape means that the grasp estimation is conducted on a complete shape. When the target object is known, the 6D object pose could be computed. When the target shape is unknown, it can be reconstructed from single-view point clouds, and grasp estimation could be conducted on the reconstructed complete 3D shape. With the joint development of the above aspects, the kinds of objects that could be grasped, the robustness of the grasp and the affordable complexity of the grasp scenario all have improved a lot, which affords many more applications in industrial as well as domestic applications.

Aiming at these tasks mentioned above, there have been some works (Sahbani et al. 2012; Bohg et al. 2014; Caldera et al. 2018) concentrating on one or a few tasks, while there is still lack of a comprehensive introduction on these tasks. These tasks are reviewed elaborately in this paper, and a taxonomy of these tasks is shown in Fig. 4. To the best of our knowledge, this is the first review that broadly summarizes the progress and promises new directions in vision-based robotic grasping. We believe that this contribution will serve as an insightful reference to the robotic community.

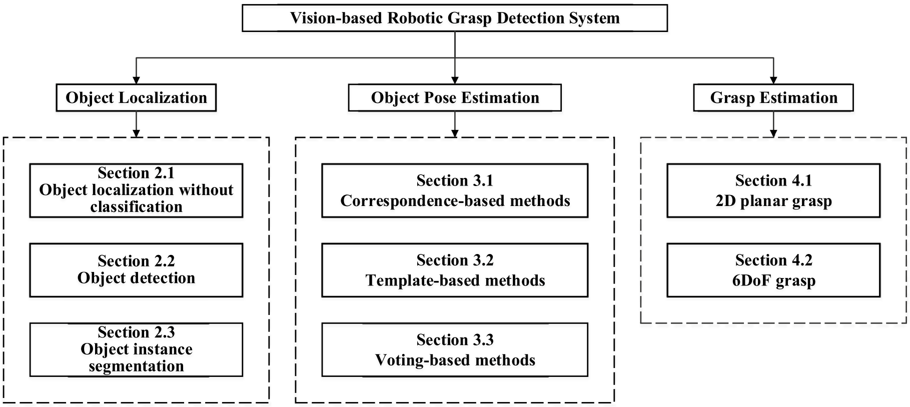

Fig. 4 A taxonomy of tasks in vision-based robotic grasp detection system

The remainder of the paper is arranged as follows. Section 2 reviews the methods for object localization. Section 3 reviews the methods for 6D object pose estimation. Section 4 reviews the methods for grasp estimation. The related datasets, evaluation metrics and comparisons are also reviewed in each section. Finally, challenges and future directions are summarized in Sect. 5.

2 Object localization

Most of the robotic grasping approaches require the target object's location in the input data first. This involves three different situations: object localization without classification, object detection and object instance segmentation. Object localization without classification only outputs the potential regions of the target objects without knowing their categories. Object detection provides bounding boxes of the target objects as well as their categories. Object instance segmentation further provides the pixel-level or point-level regions of the target objects along with their categories.

2.1 Object localization without classification

In this situation, the task is to find potential locations of the target object without knowing the category of the target object. There exist two cases: if you known the concrete shapes of the target object, you can fit primitives to obtain the locations. If you can not ensure the shapes of the target object, salient object detection(SOD) could be conducted to find the salient regions of the target object. Based on 2D or 3D inputs, the methods are summarized in Table 1.

2.1.1 2D localization without classification

This kind of methods deal with 2D image inputs, which are usually RGB images. According to whether the object's contour shape is known or not, methods can be divided into methods of fitting shape primitives and methods of salient object detection. Typical functional flow-chart of 2D object localization without classification is illustrated in Fig. 5.

Table 1 Methods of object localization without classification

Methods	Fitting shape primitives	Salient object detection
2D localization	Fitting ellipse Fitzgibbon and Fisher (1996), Fitting polygons Douglas and Peucker (1973)	Jiang et al. (2013), Zhu et al. (2014), Peng et al. (2016), Cheng et al. (2014), Wei et al. (2012), Shi et al. (2015), Yang et al. (2013), Wang et al. (2016), Guo et al. (2017), Zhao et al. (2015), Zhang et al. (2016), DHSNet Liu and Han (2016), Hou et al. (2017), PICANet Liu et al. (2018), Liu et al. (2019), Qi et al. (2019)
3D localization	Rabbani and Van Den Heuvel (2005), Rusu et al. (2009), Goron et al. (2012), Jiang and Xiao (2013), Khan et al. (2015), Zapata-Impata et al. (2019)	Peng et al. (2014), Ren et al. (2015), Qu et al. (2017), Han et al. (2018), Chen et al. (2019); Chen and Li (2019), Chen and Li (2018), Piao et al. (2019), Kim et al. (2008), Bhatia and Chalup (2013), Pang et al. (2020)

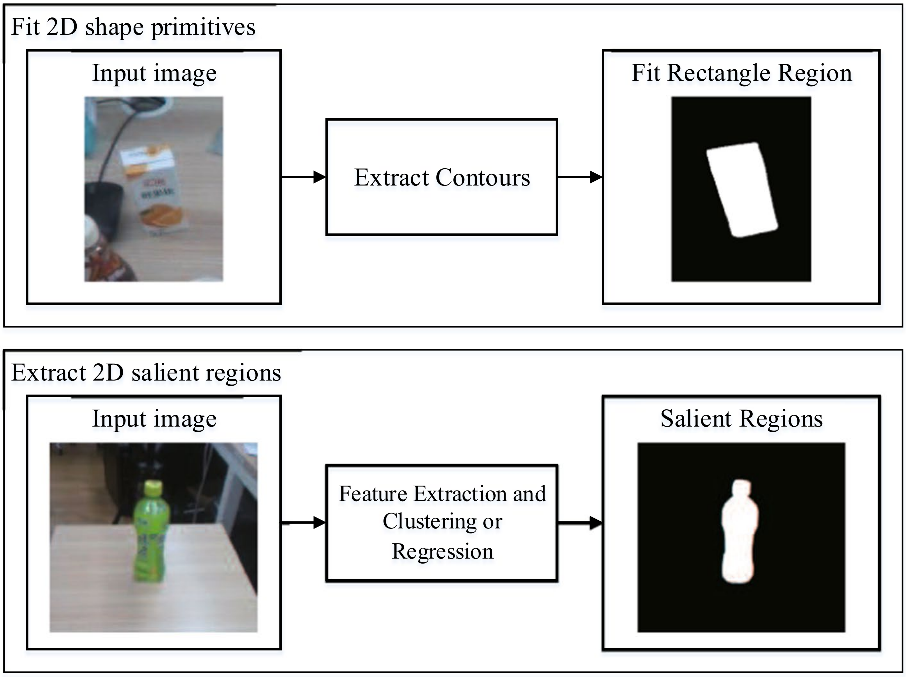

Fig. 5 Typical functional flow-chart of 2D object localization without classification

Fitting 2D shape primitives The shape of the target object could be an eclipse, a polygon or a rectangle, and these shapes could be regarded as shape primitives. Through fitting methods, the target object could be located. General procedures of this kind of methods usually contain enclosed contour extraction and primitive fitting. There exist many algorithms integrated in OpenCV Bradski and Kaehler (2008) for primitives fitting, such as fitting ellipse Fitzgibbon and Fisher (1996) and fitting polygons Douglas and Peucker (1973). This kind of methods are usually used in 2D planar robotic grasping tasks, where the object are viewed from a fixed angle, and the target object are constrained with some known shapes.

2D salient object detection Compared with shape primitives, salient object regions could be represented in arbitrary shapes. 2D salient object detection(SOD) aims to locate and segment the most visually distinctive object regions in a given image, which is more like a segmentation task without object classification. Non-deep learning SOD methods exploit low-level feature representations (Jiang et al. 2013; Zhu et al. 2014; Peng et al. 2016) or rely on certain heuristics such as color contrast Cheng et al. (2014), background prior Wei et al. (2012). Some other methods conduct an over-segmentation process that generates regions Shi et al. (2015), super-pixels (Yang et al. 2013; Wang et al. 2016), or object proposals Guo et al. (2017) to assist the above methods.

Deep learning-based SOD methods have shown superior performance over traditional solutions since 2015. Generally, they can be divided into three main categories, which are Multi-Layer Perceptron (MLP)-based methods, Fully Convolutional Network (FCN)-based methods and Capsule-based methods. MLP-based methods typically extract deep features for each processing unit of an image to train an MLP-classifier for saliency score prediction. Zhao et al. (2015) proposed a unified multi-context deep learning framework which involves global context and local context, which are fed into an MLP for foreground/background classification to model saliency of objects in images. Zhang et al. (2016) proposed a salient object detection system which outputs compact detection windows for unconstrained images, and a maximum a posteriori (MAP)-based subset optimization formulation for filtering bounding box proposals. The MLP-based SOD methods cannot capture well critical spatial information and are time-consuming. Inspired by Fully Convolutional Network (FCN) Long et al. (2015), lots of methods directly output whole saliency maps. Liu and Han (2016) proposed an end-to-end saliency detection model called DHSNet, which can simultaneously refine the coarse saliency map. Hou et al. (2017) introduced short connections to the skip-layer structures, which provides rich multi-scale feature maps at each layer. Liu et al. (2018) proposed a pixel-wise contextual attention network called PiCANet, which generates an attention map for each pixel and each attention weight corresponds to the contextual relevance at each context location. With the raise of Capsule Network (Hinton et al. 2011; Sabour et al. 2017, 2018), some capsule-based methods are proposed. Liu et al. (2019) incorporated the part-object relationships in salient object detection, which is implemented by the Capsule Network. Qi et al. (2019) proposed CapSalNet, which includes a multi-scale capsule attention module and multi-crossed layer connections for salient object detection. Readers could refer to some surveys (Borji et al. 2019; Wang et al. 2019) for comprehensive understandings of 2D salient object detection.

Discussions The 2D object localization without classification are widely used in robotic grasping tasks but in a junior level. During industrial scenarios, the mechanical components are usually with fixed shapes, and many of them could be localized through fitting shape primitives. In some other grasping scenarios, the background priors or color contract is utilized to obtain the salient object for grasping. In Dexnet 2.0 Mahler et al. (2017), the target objects are laid on a workspace with green color, and they are easily segmented using color background subtraction.

2.1.2 3D localization without classification

This kind of methods deal with 3D point cloud inputs, which are usually partial point clouds reconstructed from single-view depth images in robotic grasping tasks. According to whether the object's 3D shape is known or not, methods can also be divided into methods of fitting 3D shape primitives and methods of salient 3D object detection. Typical functional flow-chart of 3D object localization without classification is illustrated in Fig. 6.

Fitting 3D shape primitives The shape of the target object could be a sphere, a cylinder or a box, and these shapes could be regarded as 3D shape primitives. There exist lots of methods aiming at fitting 3D shape primitives, such as RANdom SAmple Consensus (RANSAC) Fischler and Bolles (1981)-based methods, Hough-like voting methods Rabbani and Van Den Heuvel (2005) and other clustering techniques (Rusu et al. 2009; Goron et al. 2012). These methods deal with different kinds of inputs and have been applied in areas like modeling, rendering and animation. Aiming at object localization and robotic grasping tasks, the input data is a partial point cloud, where the object is incomplete, and the ambition is to find the points that can constitute one of the 3D shape primitives. Some methods (Jiang and Xiao 2013; Khan et al. 2015) detect planes at object boundaries and assemble them. Jiang and Xiao (2013) and Khan et al. (2015) explored the 3D structures in an indoor scene and estimated their geometry using cuboids. Rabbani and Van Den Heuvel (2005) presented an efficient Hough transform for automatic detection of cylinders in

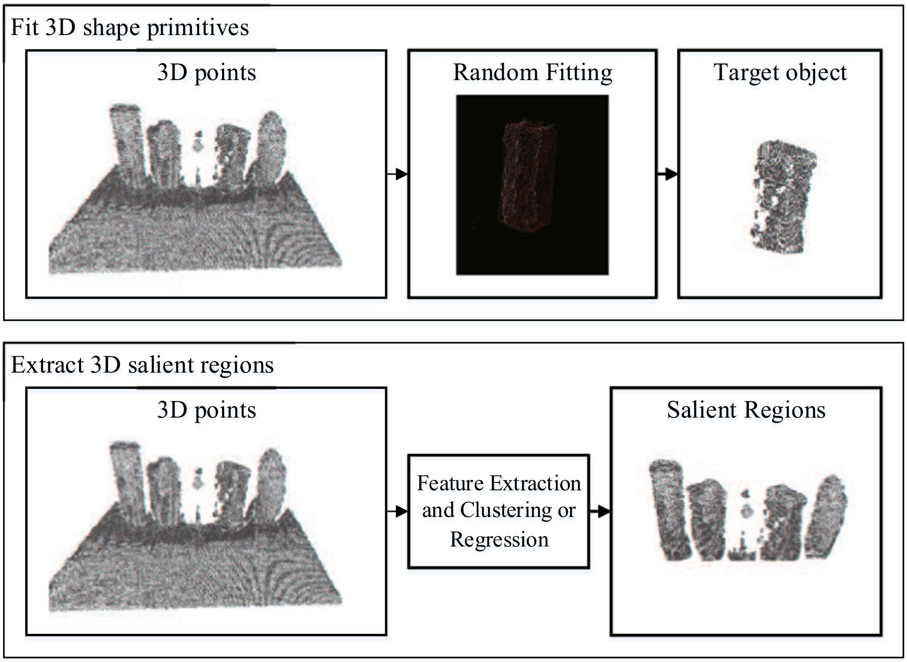

Fig. 6 Typical functional flow-chart of 3D object localization without classification

point clouds. Some methods (Rusu et al. 2009; Goron et al. 2012) conduct primitive fitting after segmenting the scene. Rusu et al. (2009) used a combination of robust shape primitive models with triangular meshes to create a hybrid shape-surface representation optimal for robotic grasping. Goron et al. (2012) presented a method to locate the best parameters for cylindrical and box-like objects in a cluttered scene. They increased the robustness of RANSAC fits when dealing with clutter through employing a set of inlier filters and the use of Hough voting. They provided robust results and models that are relevant for grasp estimation. Readers could refer to the survey Kaiser et al. (2019) for more details.

3D salient object detection Compared with 2D salient object detection, 3D salient object detection consumes many kinds of 3D data, such as depth image and point cloud. Although above 2D salient object detection methods have achieved superior performance, they still remain challenging in some complex scenarios, where depth information could provide much assistance. RGB-D saliency detection methods usually utilize hand-crafted or deep learning-based features from RGB-D images and fuse them in different ways. Peng et al. (2014) proposed a simple fusion strategy which extends RGB-based saliency models by incorporating depth-induced saliency. Ren et al. (2015) exploited the normalized depth prior and the global-context surface orientation prior for salient object detection. Qu et al. (2017) trained a CNN-based model which fuses different low level saliency cues into hierarchical features for detecting salient objects in RGB-D images. Chen et al. (Chen et al. 2019; Chen and Li 2019) utilized two-stream CNNs-based models with different fusion structures. Chen and Li (2018) further proposed a progressively complementarity-aware fusion network for RGB-D salient object detection, which is more effective than early-fusion methods Hou et al. (2017) and late-fusion methods Han et al. (2018). Piao et al. (2019) proposed a depth-induced multi-scale recurrent attention network (DMRANet) for saliency detection, which achieves dramatic performance especially in complex scenarios. Pang et al. (2020) proposed a hierarchical dynamic filtering network (HDFNet) and a hybrid enhanced loss. Li et al. (2020) proposed a Cross-Modal Weighting (CMW) strategy to encourage comprehensive interactions between RGB and depth channels. These methods demonstrate remarkable performance of RGB-D SOD.

Aiming at 3D point cloud input, lots of methods are proposed to detect saliency maps of a complete object model Zheng et al. (2019), whereas, our ambitious is to locate the salient object from the 3D scene inputs. Kim et al. (2008) described a segmentation method for extracting salient regions in outdoor scenes using both 3D point clouds and RGB image. Bhatia and Chalup (2013) proposed a top-down approach for extracting salient objects/regions in 3d point clouds of indoor scenes.They first segregates significant planar regions, and extracts isolated objects present in the residual point cloud. Each object is then ranked for saliency based on higher curvature complexity of the silhouette.

Discussions 3D object localization is widely used in robotic grasping tasks but also in a junior level. In Rusu et al. (2009) and Goron et al. (2012), fitting 3D shape primitives has been successfully applied into robotic grasping tasks. In Zapata-Impata et al. (2019), the background is first filtered out using the height constraint, and the table is filtered out by fitting a plane using RANSAC Fischler and Bolles (1981). The remained point cloud is clustered and K object's clouds are achieved finally. There also exist some other ways to remove the background points through fitting background points using existing full 3D point cloud. These methods are successfully applied into robotic grasping tasks.

2.2 Object detection

The task of object detection is to detect instances of objects of a certain class, which can be treated as a localization task plus a classification task. Usually, the shapes of the target objects are unknown, and accurate salient regions are hardly achieved. Therefore, the regularly bounding boxes are used for general object localization and classification tasks, and the outputs of object detection are bounding boxes with class labels. Based on whether using region proposals or not, the methods can be divided into two-stage methods and one-stage methods. These methods are summarized respectively in Table 2 aiming at 2D or 3D inputs.

2.2.1 2D object detection

2D object detection means detecting the target objects in 2D images by computing their 2D bounding boxes and categories. The most popular way of 2D detection is to generate object proposals and conduct classification, which is the two-stage methods. With the development of deep learning networks, especially Convolutional Neural Network (CNN), two-stage methods are improved extremely. In addition, large number of one-stage methods are

Table 2 Methods of object detection

Methods	Two-stage methods	One-stage methods
2D detection	SIFT Lowe (1999), FAST Rosten and Drummond (2005), SURF Bay et al. (2006), ORB Rublee et al. (2011), OverFeat Sermanet et al. (2013), Erhan et al. (2014), Szegedy et al. (2014), RCNN Girshick et al. (2014), Fast R-CNN Girshick (2015), Faster RCNN Ren et al. (2015), R-FCN Dai et al. (2016), FPN Lin et al. (2017)	YOLO Redmon et al. (2016), SSD Liu et al. (2016), YOLOv2 Redmon and Farhadi (2017), RetinaNet Lin et al. (2017), YOLOv3 Redmon and Farhadi (2018), FCOS Tian et al. (2019), CornerNet Law and Deng (2018), ExtremeNet Zhou et al. (2019), CenterNet Zhou et al. (2019); Duan et al. (2019), CentripetalNet Dong et al. (2020), YOLOv4 Bochkovskiy et al. (2020)
3D detection	Spin Images Johnson (1997), 3D Shape Context Frome et al. (2004), FPFH Rusu et al. (2009), CVFH Aldoma et al. (2011), SHOT Salti et al. (2014), Sliding Shapes Song and Xiao (2014), Frustum PointNets Qi et al. (2018), PointFusion Xu et al. (2018), FrustumConvNet Wang and Jia (2019), Deep Sliding Shapes Song and Xiao (2016), MV3D Chen et al. (2017), MMF Liang et al. (2019), Part-A^2 Shi et al. (2020), PV-RCNN Shi et al. (2020), PointRCNN Shi et al. (2019), STD Yang et al. (2019), VoteNet Qi et al. (2019), MLCVNet Xie et al. (2020), H3DNet Zhang et al. (2020), ImVoteNet Qi et al. (2020)	VoxelNet Zhou and Tuzel (2018), SECOND Yan et al. (2018), PointPillars Lang et al. (2019), TANet Liu et al. (2020), HVNet Ye et al. (2020), 3DSSD Yang et al. (2020), Point-GNN Shi and Rajkumar (2020), DOPS Najibi et al. (2020), Associate-3Ddet Du et al. (2020)

proposed which achieved high accuracies with high speed. Typical functional flow-chart of 2D object detection is illustrated in Fig. 7.

Two-stage methods The two-stage methods can be referred as region proposal-based methods. Most of the traditional methods utilize the sliding window strategy to obtain the bounding boxes first, and then utilize feature descriptions of the bounding boxes for classification. Large number of hand-crafted global descriptors and local descriptors are proposed, such as SIFT Lowe (1999), FAST Rosten and Drummond (2005), SURF Bay et al. (2006), ORB Rublee et al. (2011), and so on. Based on these descriptors, researchers trained classifiers, such as neural networks, Support Vector Machine (SVM) or Adaboost, to conduct 2D detection. There exist some disadvantages of traditional detection methods. For example, the sliding windows should be predefined for specific objects, and the hand-crafted features are not representative enough for a strong classifier.

With the development of deep learning, region proposals could be computed with a deep neural network. OverFeat Sermanet et al. (2013) trained a fully connected layer to predict the box coordinates for the localization task that assumes a single object. Erhan et al. (2014) and Szegedy et al. (2014) generated region proposals from a network whose last fully connected layer simultaneously predicts multiple boxes. Besides, deep neural networks extract more representative features than hand-crafted features, and training classifiers using CNN Krizhevsky et al. (2012) features highly improved the performance. R-CNN Girshick et al. (2014) uses Selective Search (SS) Uijlings et al. (2013) methods to generate region proposals, uses CNN to extract features and trains classifiers using SVM. This traditional classifier is replaced by directly regressing the bounding boxes using the Region of Interest (ROI) feature vector in Fast R-CNN Girshick (2015). Faster R-CNN Ren et al. (2015) is further proposed by replacing SS with the Region Proposal Network (RPN),

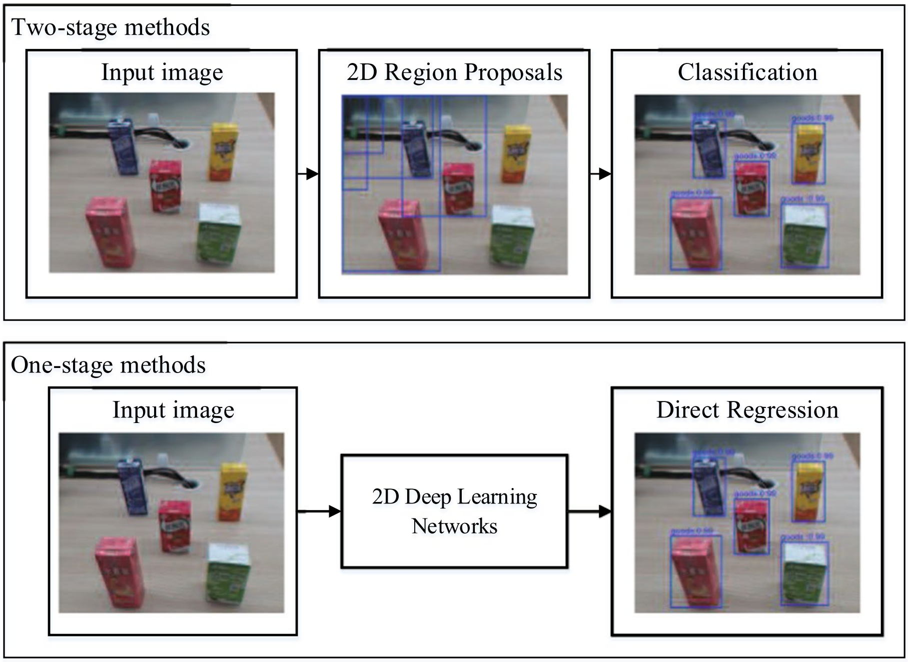

Fig. 7 Typical functional flow-chart of 2D object detection

which is a kind of fully convolutional network (FCN) Long et al. (2015) and can be trained end-to-end specifically for the task of generating detection proposals. This design is also adopted in other two-stage methods, such as R-FCN Dai et al. (2016), FPN Lin et al. (2017). Generally, two-stage methods achieve a higher accuracy, whereas need more computing resources or computing time.

One-stage methods The one-stage methods can also be referred as regression-based methods. Compared to two-stage approaches, the single-stage pipeline skips separate object proposal generation and predicts bounding boxes and class scores in one evaluation. YOLO Redmon et al. (2016) conducts joint grid regression, which simultaneously predicts multiple bounding boxes and class probabilities for those boxes. YOLO is not suitable for small objects, since it only regress two bounding boxes for each grid. SSD Liu et al. (2016) predicts category scores and box offsets for a fixed set of anchor boxes produced by the sliding window. Compared with YOLO, SSD is faster and much more accurate. YOLOv2 Redmon and Farhadi (2017) also adopts sliding window anchors for classification and spatial location prediction so as to achieve a higher recall than YOLO. RetinaNet Lin et al. (2017) proposed the focal loss function by reshaping the standard cross entropy loss so that detector will put more focus on hard, misclassified examples during training. RetinaNet achieved comparable accuracy of two-stage detectors with high detection speed. Compare with YOLOv2, YOLOv3 Redmon and Farhadi (2018) and YOLOv4 Bochkovskiy et al. (2020) are further improved with a bunch of improvements, which shows large performance improvements without sacrificing the speed, and is more robust in dealing with small objects. There also exist some anchor-free methods, which doesn't utilize the anchor bounding boxes, such as FCOS Tian et al. (2019), CornerNet Law and Deng (2018), ExtremeNet Zhou et al. (2019), CenterNet Zhou et al. (2019); Duan et al. (2019) and CentripetalNet Dong et al. (2020). Further reviews of these works can refer to recent surveys (Zou et al. 2019; Zhao et al. 2019; Liu et al. 2020; Sultana et al. 2020).

Discussions The 2D object detection methods are widely used in 2D planar robotic grasping tasks. This part can refer to Sect. 4.1.2.

2.2.2 3D object detection

3D object detection aims at finding the amodel 3D bounding box of the target object, which means finding the 3D bounding box that a complete target object occupies. 3D object detection is deeply explored in outdoor scenes and indoor scenes. Aiming at robotic grasping tasks, we can obtain the 2D and 3D information of the scene through RGB-D data, and general 3D object detection methods could be used. Similar with 2D object detection tasks, two-stage methods and one-stage methods both exist. The two-stage methods refer to region proposal-based methods and one-stage methods refer to regression-based methods. Typical functional flow-chart of 3D object detection is illustrated in Fig. 8.

Two-stage methods Traditional 3D detection methods usually aim at objects with known shapes. The 3D object detection problem is transformed into a detection and 6D object pose estimation problem. Many hand-crafted 3D shape descriptors, such as Spin Images Johnson (1997), 3D Shape Context Frome et al. (2004), FPFH Rusu et al. (2009), CVFH Aldoma et al. (2011), SHOT Salti et al. (2014), are proposed, which can locate the object proposals. In addition, the accurate 6D pose of the target object could be achieved through local registration. This part is introduced in Sect. 3.1.2. However, these methods face difficulties in general 3D object detection tasks. Aiming at general 3D object detection tasks, the 3D region proposals are widely used. Traditional methods train classifiers,

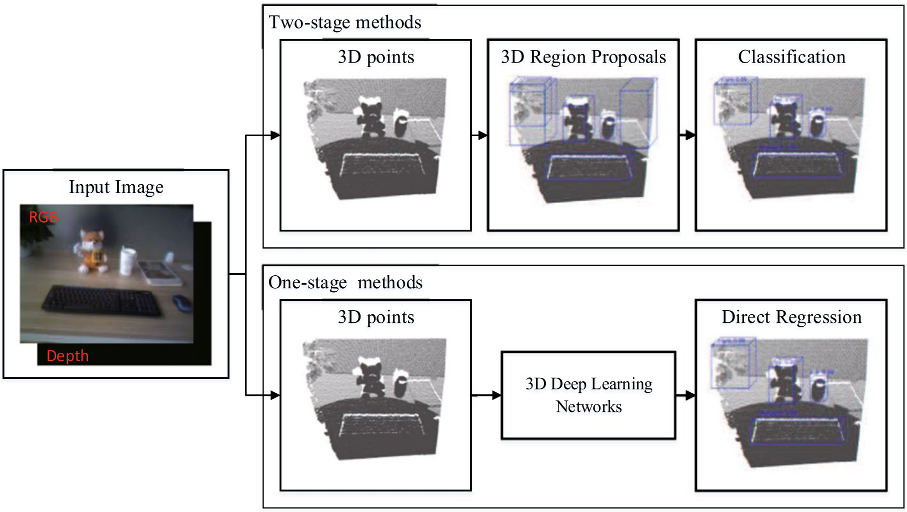

Fig. 8 Typical functional flow-chart of 3D object detection

such as SVM, based on the 3D shape descriptors. Sliding Shapes Song and Xiao (2014) is proposed which slides a 3D detection window in 3D space and extract features from the 3D point cloud to train an Exemplar-SVM classifier Malisiewicz et al. (2011). With the development of deep learning, the 3D region proposals could be generated efficiently, and the 3D bounding boxes could be regressed using features from deep neural networks rather than training traditional classifiers. There exist various methods of generating 3D object proposals, which can be roughly divided into three kinds, which are frustum-based methods (Qi et al. 2018; Xu et al. 2018; Wang and Jia 2019), global regression-based methods (Song and Xiao 2016; Chen et al. 2017; Liang et al. 2019) and local regression-based methods.

Frustum-based methods generate object proposals using mature 2D object detectors, which is a straightforward way. Frustum PointNets Qi et al. (2018) leverages a 2D CNN object detector to obtain 2D regions, and the lifted frustum-like 3D point clouds become 3D region proposals. The amodel 3D bounding boxes are regressed from features of the segmented points within the proposals based on PointNet Qi et al. (2017). PointFusion Xu et al. (2018) utilized Faster R-CNN Ren et al. (2015) to obtain the image crop first, and deep features from the corresponding image and the raw point cloud are densely fused to regress the 3D bounding boxes. FrustumConvNet Wang and Jia (2019) also utilizes the 3D region proposals lifted from the 2D region proposal and generates a sequence of frustums for each region proposal.

Global regression-based methods generate 3D region proposals from feature representations extracted from single or multiple inputs. Deep Sliding Shapes Song and Xiao (2016) proposed the first 3D Region Proposal Network (RPN) using 3D convolutional neural networks (ConvNets) and the first joint Object Recognition Network (ORN) to extract geometric features in 3D and color features in 2D to regress 3D bounding boxes. MV3D Chen et al. (2017) represents the point cloud using the bird's-eye view and employs 2D convolutions to generate 3D proposals. The region-wise features obtained via ROI pooling for multi-view data are fused to jointly predict the 3D bounding boxes. MMF Liang et al.

(2019) proposed a multi-task multi-sensor fusion model for 2D and 3D object detection, which generates a small number of high-quality 3D detections using multi-sensor fused features, and applies ROI feature fusion to regress more accurate 2D and 3D boxes. Part-A^2 Shi et al. (2020) predicts intra-object part locations and generates 3D proposals by feeding the point cloud to an encoder-decoder network. A RoI-aware point cloud pooling is proposed to aggregate the part information from each 3D proposal, and a part-aggregation network is proposed to refine the results. PV-RCNN Shi et al. (2020) utilized voxel CNN with 3D sparse convolution (Graham and van der Maaten 2017; Graham et al. 2018) for feature encoding and proposals generation, and proposed a voxel-to-keypoint scene encoding via voxel set abstraction and a keypoint-to-grid RoI feature abstraction for proposal refinement. PV-RCNN achieved remarkable 3D detection performance on outdoor scene datasets.

Local regression-based methods mean generating point-wise 3D region proposals. PointRCNN Shi et al. (2019) extracts point-wise feature vectors from the input point cloud and generates 3D proposal from each foreground point computed through segmentation. Point cloud region pooling and canonical 3D bounding box refinement are then conducted. STD Yang et al. (2019) designs spherical anchors and a strategy in assigning labels to anchors to generate accurate point-based proposals, and a PointsPool layer is proposed to generate dense proposal features for the final box prediction. VoteNet Qi et al. (2019) proposed a deep hough voting strategy to generate 3D vote points from sampled 3D seeds points. The 3D vote points are clustered to obtain object proposals which will be further refined. MLCVNet Xie et al. (2020) proposed Multi-level Context VoteNet which considers the contextual information between the objects. H3DNet Zhang et al. (2020) predicts a hybrid set of geometric primitives such as centers, face centers and edge centers of the 3d bounding boxes, and formulates 3D object detection as regressing and aggregating these geometric primitives. A matching and refinement module is then utilized to classify object proposals and fine-tune the results. Compared with point cloud input-only VoteNet Qi et al. (2019), ImVoteNet Qi et al. (2020) additionally extracts geometric and semantic features from the 2D images, and fuses the 2D features into the 3D detection pipeline, which achieved remarkable 3D detection performance on indoor scene datasets.

One-stage methods One-stage methods directly predict class probabilities and regress the 3D amodal bounding boxes of the objects using a single-stage network. These methods do not need region proposal generation or post-processing. VoxelNet Zhou and Tuzel (2018) divides a point cloud into equally spaced 3D voxels and transforms a group of points within each voxel into a unified feature representation. Through convolutional middle layers and the region proposal network, the final results are obtained. Compared with VoxelNet, SECOND Yan et al. (2018) applies sparse convolution layers Graham et al. (2018) for parsing the compact voxel features. PointPillars Lang et al. (2019) converts a point cloud to a sparse pseudo-image, which is processed into a high-level representation through a 2D convolutional backbone. The features from the backbone are used by the detection head to predict 3D bounding boxes for objects. TANet Liu et al. (2020) proposed a Triple Attention (TA) module and a Coarse-to-Fine Regression (CFR) module, which focuses on the 3D detection of hard objects and the robustness to noisy points. HVNet Ye et al. (2020) proposed a hybrid voxel network which fuses voxel feature encoder (VFE) of different scales at point-wise level and projects into multiple pseudo-image feature maps. Above methods are mainly voxel-based 3D single stage detectors, and Yang et al. (2020) proposed a point-based 3D single stage object detector called 3DSSD, which contain a fusion sampling strategy in the downsampling process, a candidate generation layer, and an anchor-free regression head with a 3D center-ness assignment strategy. They achieved a good balance between accuracy and efficiency. Point-GNN Shi and Rajkumar (2020) utilized graph neural network on the point cloud and designed a graph neural network with an auto-registration mechanism which detects multiple objects in a single shot. DOPS Najibi et al. (2020) proposed an object detection pipeline which utilizes a 3D sparse U-Net Graham and van der Maaten (2017) and a graph convolution module. Their method can jointly predict the 3D shapes of the objects. Associate-3Ddet Du et al. (2020) learns to associate feature extracted from the real scene with more discriminative feature from class-wise conceptual models. Comprehensive review about 3D object detection could refer to the survey Guo et al. (2019).

Discussions 3D object detection only presents the general shape of the target object, which is not sufficient to conduct a robotic grasp, and it is mostly used in autonomous driving areas. However, the estimated 3D bounding boxes could provide approximate grasp positions and provide valuable information for the collision detection.

2.3 Object instance segmentation

Object instance segmentation refers to detecting the pixel-level or point-level instance objects of a certain class, which is closely related to object detection and semantic segmentation tasks. Two kinds of methods also exist, which are two-stage methods and one-stage methods. The two-stage methods refer to region proposal-based methods and one-stage methods refer to regression-based methods. The representative works of the two methods are shown in Table 3 aiming at 2D inputs and 3D inputs.

2.3.1 2D object instance segmentation

2D object instance segmentation means detecting the pixel-level instance objects of a certain class from an input image, which is usually represented as masks. Two-stage methods follow the mature object detection frameworks, while one-stage methods conduct regression from the whole input image directly. Typical functional flow-chart of 2D object instance segmentation is illustrated in Fig. 9.

Two-stage methods This kind of methods could also be referred as region proposal-based methods. The mature 2D object detectors are used to generate bounding boxes or region proposals, and the object masks are then predicted within the bounding boxes. Lots of methods are based on convolutional neural networks (CNN). SDS Hariharan et al. (2014) uses CNN to classify category-independent region proposals. MNC Dai et al. (2016) conducts instance segmentation via three networks, respectively differentiating instances, estimating masks, and categorizing objects. Path Aggregation Network (PANet) Liu et al. (2018) was proposed which boosts the information flow in the proposal-based instance segmentation framework. Mask R-CNN He et al. (2017) extends Faster R-CNN Ren et al. (2015) by adding a branch for predicting an object mask in parallel with the existing branch for bounding box recognition, which achieved promising results. MaskLab Chen et al. (2018) also builds on top of Faster R-CNN Ren et al. (2015) and additionally produces semantic and instance center direction outputs. Chen et al. (2019) proposed a framework called Hybrid Task Cascade (HTC), which performs cascaded refinement on object detection and segmentation jointly and adopts a fully convolutional branch to provide spatial context. PointRend Kirillov et al. (2020) performs point-based segmentation predictions at adaptively selected locations based on an iterative subdivision algorithm. PointRend can be flexibly applied to instance

Table 3 Methods of object instance segmentation

Methods	Two-stage methods	One-stage methods
2D instance segmentation	SDS Hariharan et al. (2014), MNC Dai et al. (2016), PANet Liu et al. (2018), Mask R-CNN He et al. (2017), MaskLab Chen et al. (2018), HTC Chen et al. (2019), PointRend Kirillov et al. (2020), FGN Fan et al. (2020)	DeepMask Pinheiro et al. (2015), SharpMask Pinheiro et al. (2016), InstanceFCN Dai et al. (2016), FCIS Li et al. (2017), TensorMask Chen et al. (2019), YOLACT Bolya et al. (2019), YOLACT++ Bolya et al. (2019), PolarMask Xie et al. (2020), SOLO Wang et al. (2019), CenterMask Lee and Park (2020), BlendMask Chen et al. (2020)
3D instance segmentation	GSPN Yi et al. (2019), 3D-SIS Hou et al. (2019), 3D-MPA Engelmann et al. (2020)	SGPN Wang et al. (2018), MASC Liu and Furukawa (2019), ASIS Wang et al. (2019), JSIS3D Pham et al. (2019), JSNet Zhao and Tao (2020), 3D-BoNet Yang et al. (2019), LiDARSeg Zhang et al. (2020), OccuSeg Han et al. (2020)

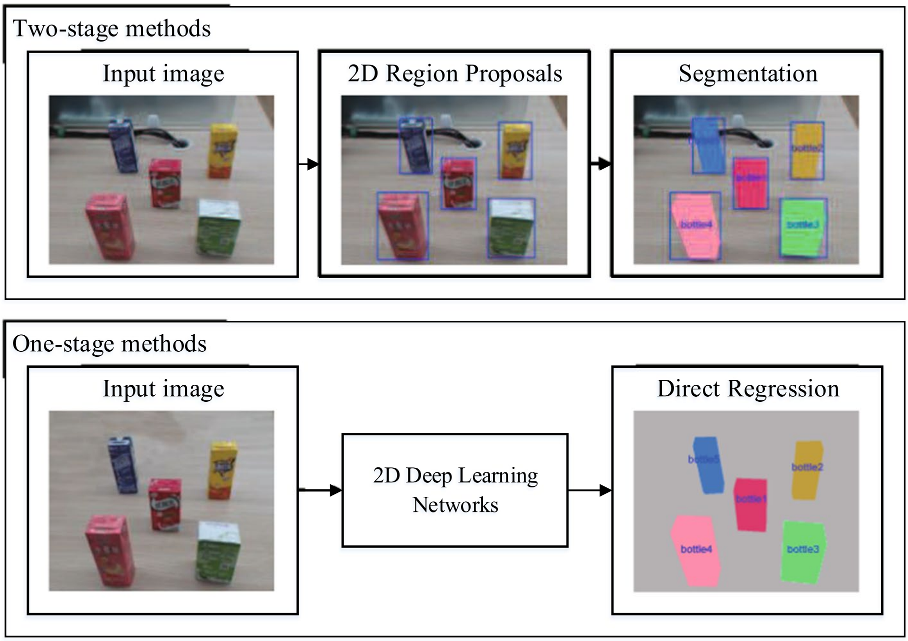

Fig. 9 Typical functional flow-chart of 2D object instance segmentation

segmentation tasks by building on top of them, and yields significantly more detailed results. FGN Fan et al. (2020) proposed a Fully Guided Network (FGN) for few-shot instance segmentation, which introduces different guidance mechanisms into the various key components in Mask R-CNN He et al. (2017).

Single-stage methods This kind of methods could also be referred as regression-based methods, where the segmentation masks are predicted as well the objectness score. DeepMask Pinheiro et al. (2015), SharpMask Pinheiro et al. (2016) and InstanceFCN Dai et al. (2016) predict segmentation masks for the the object located at the center. FCIS Li et al. (2017) was proposed as the fully convolutional instance-aware semantic segmentation method, where position-sensitive inside/outside score maps are used to perform object segmentation and detection. TensorMask Chen et al. (2019) uses structured 4D tensors to represent masks over a spatial domain and presents a framework to predict dense masks. YOLACT Bolya et al. (2019) breaks instance segmentation into two parallel subtasks, which are generating a set of prototype masks and predicting per-instance mask coefficients. YOLACT is the first real-time one-stage instance segmentation method and is improved by YOLACT++ Bolya et al. (2019). PolarMask Xie et al. (2020) formulates the instance segmentation problem as predicting contour of instance through instance center classification and dense distance regression in a polar coordinate. SOLO Wang et al. (2019) introduces the notion of instance categories, which assigns categories to each pixel within an instance according to the instance's location and size, and converts instance mask segmentation into a classification-solvable problem. CenterMask Lee and Park (2020) adds a novel spatial attention-guided mask (SAG-Mask) branch to anchor-free one stage object detector (FCOS Tian et al. (2019)) in the same vein with Mask R-CNN He et al. (2017). BlendMask Chen et al. (2020) also builds upon the FCOS Tian et al. (2019) object detector, which uses a blender module to effectively predict dense per-pixel position-sensitive instance features and learn attention maps for each instance. Detailed reviews refer to the survey (Minaee et al. 2020; Sultana et al. 2020; Hafiz and Bhat 2020).

Discussions 2D object instance segmentation is widely used in robotic grasping tasks. For example, SegICP Wong et al. (2017) utilize RGB-based object segmentation to obtain the points belong to the target objects. Xie et al. (2020) separately leverage RGB and Depth for unseen object instance segmentation. Danielczuk et al. (2019) segments unknown 3d objects from real depth images using Mask R-CNN He et al. (2017) trained on synthetic data.

2.3.2 3D object instance segmentation

3D object instance segmentation means detecting the point-level instance objects of a certain class from an input 3D point cloud. Similar to 2D object instance segmentation, two-stage methods need region proposals, while one-stage methods are proposal-free. Typical functional flow-chart of 3D object instance segmentation is illustrated in Fig. 10.

Two-stage methods This kind of methods could also be referred as proposal-based methods. General methods utilize the 2D or 3D detection results and conduct foreground or background segmentation in the corresponding frustum or bounding boxes. GSPN Yi et al. (2019) proposed the Generative Shape Proposal Network (GSPN) to generates 3D object proposals and the Region-PointNet framework to conduct 3D object instance segmentation. 3D-SIS Hou et al. (2019) leverages joint 2D and 3D end-to-end feature learning on both geometry and RGB input for 3D object bounding box detection and semantic instance segmentation. 3D-MPA Engelmann et al. (2020) predicts dense object centers based on learned semantic features from a sparse volumetric backbone, employes a graph convolutional network to explicitly model higher-order interactions between neighboring proposal features, and utilizes a multi proposal aggregation strategy other than NMS to obtain the final results.

Single-stage methods This kind of methods could also be referred as regression-based methods. Lots of methods learn to group per-point features to segment 3D instances. SGPN Wang et al. (2018) proposed the Similarity Group Proposal Network (SGPN) to

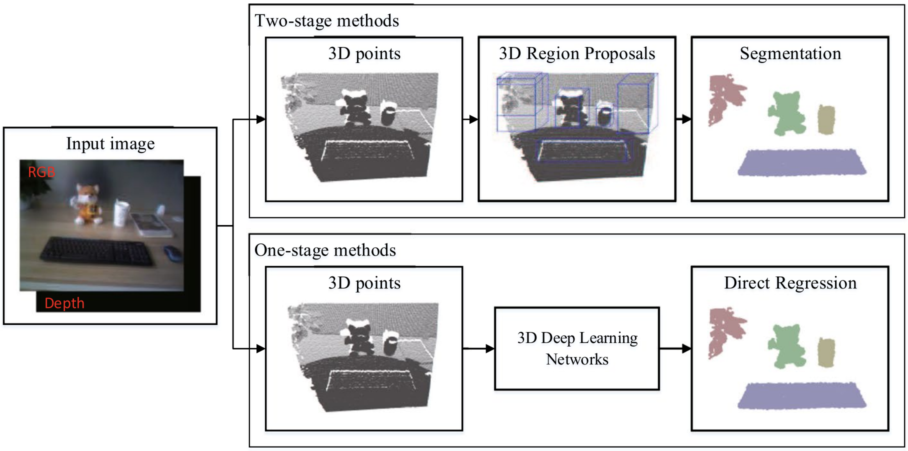

Fig. 10 Typical functional flow-chart of 3D object instance segmentation

predict point grouping proposals and a corresponding semantic class for each proposal, from which we can directly extract instance segmentation results. MASC Liu and Furukawa (2019) utilizes the sub-manifold sparse convolutions (Graham and van der Maaten 2017; Graham et al. 2018) to predict semantic scores for each point as well as the affinity between neighboring voxels at different scales. The points are then grouped into instances based on the predicted affinity and the mesh topology. ASIS Wang et al. (2019) learns semantic-aware point-level instance embedding and semantic features of the points belonging to the same instance are fused together to make per-point semantic predictions. JSIS3D Pham et al. (2019) proposed a multi-task point-wise network (MT-PNet) that simultaneously predicts the object categories of 3D points and embeds these 3D points into high dimensional feature vectors that allow clustering the points into object instances. JSNet Zhao and Tao (2020) also proposed a joint instance and semantic segmentation (JISS) module and designed an efficient point cloud feature fusion (PCFF) module to generate more discriminative features. 3D-BoNet Yang et al. (2019) was proposed to directly regress 3D bounding boxes for all instances in a point cloud, while simultaneously predicting a point-level mask for each instance. LiDARSeg Zhang et al. (2020) proposed a dense feature encoding technique, a solution for single-shot instance prediction and effective strategies for handling severe class imbalances. OccuSeg Han et al. (2020) proposed an occupancy-aware 3D instance segmentation scheme, which predicts the number of occupied voxels for each instance. The occupancy signal guides the clustering stage of 3D instance segmentation and OccuSeg achieves remarkable performance.

Discussions 3D object instance segmentation is quite important in robotic grasping tasks. However, current methods mainly leverage 2D instance segmentation methods to obtain the 3D point cloud of the target object, which utilizes the advantages of RGB-D images. Nowadays 3D object instance segmentation is still a fast developing area, and it will be widely used in the future if its performance and speed improve a lot.

3 Object pose estimation

In some 2D planar grasps, the target objects are constrained in the 2D workspace and are not piled up, the 6D object pose can be represented as the 2D position and the in-plane rotation angle. This case is relatively simple and is addressed quite well based on matching 2D feature points or 2D contour curves. In other 2D planar grasp and 6DoF grasp scenarios, the 6D object pose is mostly needed, which helps a robot to get aware of the 3D position and 3D orientation of the target object. The 6D object pose transforms the object from the object coordinate into the camera coordinate. We mainly focus on 6D object pose estimation in this section and divide 6D object pose estimation into three kinds, which are correspondence-based, template-based and voting-based methods. During the review of each kind of methods, both traditional methods and deep learning-based methods are reviewed.

3.1 Correspondence-based methods

Correspondence-based 6D object pose estimation involves methods of finding correspondences between the observed input data and the existing complete 3D object model. When we want to solve this problem based on the 2D RGB image, we need to find correspondences between 2D pixels and 3D points of the existing 3D model. The 6D object pose can thus be recovered through Perspective-n-Point (PnP) algorithms Lepetit et al. (2009). When we want to solve this problem based on the 3D point cloud lifted from the depth image, we need to find correspondences of 3D points between the observed partial-view point cloud and the complete 3D model. The 6D object pose can thus recovered through least square methods. The methods of correspondence-based methods are summarized in Table 4.

3.1.1 2D image-based methods

When using the 2D RGB image, correspondence-based methods mainly target on the objects with rich texture through the matching of 2D feature points, as shown in Fig. 11. Multiple images are first rendered by projecting the existing 3D models from various angles and each object pixel in the rendered images corresponds to a 3D point. Through matching 2D feature points on the observed image and the rendered images (Vacchetti et al. 2004; Lepetit et al. 2005), the 2D–3D correspondences are established. Other than rendered images, the keyframes in keyframe-based SLAM approaches Mur-Artal et al. (2015) could also provide 2D–3D correspondences for 2D keypoints. The common 2D descriptors such as SIFT Lowe (1999), FAST Rosten and Drummond (2005), SURF Bay et al. (2006), ORB Rublee et al. (2011), etc., are usually utilized for the 2D feature matching. Based on the 2D–3D correspondences, the 6D object pose can be calculated with Perspective-n-Point (PnP) algorithms Lepetit et al. (2009). However, these 2D feature-based methods fail when the objects do not have rich texture.

With the development of deep neural networks such as CNN, representative features could be extracted from the image. A straightforward way is to extract discriminative feature points (Yi et al. 2016; Truong et al. 2019) and match them using the representative CNN features. Yi et al. (2016) presented a SIFT-like feature descriptor. Truong et al. (2019) presented a method to greedily learn accurate match points. Superpoint DeTone et al. (2018) proposed a self-supervised framework for training interest point detectors and descriptors, which shows advantages over a few traditional feature detectors and descriptors. LCD Pham et al. (2020) particularly learns a local cross-domain descriptor for 2D image and 3D point cloud matching, which contains a dual auto-encoder neural network that maps 2D and 3D inputs into a shared latent space representation.

There exists another kind of methods (Rad and Lepetit 2017; Tekin et al. 2018; Crivellaro et al. 2017; Hu et al. 2019), which uses the representative CNN features to predict the 2D locations of 3D points, as shown in Fig. 11. Since it's difficult to selected the 3D points to be projected, many methods utilize the eight vertices of the object's 3D bounding box. Rad and Lepetit (2017) predicts 2D projections of the corners of their 3D bounding boxes and obtains the 2D–3D correspondences. Different with them, Tekin et al. (2018) proposed a single-shot deep CNN architecture that directly detects the 2D projections of the 3D bounding box vertices without posteriori refinements. Some other methods utilize feature points of the 3D object. Crivellaro et al. (2017) predicts the pose of each part of the object in the form of the 2D projections of a few control points with the assistance of a Convolutional Neural Network (CNN). KeyPose Liu et al. (2020) predicts object poses using 3D keypoints from stereo input, and is suitable for transparent objects. Hu et al. (2020) further predicts the 6D object pose from a group of candidate 2D–3D correspondences using deep learning networks in a single-stage manner, instead of RANSAC-based Perspective-n-Point (PnP) algorithms. HybridPose Song et al. (2020) predicts a hybrid intermediate representation to express different geometric information in the input image,

Table 4 Summary of correspondence-based 6D object pose estimation methods

Methods	Descriptions	Traditional methods	Deep learning-based methods
2D image-based methods	Find correspondences between 2D pixels and 3D points, and use PNP methods	SIFT Lowe (1999), FAST Rosten and Drummond (2005), SURF Bay et al. (2006), ORB Rublee et al. (2011)	LCD Pham et al. (2020), BB8 Rad and Lepetit (2017), Tekin et al. (2018), Crivellaro et al. (2017), KeyPose Liu et al. (2020), Hu et al. (2020), HybridPose Song et al. (2020), Hu et al. (2019), DPOD Zakharov et al. (2019), Pix2pose Park et al. (2019), EPOS Hodan et al. (2020)
3D point cloud-based methods	Find correspondences between 3D points	Spin Images Johnson (1997), 3D Shape Context Frome et al. (2004), FPFH Rusu et al. (2009), CVFH Aldoma et al. (2011), SHOT Salti et al. (2014)	3DMatch Zeng et al. (2017a), 3DFeat-Net Yew and Lee (2018), Gojcic et al. (2019), Yuan et al. (2020), StickyPillars Simon et al. (2020)

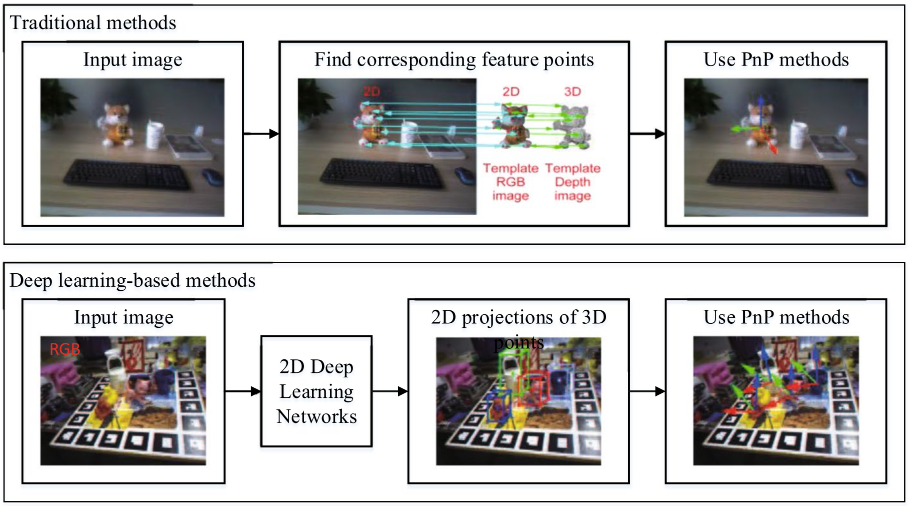

Fig. 11 Typical functional flow-chart of 2D correspondence-based 6D object pose estimation methods. Data from the lineMod dataset Hinterstoisser et al. (2012)

including keypoints, edge vectors, and symmetry correspondences. Some other methods predict 3D positions for all the pixels of the object. Hu et al. (2019) proposed a segmentation-driven 6D pose estimation framework where each visible part of the object contributes to a local pose prediction in the form of 2D keypoint locations. The pose candidates are them combined into a robust set of 2D–3D correspondences from which the reliable pose estimation result is computed. DPOD Zakharov et al. (2019) estimates dense multi-class 2D–3D correspondence maps between an input image and available 3D models. Pix-2pose Park et al. (2019) regresses pixel-wise 3D coordinates of objects from RGB images using 3D models without textures. EPOS Hodan et al. (2020) represents objects by surface fragments which allows to handle symmetries, predicts a data-dependent number of precise 3D locations at each pixel, which establishes many-to-many 2D–3D correspondences, and utilizes an estimator for recovering poses of multiple object instances.

3.1.2 3D point cloud-based methods

Typical functional flow-chart of 3D correspondence-based 6D object pose estimation methods is illustrated in Fig. 12. When using the 3D point cloud lifted from the depth image, 3D geometric descriptors could be utilized for matching, which eliminates the influence of the texture. The 6D object pose could then be achieved by computing the transformations based on 3D-3D correspondences directly. The widely used 3D local shape descriptors, such as Spin Images Johnson (1997), 3D Shape Context Frome et al. (2004), FPFH Rusu et al. (2009), CVFH Aldoma et al. (2011), SHOT Salti et al. (2014), can be utilized to find correspondences between the object's partial 3D point cloud and full point cloud to obtain the 6D object pose. Some other 3D local descriptors could refer to the survey Guo et al. (2016). However, this kind of methods require that the target objects have rich geometric features.

There also exist deep learning-based 3D descriptors (Zeng et al. 2017a; Yew and Lee 2018) aiming at matching 3D points, which are representative and discriminative.

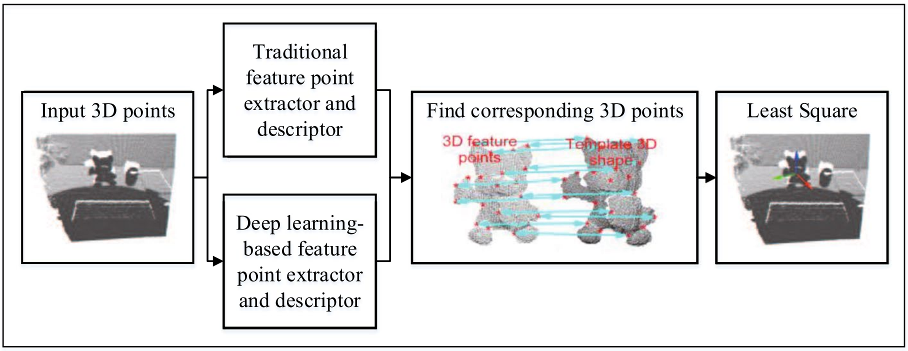
Input 3D points
Traditional feature point extractor and descriptor
Deep learning-based feature point extractor and descriptor
Find corresponding 3D points
3D feature points
Template 3D shape
Least Square

Fig. 12 Typical functional flow-chart of 3D correspondence-based 6D object pose estimation methods

3DMatch Zeng et al. (2017a) is proposed to match 3D feature points using 3D voxel-based deep learning networks. 3DFeat-Net Yew and Lee (2018) proposed a weakly supervised network that holistically learns a 3D feature detector and descriptor using only GPS/INS tagged 3D point clouds. Gojcic et al. (2019) proposed 3DSmoothNet, which matches 3D point clouds with a siamese deep learning architecture and fully convolutional layers using a voxelized smoothed density value (SDV) representation. Yuan et al. (2020) proposed a self-supervised learning method for descriptors in point clouds, which requires no manual annotation and achieves competitive performance. StickyPillars Simon et al. (2020) proposed an end-to-end trained 3D feature matching approach based on a graph neural network, and they perform context aggregation with the aid of transformer based multi-head self and cross attention.

3.2 Template-based methods

Template-based 6D object pose estimation involves methods of finding the most similar template from the templates that are labeled with Ground Truth 6D object poses. In 2D case, the templates could be projected 2D images from known 3D models, and the objects within the templates have corresponding 6D object poses in the camera coordinate. The 6D object pose estimation problem is thus transformed into an image retrieval problem. In 3D case, the template could be the full point cloud of the target object. We need to find the best 6D pose that aligns the partial point cloud to the template and thus the 6D object pose estimation becomes a part-to-whole coarse registration problem. The methods of template-based methods are summarized in Table 5.

3.2.1 2D image-based methods

Traditional 2D feature-based methods could be used to find the most similar template image and 2D correspondence-based methods could be utilized if discriminative feature points exist. Therefore, this kind of methods mainly aim at texture-less or non-texture objects that correspondence-based methods can hardly deal with. In these methods, the gradient information is usually utilized. Typical functional flow-chart of 2D template-based 6D object pose estimation methods is illustrated in Fig. 13. Multiple images which are generated by projecting the existing complete 3D model from various angles

Table 5 Summary of template-based 6D object pose estimation methods

Methods	Descriptions	Traditional methods	Deep learning-based methods
2D image-based methods	Retrieve the template image that is most similar with the observed image	LineMod Hinterstoisser et al. (2012), Hodaň et al. (2015)	AAE Sundermeyer et al. (2018), PoseCNN Xiang et al. (2018), SSD6D Kehl et al. (2017), Deep-6DPose Do et al. (2018), Liu et al. (2019), CDPN Li et al. (2019), Tian et al. (2020), NOCS Wang et al. (2019), LatentFusion Park et al. (2020), Chen et al. (2020)
3D point cloud-based methods	Find the pose that best aligns the observed partial 3D point cloud with the template full 3D model	Super4PCS Mellado et al. (2014), Go-ICP Yang et al. (2015)	PCRNet Sarode et al. (2019), DCP Wang and Solomon (2019), PointNetLK Aoki et al. (2019), PRNet Wang and Solomon (2019), DeepICP Lu et al. (2019), Sarode et al. (2019), TEASER Yang et al. (2020), DGR Choy et al. (2020), G2L-Net Chen et al. (2020), Gao et al. (2020)

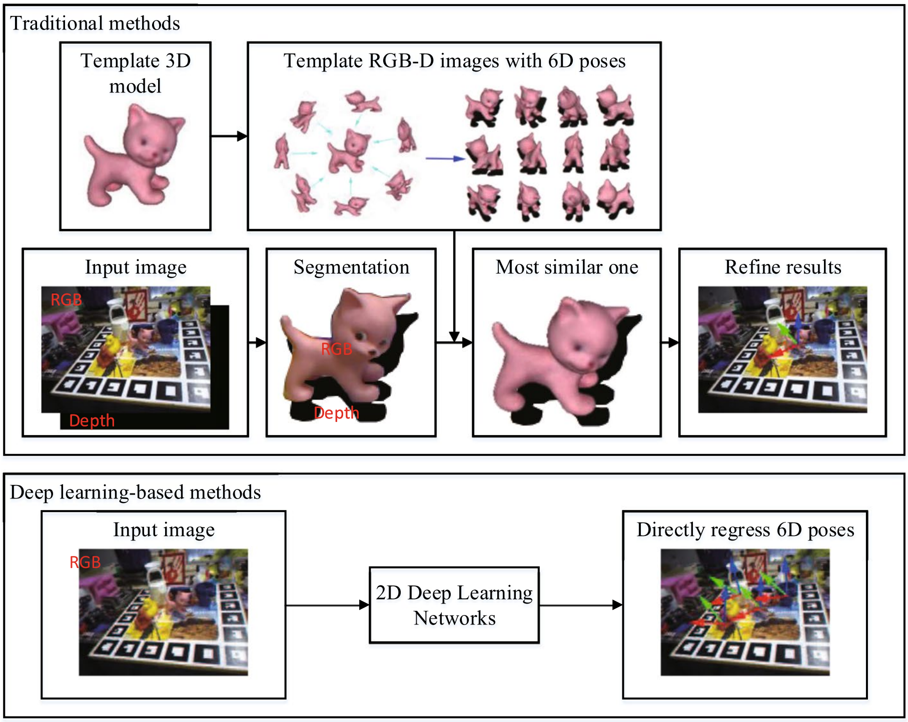

Fig. 13 Typical functional flow-chart of 2D template-based 6D object pose estimation methods. Data from the lineMod dataset Hinterstoisser et al. (2012)

are regarded as the templates. Hinterstoisser et al. (2012) proposed a novel image representation by spreading image gradient orientations for template matching and represented a 3D object with a limited set of templates. The accuracy of the estimated pose was improved by taking into account the 3D surface normal orientations which are computed from the dense point cloud obtained from a dense depth sensor. Hodaň et al. (2015) proposed a method for the detection and accurate 3D localization of multiple texture-less and rigid objects depicted in RGB-D images. The candidate object instances are verified by matching feature points in different modalities and the approximate object pose associated with each detected template is used as the initial value for further optimization. There exist deep learning-based image retrieval methods Gordo et al. (2016), which could assist the template matching process. However, seldom of them were used in template-based methods and perhaps the number of templates is too small for deep learning methods to learn representative and discriminative features.

Above methods find the most similar template explicitly, and there also exist some implicitly ways. Sundermeyer et al. (2018) proposed Augmented Autoencoders (AAE), which learns the 3D orientation implicitly. Thousands of template images are rendered from a full 3D model and these template images are encoded into a codebook. The input image will be encoded into a new code and matched with the codebook to find the most similar template image, and the 6D object pose is thus obtained.

There also exist methods (Xiang et al. 2018; Do et al. 2018; Liu et al. 2019) that directly estimate the 6D pose of the target object from the input image, which can be regarded as finding the most similar image from the pre-trained and labeled images implicitly. Different from correspondence-based methods, this kind of method learns the immediate mapping from an input image to a parametric representation of the pose, and the 6D object pose can thus be estimated combined with object detection Patil and Rabha (2018). Xiang et al. (2018) proposed PoseCNN for direct 6D object pose estimation. The 3D translation of an object is estimated by localizing the center in the image and predicting the distance from the camera, and the 3D rotation is computed by regressing a quaternion representation. Kehl et al. (2017) presented a similar method by making use of the SSD network. Do et al. (2018) proposed an end-to-end deep learning framework named Deep-6DPose, which jointly detects, segments, and recovers 6D poses of object instances form a single RGB image. They extended the instance segmentation network Mask R-CNN He et al. (2017) with a pose estimation branch to directly regress 6D object poses without any post-refinements. Liu et al. (2019) proposed a two-stage CNN architecture which directly outputs the 6D pose without requiring multiple stages or additional post-processing like PnP. They transformed the pose estimation problem into a classification and regression task. CDPN Li et al. (2019) proposed the Coordinates-based Disentangled Pose Network (CDPN), which disentangles the pose to predict rotation and translation separately. Tian et al. (2020) also proposed a discrete-continuous formulation for rotation regression to resolve this local-optimum problem. They uniformly sample rotation anchors in $SO(3)$, and predict a constrained deviation from each anchor to the target.

There also exist methods that build a latent representation for category-level objects. This kind of methods can also be treated as the template-based methods, and the template could be implicitly built from multiple images. NOCS Wang et al. (2019), LatentFusion Park et al. (2020) and Chen et al. (2020) are the representative methods.

3.2.2 3D point cloud-based methods

Typical functional flow-chart of 3D template-based 6D object pose estimation methods is illustrated in Fig. 14. Traditional partial registration methods aim at finding the 6D transformation that best aligns the partial point cloud to the full point cloud. Various global

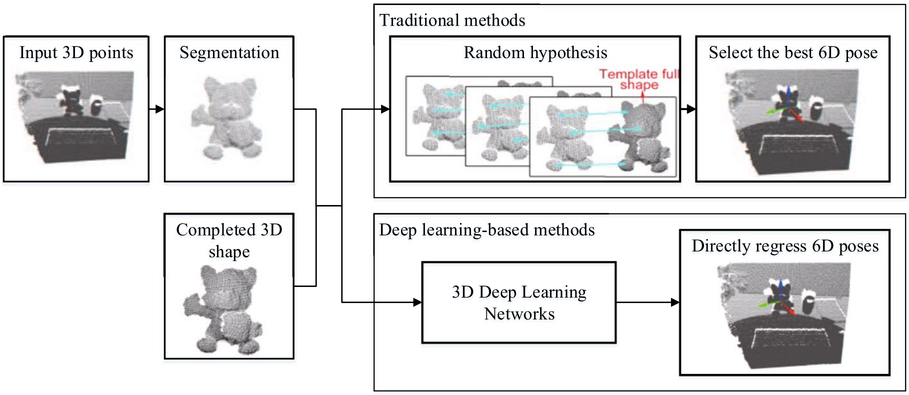

Fig. 14 Typical functional flow-chart of 3D template-based 6D object pose estimation methods

registration methods (Mellado et al. 2014; Yang et al. 2015; Zhou et al. 2016) exist which afford large variations of initial poses and are robust with large noise. However, this kind of method is time-consuming. Most of these methods utilize local registration methods such as the iterative closest points(ICP) algorithm Besl and McKay (1992) to refine the results. This part can refer to some review papers (Tam et al. 2013; Bellekens et al. 2014).

Some deep learning-based methods also exist, which can accomplish the partial registration task in an efficient way. These methods consume a pair of point clouds, extract representative and discriminative features from 3D deep learning networks, and regress the relative 6D transformations between the pair of point clouds. PCRNet Sarode et al. (2019), DCP Wang and Solomon (2019), PointNetLK Aoki et al. (2019), PRNet Wang and Solomon (2019), DeepICP Lu et al. (2019), Sarode et al. (2019), TEASER Yang et al. (2020) and DGR Choy et al. (2020) are the representative methods and readers could refer to the recent survey Villena-Martinez et al. (2020). There also exist methods (Chen et al. 2020; Gao et al. 2020) that directly regress the 6D object pose from the partial point cloud. G2L-Net Chen et al. (2020) extracts the coarse object point cloud from the RGB-D image by 2D detection, and then conducts translation localization and rotation localization. Gao et al. (2020) proposed a method which conduct 6D object pose regression via supervised learning on point clouds.

3.3 Voting-based methods

Voting-based methods mean that each pixel or 3D point contributes to the 6D object pose estimation by providing one or more votes. We roughly divide voting methods into two kinds, which are indirectly voting methods and directly voting methods. Indirectly voting methods mean that each pixel or 3D point vote for some feature points, which affords 2D–3D correspondences or 3D-3D correspondences. Directly voting methods mean that each pixel or 3D point vote for a certain 6D object coordinate or pose. These methods are summarized in Table 6.

3.3.1 Indirect voting methods

This kind of methods can be regarded as voting for correspondence-based methods. In 2D case, 2D feature points are voted and 2D–3D correspondences could be achieved. In 3D case, 3D feature points are voted and 3D-3D correspondences between the observed partial point cloud and the canonical full point cloud could be achieved. Most of this kind of methods utilize deep learning methods for their strong feature representation capabilities in order to predict better votes. Typical functional flow-chart of indirect voting-based 6D object pose estimation methods is illustrated in Fig. 15.

In 2D case, PVNet Peng et al. (2019) votes projected 2D feature points and then finds the corresponding 2D–3D correspondences to compute the 6D object pose. Yu et al. (2020) proposed a method which votes 2D positions of the object keypoints from vector-fields. They develop a differentiable proxy voting loss (DPVL) which mimics the hypothesis selection in the voting procedure. In 3D case, PVN3D He et al. (2020) votes 3D keypoints, and can be regarded as a variation of PVNet Peng et al. (2019) in 3D domain. YOLOff Gonzalez et al. (2020) utilizes a classification CNN to estimate the object's 2D location in the image from local patches, followed by a regression CNN trained to predict the 3D location of a set of keypoints in the camera coordinate system. The 6D object pose is then achieved by minimizing a registration error. 6-PACK Wang et al. (2019) predicts

Table 6 Summary of voting-based 6D object pose estimation methods

Methods	Descriptions	2D image-based methods	3D point cloud-based methods
Indirect voting methods	Voting for correspondence-based methods	PVNet Peng et al. (2019), Yu et al. (2020)	PVN3D He et al. (2020), YOLOff Gonzalez et al. (2020), 6-PACK Wang et al. (2019)
Direct voting methods	Voting for template-based methods	Brachmann et al. (2014), Tejani et al. (2014), Crivellaro et al. (2017), PPF Drost and Ilic (2012)	DenseFusion Wang et al. (2019), MoreFusion Wada et al. (2020)

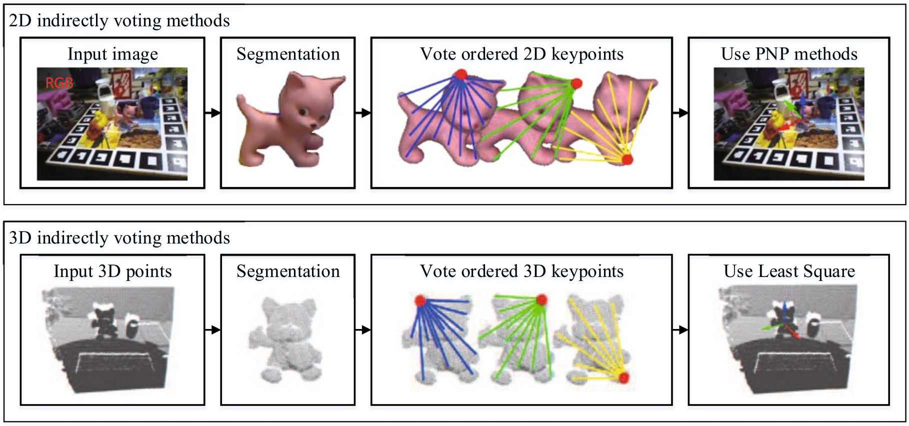
2D indirectly voting methods
Input image
RGB
Segmentation
Vote ordered 2D keypoints
Use PNP methods
3D indirectly voting methods
Input 3D points
Segmentation
Vote ordered 3D keypoints
Use Least Square

Fig. 15 Typical functional flow-chart of indirect voting-based object pose estimation methods

a handful of ordered 3D keypoints for an object based on the observation that inter-frame motion of an object instance can be estimated through keypoint matching. This method achieves category-level 6D object pose tracking on RGB-D data.

3.3.2 Direct voting methods

This kind of methods can be regarded as voting for template-based methods if we treat the voted object pose or object coordinate as the most similar template. Typical functional flow-chart of direct voting-based 6D object pose estimation methods is illustrated in Fig. 16.

In 2D case, this kind of methods are mainly used for computing the poses of objects with occlusions. For these objects, the local evidence in the image restricts the possible outcome of the desired output, and every image patch is thus usually used to cast a vote

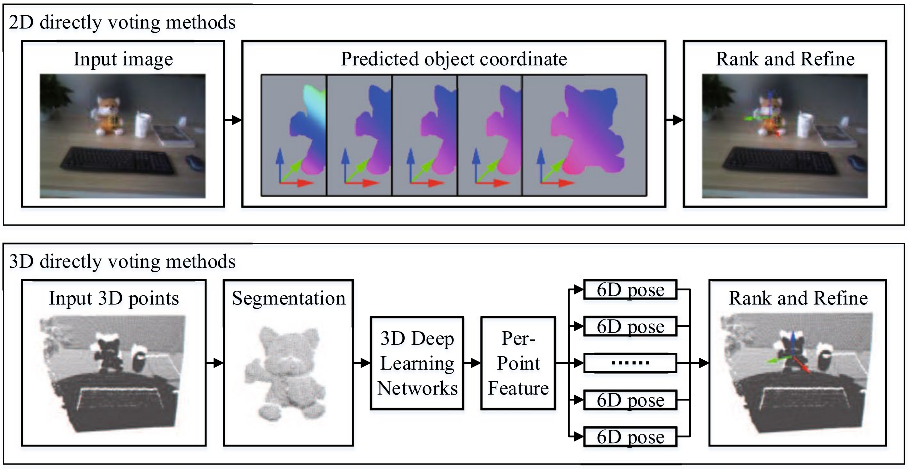

Fig. 16 Typical functional flow-chart of direct voting-based 6D object pose estimation methods

about the 6D object pose. Brachmann et al. (2014) proposed a learned, intermediate representation in the form of a dense 3D object coordinate labelling paired with a dense class labelling. Each object coordinate prediction defines a 3D-3D correspondence between the image and the 3D object model, and the pose hypotheses are generated and refined to obtain the final hypothesis. Tejani et al. (2014) trained a Hough forest for 6D pose estimation from an RGB-D image. Each tree in the forest maps an image patch to a leaf which stores a set of 6D pose votes.

In 3D case, Drost et al. (2010) proposed the Point Pair Features (PPF) to recover the 6D pose of objects from a depth image. A point pair feature contains information about the distance and normals of two arbitrary 3D points. PPF has been one of the most successful 6D pose estimation method as an efficient and integrated alternative to the traditional local and global pipelines. Hodan et al. (2018) proposed a benchmark for 6D pose estimation of a rigid object from a single RGB-D input image, and a variation of PPF Vidal et al. (2018) won the 2018 SIXD challenge.

Deep learning-based methods also assist the directly voting methods. DenseFusion Wang et al. (2019) utilizes a heterogeneous architecture that processes the RGB and depth data independently and extracts pixel-wise dense feature embeddings. Each feature embedding votes a 6D object pose and the best prediction is adopted. They further proposed an iterative pose refinement procedure to refine the predicted 6D object pose. MoreFusion Wada et al. (2020) conducts an object-level volumetric fusion and performs point-wise volumetric pose prediction that exploits volumetric reconstruction and CNN feature extraction from the image observation. The object poses are then jointly refined based on geometric consistency among objects and impenetrable space.

3.4 Comparisons and discussions

In this section, we mainly review the methods based on the RGB-D image, since 3D point cloud-based 6D object pose estimation could be regarded as a registration or alignment problem where some surveys (Tam et al. 2013; Bellekens et al. 2014) exist. The related datasets, evaluation metrics and comparisons are presented.

3.4.1 Datasets and evaluation metrics

There exist various benchmarks Hodaň et al. (2018) for 6D pose estimation, such as LineMod Hinterstoisser et al. (2012), IC-MI/IC-BIN dataset Tejani et al. (2014), T-LESS dataset Hodaň et al. (2017), RU-APC dataset Rennie et al. (2016), and YCB-Video Xiang et al. (2018), etc. Here we only reviewed the most widely used LineMod Hinterstoisser et al. (2012) dataset and YCB-Video Xiang et al. (2018) dataset. LineMod Hinterstoisser et al. (2012) provides manual annotations for around 1,000 images for each of the 15 objects in the dataset. Occlusion Linemod Brachmann et al. (2014) contains more examples where the objects are under occlusion. YCB-Video Xiang et al. (2018) contains a subset of 21 objects and comprises 133,827 images. These datasets are widely evaluated aiming at various kinds of methods.

The 6D object pose can be represented by a 4×4 matrix $\mathrm{P} = [R, t; 0, 1]$, where R is a 3×3 rotation matrix and t is a 3×1 translation vector. The rotation matrix could also be represented as quaternions or angle-axis representation. Direct comparison of the variances between the values can not provide intuitive visual understandings. The commonly used metrics are the Average Distance of Model Points (ADD) Hinterstoisser et al. (2012)

for non-symmetric objects and the average closest point distances (ADD-S) Xiang et al. (2018) for symmetric objects.

Given a 3D model M, the ground truth rotation R and translation T, and the estimated rotation $\hat{R}$ and translation $\hat{T}$, ADD means the average distance of all model points x from their transformed versions. The 6D object pose is considered to be correct if the average distance is smaller than a predefined threshold.

$$e_{ADD} = \underset{x \in M}{avg} \left\| (Rx + T) - (\hat{R}x + \hat{T}) \right\|. \tag{1}$$

ADD-S Xiang et al. (2018) is an ambiguity-invariant pose error metric which takes both symmetric and non-symmetric objects into an overall evaluation. Given the estimated pose [R̂|T̂] and the ground truth pose [R|T], ADD-S calculates the mean distance from each 3D model point transformed by [R̂|T̂] to its closest point on the target model transformed by [R|T].

Aim at the LineMOD dataset, ADD is used for asymmetric objects and ADD-S is used for symmetric objects. The threshold is usually set as 10% of the model diameter. Aiming at the YCB-Video dataset, the commonly used evaluation metric is the ADD-S metric. The percentage of ADD-S smaller than 2cm (< 2cm) is often used, which measures the predictions under the minimum tolerance for robotic manipulation. In addition, the area under the ADD-S curve (AUC) following PoseCNN Xiang et al. (2018) is also reported and the maximum threshold of AUC is set to be 10cm.

3.4.2 Comparisons and discussions

6D object pose estimation plays a pivotal role in robotic and augment reality areas. Various methods exist with different inputs, precision, speed, advantages and disadvantages. Aiming at robotic grasping tasks, the practical environment, the available input data, the available hardware setup, the target objects to be grasped, the task requirements should be analyzed first to decide which kinds of methods to use. The above mentioned three kinds of methods deal with different kinds of objects. Generally, when the target object has rich texture or geometric details, the correspondence-based method is a good choice. When the target object has weak texture or geometric detail, the template-based method is a good choice. When the object is occluded and only partial surface is visible, or the addressed object ranges from specific objects to category-level objects, the voting-based method is a good choice. Besides, the three kinds of methods all have 2D inputs, 3D inputs or mixed inputs. The results of methods with RGB-D images as inputs are summarized in Table 7 on the YCB-Video dataset, and Table 8 on the LineMOD and Occlusion LineMOD datasets. All recent methods on LineMOD achieve high accuracy since there's little occlusion. When there exist occlusions, correspondence-based and voting-based methods perform better than template-based methods. The template-based methods are more like a direct regression problem, which highly depend on the global feature extracted. Whereas, correspondence-based and voting-based methods utilize the local parts information and constitute local feature representations.

There exist some challenges for nowadays 6D object pose estimation methods. The first challenge lies in that current methods show obvious limitations in cluttered scenes in which occlusions usually occur. Although the state-of-the-art methods achieve high accuracies on the Occlusion LineMOD dataset, they still could not afford severe occluded cases since this situation may cause ambiguities even for human beings. The second one is the lack

Table 7 Accuracies of AUC and ADD-S metrics on YCB-video dataset

Category	Method	AUC	ADD-S (< 2cm)
Corre-based	Heatmaps Oberweger et al. (2018)	72.8	53.1
Template-based	PoseCNN Xiang et al. (2018) + ICP	61.0	73.8
	PoseCNN Xiang et al. (2018) + ICP	93.0	93.2
	Castro et al. (2020)	67.52	47.09
	PointFusion Xu et al. (2018)	83.9	74.1
	MaskedFusion Pereira and Alexandre (2019)	93.3	97.1
Voting-based	DenseFusion Wang et al. (2019) (per-pixel)	91.2	95.3
	DenseFusion Wang et al. (2019) (iterative)	93.1	96.8

of sufficient training data, as the sizes of the datasets presented above are relatively small. Nowadays deep learning methods show poor performance on objects which do not exist in the training datasets and perhaps the simulated datasets could be one solution. Although some category-level 6D object pose methods (Wang et al. 2019; Park et al. 2020; Chen et al. 2020) emerged recently, they still can not handle large number of categories.

4 Grasp estimation

Grasp estimation means estimating the 6D gripper pose in the camera coordinate. As mentioned before, the grasp can be categorized into 2D planar grasp and 6DoF grasp. For 2D planar grasp, where the grasp is constrained from one direction, the 6D gripper pose could be simplified into a 3D representation, which includes the 2D in-plane position and 1D rotation angle, since the height and the rotations along other axes are fixed. For 6DoF grasp, the gripper can grasp the object from various angles and the 6D gripper pose is essential to conduct the grasp. In this section, methods of 2D planar grasp and 6DoF grasp are presented in detail.

4.1 2D planar grasp

Methods of 2D planar grasp can be divided into methods of evaluating grasp contact points and methods of evaluating oriented rectangles. In 2D planar grasp, the grasp contact points can uniquely define the gripper's grasp pose, which is not the situation in 6DoF grasp. The 2D oriented rectangles can also uniquely define the gripper's grasp pose. These methods are summarized in Table 9 and typical functional flow-chart is illustrated in Fig. 17.

4.1.1 Methods of evaluating grasp contact points

This kind of methods first sample candidate grasp contact points, and use analytical methods or deep learning-based methods to evaluate the possibility of a successful grasp, which are classification-based methods. Empirical methods of robotic grasping are performed based on the premise that certain prior knowledge, such as object geometry, physics models, or force analytic, are known. The grasp database usually covers a limited amount of objects, and empirical methods will face difficulties in dealing with unknown objects. Domae et al. (2014) presented a method that estimates graspability

Table 8 Accuracies of methods using ADD(-S) metric on LineMOD and Occlusion LineMOD dataset

Category	Method	LineMOD	Occlusion
Correspondence-based methods	BB8 Rad and Lepetit (2017)	43.6	–
	BB8 Rad and Lepetit (2017) + Refine	62.7	–
	Tekin et al. (2018)	55.95	6.42
	Heatmaps Oberweger et al. (2018)	–	25.8
	Heatmaps Oberweger et al. (2018) + Refine	–	30.4
	Hu et al. (2019)	–	26.1
	Pix2pose Park et al. (2019)	72.4	32.0
	DPOD Zakharov et al. (2019)	82.98	32.79
	DPOD Zakharov et al. (2019) + Refine	95.15	47.25
	HybridPose Song et al. (2020)	94.5	79.2
Template-based methods	SSD-6D Kehl et al. (2017)	2.42	–
	SSD-6D Kehl et al. (2017) + Refine	76.7	27.5
	AAE Sundermeyer et al. (2018)	31.41	–
	AAE Sundermeyer et al. (2018) + Refine	64.7	–
	Castro et al. (2020)	59.32	–
	PoseCNN Xiang et al. (2018)	62.7	6.42
	PoseCNN Xiang et al. (2018) + Refine	88.6	78.0
	CDPN Li et al. (2019)	89.86	–
	Tian et al. (2020)	92.87	–
	MaskedFusion Pereira and Alexandre (2019)	97.3	–
Voting-based methods	Brachmann et al. (2016)	32.3	–
	Brachmann et al. (2016) + Refine	50.2	–
	PVNet Peng et al. (2019)	86.27	40.8
	DenseFusion Wang et al. (2019)(per-pixel)	86.2	
	DenseFusion Wang et al. (2019)(iterative)	94.3	
	DPVL Yu et al. (2020)	91.5	43.52
	YOLOff Gonzalez et al. (2020)	94.2	–
	YOLOff Gonzalez et al. (2020) + Refine	98.1	–
	PVN3D He et al. (2020)	95.1	-
	P^{2}GNet Yu et al. (2019)	96.2	–
	P^{2}GNet Yu et al. (2019) + Refine	97.4	–
	PointPoseNet Hagelskjær and Buch (2019)	96.3	52.6
	PointPoseNet Hagelskjær and Buch (2019) + Refine	–	75.1

Refine means methods such as ICP or DeepIM Li et al. (2018). IR is short for iterative refinement

measures on a single depth map for grasping objects randomly placed in a bin. Candidate grasp regions are first extracted and the graspability is computed by convolving one contact region mask image and one collision region mask image. Deep learning-based methods could assists in evaluating the grasp qualities of candidate grasp contact points. Mahler et al. (2017) proposed DexNet 2.0, which plans robust grasps with synthetic point clouds and analytic grasping metrics. They first segment the current points of interests from the depth image, and multiple candidate grasps are generated. The

Table 9 Summary of 2D planar grasp estimation methods

Methods	Traditional methods	Deep learning-based methods
Methods of evaluating grasp contact points	Domae et al. (2014)	Zeng et al. (2018), Mahler et al. (2017), Cai et al. (2019), GG-CNN Morrison et al. (2018), MVP Morrison et al. (2019), Wang et al. (2019)
Methods of evaluating oriented rectangles	Jiang et al. (2011), Vohra et al. (2019)	Lenz et al. (2015), Pinto and Gupta (2016), Park and Chun (2018), Redmon and Angelova (2015), Zhang et al. (2017), Kumra and Kanan (2017), Kumra et al. (2019), Zhang et al. (2018), Guo et al. (2017), Chu et al. (2018), Park et al. (2018), Zhou et al. (2018), Depierre et al. (2020)

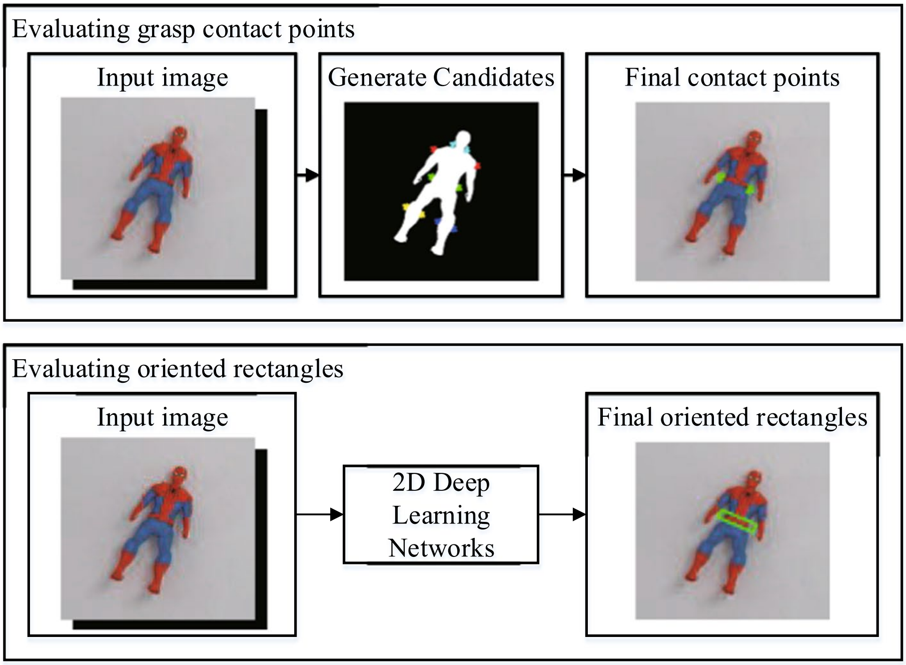
Evaluating grasp contact points
Input image
Generate Candidates
Final contact points
Evaluating oriented rectangles
Input image
2D Deep Learning Networks
Final oriented rectangles

Fig. 17 Typical functional flow-chart of 2D planar grasp methods. Data from the JACQUARD dataset Depierre et al. (2018)

grasp qualities are then measured using the Grasp Quality-CNN network, and the one with the highest quality will be selected as the final grasp. Their database have more than 50k grasps, and the grasp quality measurement network achieved relatively satisfactory performance.

Deep learning-based methods could also assist in estimating the most probable grasp contact points through estimating pixel-wise grasp affordances. Robotic affordances (Do et al. 2018; Ardón et al. 2019; Chu et al. 2019) usually aim to predict affordances of the object parts for robot manipulation, which are more like a segmentation problem. However, there exist some methods (Zeng et al. 2018; Cai et al. 2019) that predict pixel-wise affordances with respect to the grasping primitive actions. These methods generate grasp qualities for each pixel, and the pair of points with the highest affordance value is executed. Zeng et al. (2018) proposed a method which infers dense pixel-wise probability maps of the affordances for four different grasping primitive actions through utilizing fully convolutional networks. Cai et al. (2019) presented a pixel-level affordance interpreter network, which learns antipodal grasp patterns based on a fully convolutional residual network similar with Zeng et al. (2018). Both of these two methods do not segment the target object and predict pixel-wise affordance maps for each pixels. This is a way which directly estimate grasp qualities without sampling grasp candidates. Morrison et al. (2018) proposed the Generative Grasping Convolutional Neural Network (GG-CNN), which predicts the quality and pose of grasps at every pixel. Further, Morrison et al. (2019) proposed a Multi-View Picking (MVP) controller, which uses an active perception approach to choose informative viewpoints based on a distribution of grasp pose estimates. They utilized the real-time GG-CNN Morrison et al. (2018) for visual grasp detection. Wang et al. (2019) proposed a fully convolution neural network which encodes the origin input images to features and decodes these features to generate robotic grasp properties for each pixel. Unlike classification-based methods for generating multiple grasp candidates through neural network, their pixel-wise implementation directly predicts multiple grasp candidates through one forward propagation.

4.1.2 Methods of evaluating oriented rectangles

Jiang et al. (2011) first proposed to use an oriented rectangle to represent the gripper configuration and they utilized a two-step procedure, which first prunes the search space using certain features that are fast to compute and then uses advanced features to accurately select a good grasp. Vohra et al. (2019) proposed a grasp estimation strategy which estimates the object contour in the point cloud and predicts the grasp pose along with the object skeleton in the image plane. Grasp rectangles at each skeleton point are estimated, and point cloud data corresponding to the grasp rectangle part and the centroid of the object is used to decide the final grasp rectangle. Their method is simple and needs no grasp configuration sampling steps.

Aiming at the oriented rectangle-based grasp configuration, deep learning methods are gradually applied in three different ways, which are classification-based methods, regression-based methods and detection-based methods. Most of these methods utilize a five dimensional representation Lenz et al. (2015) for robotic grasps, which are rectangles with a position, orientation and size: (x,y,θ,h,w).

Classification-based methods train classifiers to evaluate candidate grasps, and the one with the highest score will be selected. Lenz et al. (2015) is the first to apply deep learning methods to robotic grasping. They presented a two-step cascaded system with two deep networks, where the top detection results from the first are re-evaluated by the second. The first network produces a small set of oriented rectangles as candidate grasps, which will be axis aligned. The second network ranks these candidates using features extracted from the color image, the depth image and surface normals. The top-ranked rectangle is selected and the corresponding grasp is executed. Pinto and Gupta (2016) predicted grasping locations by sampling image patches and predicting the grasping angle. They trained a CNN-based classifier to estimate the grasp likelihood for different grasp directions given an input image patch. Park and Chun (2018) proposed a classification based robotic grasp detection method with multiple-stage spatial transformer networks (STN). Their method allows partial observation for intermediate results such as grasp location and orientation for a number of grasp configuration candidates. The procedure of classification-based methods is straightforward, and the accuracy is relatively high. However, these methods tend to be quite slow.

Regression-based methods train a model to yield grasp parameters for location and orientation directly, since a uniform network would perform better than the two-cascaded system Lenz et al. (2015). Redmon and Angelova (2015) proposed a larger neural network, which performs a single-stage regression to obtain graspable bounding boxes without using standard sliding window or region proposal techniques. Zhang et al. (2017) utilized a multi-modal fusion architecture which combines RGB features and depth features to improve the grasp detection accuracy. Kumra and Kanan (2017) utilized deep neural networks like ResNet He et al. (2016) and further increased the performances in grasp detection. Kumra et al. (2019) proposed a novel Generative Residual Convolutional Neural Network (GR-ConvNet) model that can generate robust antipodal grasps from a n-channel image input. Rather than regressing the grasp parameters globally, some methods utilized a ROI (Region of Interest)-based or pixel-wise way. Zhang et al. (2018) utilized ROIs in the input image and regressed the grasp parameters based on ROI features.

Detection-based methods utilize the reference anchor box, which are used in some deep learning-based object detection algorithms (Ren et al. 2015; Liu et al. 2016; Redmon et al. 2016), to assist the generation and evaluation of candidate grasps. With the prior knowledge on the size of the expected grasps, the regression problem is simplified Depierre et al. (2020). Guo et al. (2017) presented a hybrid deep architecture combining the visual and tactile sensing. They introduced the reference box which is axis aligned. Their network produces a quality score and an orientation as classification between discrete angle values. Chu et al. (2018) proposed an architecture that predicts multiple candidate grasps instead of a single outcome and transforms the orientation regression to a classification task. The orientation classification contains the quality score and therefore their network predicts both grasp regression values and discrete orientation classification score. Park et al. (2018) proposed a rotation ensemble module (REM) for robotic grasp detection using convolutions that rotates network weights. Zhou et al. (2018) designed an oriented anchor box mechanism to improve the accuracy of grasp detection and employed an end-to-end fully convolutional neural network. They utilized only one anchor box with multiple orientations, rather than multiple scales or aspect ratios (Guo et al. 2017; Chu et al. 2018) for reference grasps, and predicted five regression values and one grasp quality score for each oriented reference box. Depierre et al. (2020) further extends Zhou et al. (2018) by adding a direct dependency between the regression prediction and the score evaluation. They proposed a novel DNN architecture with a scorer which evaluates the graspability of a given position and introduced a novel loss function which correlates the regression of grasp parameters with the graspability score.

Some other methods are also proposed aiming at cluttered scenes, where a robot need to know if an object is on another object in the piles of objects for a successful grasp. Guo et al. (2016) presented a shared convolutional neural network to conduct object discovery and grasp detection. Zhang et al. (2018) proposed a multi-task convolution robotic grasping network to address the problem of combining grasp detection and object detection with relationship reasoning in the piles of objects. The method of Zhang et al. (2018) consists of several deep neural networks that are responsible for generating local feature maps, grasp detection, object detection and relationship reasoning separately. In comparison, Park et al. (2019) proposed a single multi-task deep neural networks that yields the information on grasp detection, object detection and relationship reasoning among objects with a simple post-processing.

4.1.3 Comparisons and discussions

The methods of 2D planar grasp are evaluated in this section, which contain the datasets, evaluation metrics and comparisons of the recent methods.

Datasets and evaluation metrics There exist a few datasets for 2D planar grasp, which are presented in Table 10. Among them, the Cornell Grasping dataset Jiang et al. (2011) is the most widely used dataset. In addition, the dataset has the image-wise splitting and the object-wise splitting. Image-wise splitting splits images randomly and is used to test how well the method can generalize to new positions for objects it has seen previously. Object-wise splitting puts all images of the same object into the same cross-validation split and is used to test how well the method can generalize to novel objects.

Aiming at the point-based grasps and the oriented rectangle-based grasps Jiang et al. (2011), there exist two metrics for evaluating the performance of grasp detection: the point metric and the rectangle metric. The former evaluates the distance between predicted grasp center and the ground truth grasp center w.r.t. a threshold value. It has difficulties in determining the distance threshold and does not consider the grasp angle. The latter metric considers a grasp to be correct if the grasp angle is within 30° of the ground truth grasp, and the Jaccard index $J(A, B) = |A \cap B| \big/ |A \cup B|$ of the predicted grasp A and the ground truth B is greater than 25%.

Comparisons The methods of evaluating oriented rectangles are compared in Table 11 on the widely used Cornell Grasping dataset Jiang et al. (2011). From the table, we can see that the state-of-the-art methods have achieved very high accuracies on this dataset. Recent works Depierre et al. (2020) began to conduct experiments on the Jacquard Grasp dataset Depierre et al. (2018) since it has more images and the grasps are more diverse.

4.2 6DoF Grasp

Methods of 6DoF grasp can be divided into methods based on the partial point cloud and methods based on the complete shape. These methods are summarized in Table 12.

Table 10 Summaries of publicly available 2D planar grasp datasets

Dataset	Objects num	Num of RGB-D images	Num of grasps
Stanford Grasping (Saxena et al. 2008a, b)	10	13747	13747
Cornell Grasping Jiang et al. (2011)	240	885	8019
CMU dataset Pinto and Gupta (2016)	Over 150	50567	No
Dex-Net 2.0 Mahler et al. (2017)	Over 150	6.7 M(Depth only)	6.7 M
JACQUARD Depierre et al. (2018)	11619	54485	1.1 M

Table 11 Accuracies of grasp prediction on the Cornell Grasp dataset

Method	Input size	Accuracy(%)		Time
		Image split	Object split	
Jiang et al. (2011)	227 × 227	60.50	58.30	50 s
Lenz et al. (2015)	227 × 227	73.90	75.60	13.5 s
Morrison et al. (2018)	300 × 300	78.56	–	7 ms
Redmon and Angelova (2015)	224 × 224	88.00	87.1	76 ms
Zhang et al. (2017)	224 × 224	88.90	88.20	117 ms
Kumra and Kanan (2017)	224 × 224	89.21	88.96	103 ms
Park and Chun (2018)	400 × 400	89.60	–	23 ms
Asif et al. (2018)	224 × 224	90.60	90.20	24 ms
Wang et al. (2019)	400 × 400	94.42	91.02	8 ms
Chu et al. (2018)	227 × 227	96.00	96.10	120 ms
Park et al. (2018)	360 × 360	96.60	95.40	20 ms
Zhou et al. (2018)	320 × 320	97.74	96.61	118 ms
Park et al. (2019)	360 × 360	98.6	97.2	16 ms

4.2.1 Methods based on the partial point cloud

This kind of methods can be divided into two kinds. The first kind of methods estimate grasp qualities of candidate grasps and the second kind of methods transfer grasps from existing ones. Typical functional flow-chart of methods based on the partial point cloud is illustrated in Fig. 18.

Methods of estimating grasp qualities of candidate grasps This kind of methods mean that the 6DoF grasping pose is estimated through analyzing the input partial point cloud merely. Most of this kind of methods (Bohg and Kragic 2010; Pas and Platt 2015; Zapata-Impata et al. 2019; ten Pas et al. 2017; Liang et al. 2019) sample large number of candidate grasps first, and then utilize various methods to evaluate grasp qualities, which is a classification-based manner. While some novel methods (Qin et al. 2020; Zhao and Nanning 2020; Ni et al. 2020; Mousavian et al. 2019) estimate the grasp qualities implicitly and directly predict the 6DoF grasp pose in a single-shot way, which is a regression-based manner.

Bohg and Kragic (2010) applied the concept of shape context Belongie et al. (2002) to improve the performance of grasping point classification. They used a supervised learning approach and the classifier is trained with labelled synthetic images. Pas and Platt (2015) first used a geometrically necessary condition to sample a large set of high quality grasp hypotheses, which will be classified using the notion of an antipodal grasp. Zapata-Impata et al. (2019) proposed a method to find the best pair of grasping points given a partial single-view point cloud of an unknown object. They defined an improved version of the ranking metric Zapata-Impata et al. (2017) for evaluating a pair of contact points, which is parameterized by the morphology of the robotic hand in use.

3D data has different representations such as multi-view images, voxel grids or point cloud, and each representation can be processed with corresponding deep neural networks. These different kinds of neural networks have already been applied into robotic grasping tasks. GPD ten Pas et al. (2017) generates candidate grasps on the a region of interest (ROI) first. These candidate grasps are then encoded into a stacked multi-channel

Table 12 Summary of 6DoF grasp estimation methods

Methods	Descriptions	Traditional methods	Deep learning-based methods
Methods based on the partial point cloud	Estimate grasp qualities of candidate grasps	Bohg and Kragic (2010), Pas and Platt (2015), Zapata-Impata et al. (2019)	GPD ten Pas et al. (2017), PointnetGPD Liang et al. (2019), 6-DoF GraspNet Mousavian et al. (2019), S^4G Qin et al. (2020), REGNet Zhao and Nanning (2020)
	Transfer grasps from existing ones	Miller et al. (2003), Nikandrova and Kyrki Nikandrova and Kyrki (2015), Vahrenkamp et al. (2016)	Tian et al. (2019), Dense Object Nets Florence et al. (2018), DGCM-Net Patten et al. (2020)
Methods based on the complete shape	Estimate the 6D object pose	Zeng et al. (2017b)	Billings and Johnson-Roberson (2018)
	Conduct shape completion	Miller et al. (2003)	Varley et al. (2017), Lundell et al. (2019), Watkins-Valls et al. (2019), Van der Merwe et al. (2019), Wang et al. (2018), Yan et al. (2018), cite2019DataEfficientLearning, Tosun et al. (2020), kPAM-SC Gao and Tedrake (2019), ClearGrasp Sajjan et al. (2019)

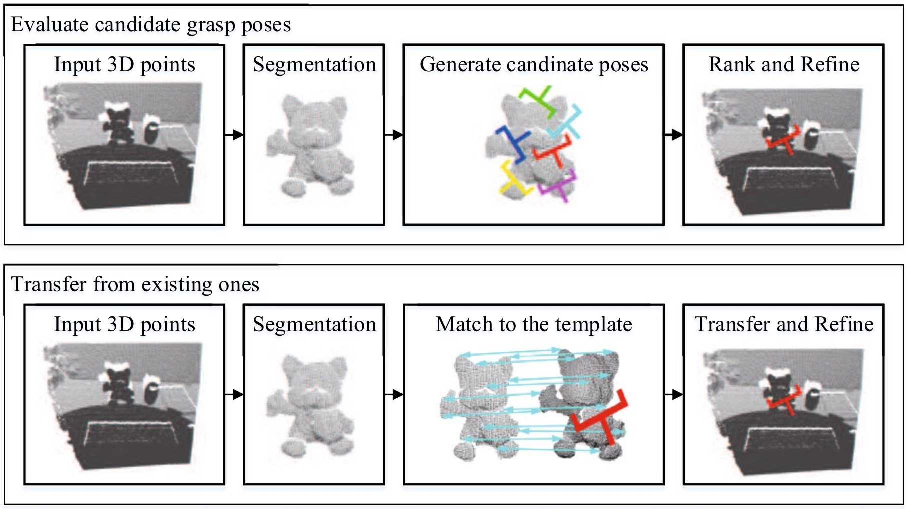

Fig. 18 Typical functional flow-chart of 6DoF grasp methods based on the partial point cloud

image. Each candidate is evaluated to obtain a score using a four-layer convolutional neural network finally. Lou et al. (2019) proposed an algorithm that uniformly samples over the entire 3D space first to generate candidate grasps, predicts the grasp stability using 3D CNN together with a grasping reachability using the candidate grasp pose, and obtains the final grasping success probability. PointnetGPD Liang et al. (2019) randomly samples candidate grasps, and evaluates the grasp quality by direct point cloud analysis with the 3D deep neural network PointNet Qi et al. (2017). During the generation of training datasets, the grasp quality is evaluated through combining the force-closure metric and the Grasp Wrench Space (GWS) analysis Kirkpatrick et al. (1992). Mousavian et al. (2019) proposed an algorithm called 6-DoF GraspNet, which samples grasp proposals using a variational auto-encoder and refines the sampled grasps using a grasp evaluator model. Pointnet++ Qi et al. (2017) is used to generate and evaluate grasps. Murali et al. (2019) further improved 6-DoF GraspNet by introducing a learned collision checker conditioned on the gripper information and on the raw point cloud of the scene, which affords a higher success rate in cluttered scenes.

Qin et al. (2020) presented an algorithm called S^4G, which utilizes a single-shot grasp proposal network trained with synthetic data using Pointnet++ Qi et al. (2017) and predicts amodal grasp proposals efficiently and effectively. Each grasp proposal is further evaluated with a robustness score. The core novel insight of S^4G is that they learn to propose possible grasps by regression, rather than using a sliding windows-like style. S^4G generates grasp proposals directly, while 6-DoF GraspNet uses an encode and decode way. Ni et al. (2020) proposed Pointnet++Grasping, which is also an end-to-end approach to directly predict the poses, categories and scores of all the grasps. Further, Zhao et al. (2020) proposed an end-to-end single-shot grasp detection network called REGNet, which takes one single-view point cloud as input for parallel grippers. There network contains three stages, which are the Score Network (SN) to select positive points with high grasp confidence, the Grasp Region Network (GRN) to generate a set of grasp proposals on selected positive points, and the Refine Network (RN) to refine the detected grasps based on local grasp features. REGNet is the state-of-the-art method for grasp detection in 3D space and outperforms several methods including GPD ten Pas et al. (2017), PointnetGPD Liang et al. (2019) and S^{4}G Qin et al. (2020). Fang et al. (2020) proposed a large-scale grasp pose detection dataset called GraspNet-1Billion, which contains 97,280 RGB-D image with over one billion grasp poses. They also proposed an end-to-end grasp pose prediction network that learns approaching direction and operation parameters in a decoupled manner.

Methods of transferring grasps from existing ones This kind of methods transfer grasps from existing ones, which means finding correspondences from the observed single-view point cloud to the existing complete one if we know that they come from one category. In most cases, target objects are not totally the same with the objects in the existing database. If an object comes from a class that is involved in the database, it is regarded as a similar object. After the localization of the target object, correspondence-based methods can be utilized to transfer the grasp points from the similar and complete 3D object to the current partial-view object. These methods learn grasps by observing the object without estimating its 6D pose, since the current target object is not totally the same with the objects in the database.

Different kinds of methods are utilized to find the correspondences based on taxonomy, segmentation, and so on. Miller et al. (2003) proposed a taxonomy-based approach, which classified objects into categories that should be grasped by each canonical grasp. Nikandrova and Kyrki (2015) presented a probabilistic approach for task-specific stable grasping of objects with shape variations inside the category. An optimal grasp is found as a grasp that is maximally likely to be task compatible and stable taking into account shape uncertainty in a probabilistic context. Their method requires partial models of new objects, and few models and example grasps are used during the training. Vahrenkamp et al. (2016) presented a part-based grasp planning approach to generate grasps that are applicable to multiple familiar objects. The object models are segmented according to their shape and volumetric information, and the objet parts are labeled with semantic and grasping information. A grasp transferability measure is proposed to evaluate how successful planned grasps are applied to novel object instances of the same object category. Tian et al. (2019) proposed a method to transfer grasp configurations from prior example objects to novel objets, which assumes that the novel and example objects have the same topology and similar shapes. They perform 3D segmentation on the objects considering geometric and semantic shape characteristics, compute a grasp space for each part of the example object using active learning, and build bijective contact mappings between the model parts and the corresponding grasps for novel objects. Florence et al. (2018) proposed Dense Object Nets, which is built on self-supervised dense descriptor learning and takes dense descriptors as a representation for robotic manipulation. They could grasp specific points on objects across potentially deformed configurations, grasp objects with instance-specificity in clutter, or transfer specific grasps across objects in class. Patten et al. (2020) presented DGCM-Net, a dense geometrical correspondence matching network for incremental experience-based robotic grasping. They apply metric learning to encode objects with similar geometry nearby in feature space, and retrieve relevant experience for an unseen object through a nearest neighbour search. DGCM-Net also reconstructs 3D-3D correspondences using the view-dependent normalized object coordinate space to transform grasp configurations from retrieved samples to unseen objects. Their method could be extended for semantic grasping by guiding grasp selection to the parts of objects that are relevant to the object's functional use.

Comparisons and discussions Methods of estimating grasp qualities of candidate grasps gain much attentions since this is the direct manner to obtain the 6D grasp pose. Aiming at 6DoF grasp, the evaluation metrics for 2D planar grasp are not suitable. The commonly used metric is the Valid Grasp Ratio (VGR) proposed by REGNet Zhao and Nanning (2020). VGR is defined as the quotient of antipodal and collision-free grasps and all grasps. The usually used grasp dataset for evaluation is the YCB-Video Xiang et al. (2018) dataset. Comparisons with recent methods are shown in Table 13.

Methods of transferring grasps from existing ones have potential usages in high-level robotic manipulation tasks. Not only the grasps could be transferred, the manipulation skills could also be transferred. Lots of methods (Berscheid et al. 2019; Yang et al. 2019) that learn grasps from demonstration usually utilize this kind of methods.

4.2.2 Methods based on the complete shape

Methods based on the partial point cloud are suitable for unknown objects, since these methods have no identical 3D models to use. Aiming at known objects, their 6D poses can be estimated and the 6DoF grasp poses estimated on the complete 3D shape could be transformed from the object coordinate to the camera coordinate. In another perspective, the 3D complete object shape under the camera coordinate could also be completed from the observed single-view point cloud. And the 6DoF grasp poses could be estimated based on the completed 3D object shape in the camera coordinate. We consider these two kinds of methods as complete shape-based methods since 6DoF grasp poses are estimated based on complete object shapes. Typical functional flow-chart of 6DoF grasp methods based on the complete shape is illustrated in Fig. 19.

Methods of estimating the 6D object pose The 6D object pose could be accurately estimated from the RGB-D data if the target object in known as mentioned in Sect. 3, and 6DoF grasp poses can be obtained via offline pre-computation or online generation. This is the most popular method used for the grasping systems. If the 6DoF grasp poses exist in the database, the current 6DoF grasp pose could be retrieved from the knowledge base, or obtained by sampling and ranking them through comparisons with existing grasps. If the 6DoF grasp poses do not exist in the database, analytical methods are utilized to compute the grasp poses. Analytical methods consider kinematics and dynamics formulation in determining grasps Sahbani et al. (2012). Force-closure is one of the main conditions in completing the grasping tasks and there exist many force-closure grasp synthesis methods for 3D objects. Among them, the polyhedral objects are first dealt with, as they are composed of a finite number of flat faces. The force-closure condition is reduced into the test of the angles between the faces normals Nguyen (1987) or using the linear model to derive analytical formulation for grasp characterization Ponce et al. (1993). To handle the commonly used objects which usually have more complicated shapes, methods of observing different contact points are proposed Ding et al. (2001). These methods try to find contact points on a 3D object surface to ensure force-closure and compute the optimal grasp by minimizing an objective energy function according to a predefined grasp quality

Table 13 Accuracies of grasp prediction on the Cornell Grasp dataset

Method	VGR (%)	Time (ms)
GPD ten Pas et al. (2017) (3 channels)	79.34	2077.12
GPD ten Pas et al. (2017) (12 channels)	80.22	2702.38
PointNetGPD Liang et al. (2019)	81.48	1965.60
S^4G Qin et al. (2020)	77.63	679.04
REGNet Zhao and Nanning (2020)	92.47	686.31

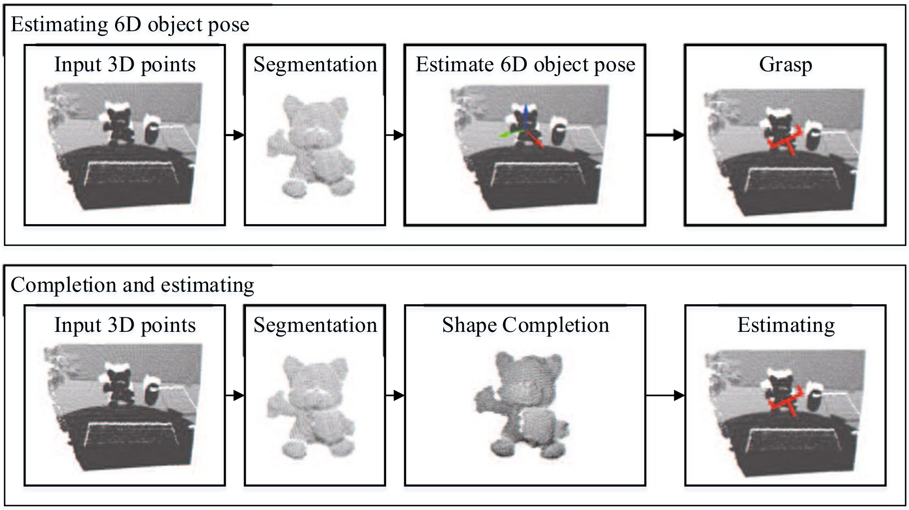

Fig. 19 Typical functional flow-chart of 6DoF grasp methods based on the complete shape

criterion Mirtich and Canny (1994). However, searching the grasp solution space is a complex problem which is quite time-consuming. Some heuristical techniques were then proposed to reduce the search space by generating a set of grasp candidates according to a predefined procedure Borst et al. (2003), or by defining a set of rules to generate the starting positions Miller et al. (2003). A few robotic grasping simulators, such as GraspIt! Miller and Allen (2004), assist the generation of the best gripper pose to conduct a successful grasp. Miller and Allen (2004) proposed GraspIt!, which is a versatile simulator for robotic grasping. GraspIt! supports the loading of objects and obstacles of arbitrary geometry to populate a complete simulation world. It allows a user to interactively manipulate a robot or an object and create contacts between them. Xue et al. (2009) implemented a grasping planning system based on GraspIt! to plan high-quality grasps. Leόn et al. (2010) presented OpenGRASP, a toolkit for simulating grasping and dexterous manipulation. It provides a holistic environment that can deal with a variety of factors associated with robotic grasping. These methods produce successful grasps and detailed reviews could be found in the survey Sahbani et al. (2012).

Both traditional and deep learning-based 6D object pose estimation algorithms are utilized to assist the robotic grasping tasks. Most of the methods Zeng et al. (2017b) presented in the Amazon picking challenge utilize the 6D poses estimated through partial registration first. Zeng et al. (2017b) proposed an approach which segments and labels multiple views of a scene with a fully convolutional neural network, and then fits pre-scanned 3D object models to the segmentation results to obtain the 6D object poses. Besides, Billings and Johnson-Roberson (2018) proposed a method which jointly accomplish object pose estimation and grasp point selection using a Convolutional Neural Network (CNN) pipeline. Wong et al. (2017) proposed a method which integrated RGB-based object segmentation and depth image-based partial registration to obtain the pose of the target object. They presented a novel metric for scoring model registration quality, and conducted multi-hypothesis registration, which achieved accurate pose estimation with $1cm$ position error and $< 5°$ angle error. Using this accurate 6D object pose, grasps are conducted with a high success rate. A few deep learning-based 6D object pose estimation approaches such as DenseFusion Wang et al. (2019) also illustrate high successful rates in conducting practical robotic grasping tasks.

Methods of conducting shape completion There also exist one kind of methods, which conduct 3D shape completion for the partial point cloud, and then estimate grasps. 3D shape completion provides the complete geometry of objects from partial observations, and estimating 6DoF grasp poses on the completed shape is more precise. Most of this kind of methods estimate the object geometry from partial point cloud (Varley et al. 2017; Lundell et al. 2019; Van der Merwe et al. 2019; Watkins-Valls et al. 2019; Tosun et al. 2020), and some other methods (Wang et al. 2018; Yan et al. 2018, 2019; Gao and Tedrake 2019; Sajjan et al. 2019) utilize the RGB-D images. Many of them (Wang et al. 2018; Watkins-Valls et al. 2019) also combine tactile information for better prediction.

Varley et al. (2017) proposed an architecture to enable robotic grasp planning via shape completion. They utilized a 3D convolutional neural network (CNN) to complete the shape, and created a fast mesh for objects not to be grasped, a detailed mesh for objects to be grasped. The grasps are finally estimated on the reconstructed mesh in GraspIt! Miller and Allen (2004) and the grasp with the highest quality is executed. Lundell et al. (2019) proposed a shape completion DNN architecture to capture shape uncertainties, and a probabilistic grasp planning method which utilizes the shape uncertainty to propose robust grasps. Van der Merwe et al. (2019) proposed PointSDF to learn a signed distance function implicit surface for a partially viewed object, and proposed a grasp success prediction learning architecture which implicitly learns geometrically aware point cloud encodings. Watkins-Valls et al. (2019) also incorporated depth and tactile information to create rich and accurate 3D models useful for robotic manipulation tasks. They utilized both the depth and tactile as input and fed them directly into the model rather than using the tactile information to refine the results. Tosun et al. (2020) utilized a grasp proposal network and a learned 3D shape reconstruction network, where candidate grasps generated from the first network are refined using the 3D reconstruction result of the second network. These above methods mainly utilize depth data or point cloud as inputs.

Wang et al. (2018) perceived accurate 3D object shape by incorporating visual and tactile observations, as well as prior knowledge of common object shapes learned from large-scale shape repositories. They first applied neural networks with learned shape priors to predict an object's 3D shape from a single-view color image and the tactile sensing was used to refine the shape. Yan et al. (2018) proposed a deep geometry-aware grasping network (DGGN), which first learn a 6DoF grasp from RGB-D input. DGGN has a shape generation network and an outcome prediction network. Yan et al. (2019) further presented a self-supervised shape prediction framework that reconstructs full 3D point clouds as representation for robotic applications. They first used an object detection network to obtain object-centric color, depth and mask images, which will be used to generate a 3D point cloud of the detected object. A grasping critic network is then used to predict a grasp. Gao and Tedrake (2019) proposed a new hybrid object representation consisting of semantic keypoints and dense geometry (a point cloud or mesh) as the interface between the perception module and motion planner. Leveraging advances in learning-based keypoint detection and shape completion, both dense geometry and keypoints can be perceived from raw sensor input. Sajjan et al. (2019) presented ClearGrasp, a deep learning approach for estimating accurate 3D geometry of transparent objects from a single RGB-D image for robotic manipulation. ClearGrasp uses deep convolutional networks to infer surface normals, masks of transparent surfaces, and occlusion boundaries, which will refine the initial depth estimates for all transparent surfaces in the scene.

Comparisons and Discussions When accurate 3D models are available, the 6D object pose could be achieved, which affords the generation of grasps for the target object. However, when existing 3D models are different from the target one, the 6D poses will have a large deviation, and this will lead to the failure of the grasp. In this case, we can complete the partial-view point cloud and triangulate it to obtain the complete shape. The grasps could be generated on the reconstructed and complete 3D shape. Various grasp simulation toolkits are developed to facilitate the grasps generation.

Aiming at methods of estimating the 6D object pose, there exist some challenges. Firstly, this kind of methods highly rely on the accuracy of object segmentation. However, training a network which supports a wide range of objects is not easy. Meanwhile, these methods require the 3D object to grasp be similar enough to those of the annotated models such that correspondences can be found. It is also challenging to compute grasp points with high qualities for objects in cluttered environments where occlusion usually occurs. Aiming at methods of conducting shape completion, there also exist some challenges. The lack of information, especially the geometry on the opposite direction from the camera, extremely affect the completion accuracy. However, using multi-source data would be a future direction.

5 Challenges and future directions

In this survey, we review related works on vision-based robotic grasping from three key aspects: object localization, object pose estimation and grasp estimation. The purpose of this survey is to allow readers to get a comprehensive map about how to detect a successful grasp given the initial raw data. Various subdivided methods are introduced in each section, as well as the related datasets and comparisons. Comparing with existing literatures, we present an end-to-end review about how to conduct a vision-based robotic grasp detection system.

Although so many intelligent algorithms are proposed to assist the robotic grasping tasks, challenges still exist in practical applications, such as the insufficient information in data acquisition, the insufficient amounts of training data, the generalities in grasping novel objects and the difficulties in grasping transparent objects.

The first challenge is the insufficient information in data acquisition. Currently, the mostly used input to decide a grasp is one RGB-D image from one fixed position, which lacks the information backwards. It's really hard to decide the grasp when we do not have the full object geometry. Aiming at this challenge, some strategies could be adopted. The first strategy is to utilize multi-view data. A more widely perspective data is much better since the partial views are not enough to get a comprehensive knowledge of the target object. Methods based on poses of the robotic arms (Blomqvist et al. 2020) or the slam methods Dai et al. (2017) can be adopted to merge the multi-view data. Instead of fusing multi-view data, the best grasping view could also be chosen explicitly Morrison et al. (2019). The second one is to involve multi-sensor data such as the haptic information. There exist some works (Lee et al. 2019; Falco et al. 2019; Hogan et al. 2020) which already involve the tactile data to assist the robotic grasping tasks.

The second challenge is the insufficient amounts of training data. The requirements for the training data is extremely large if we want to build an intelligent enough grasp detection system. The amount of open grasp datasets is really small and the involved objects are mostly instance-level, which is too small compared with the objects in our daily life.

Aiming at this challenges, some strategies could be adopted. The first strategy is to utilize simulated environments to generate virtual data Tremblay et al. (2018). Once the virtual grasp environments are built, large amounts of virtual data could be generated by simulating the sensors from various angles. Since there exists gaps from the simulation data to the practical one, many domain adaptation methods (Bousmalis et al. 2018; Fang et al. 2018; Zhao et al. 2020) have been proposed. The second strategy is to utilize the semi-supervised learning approaches (Mahajan et al. 2020; Yokota et al. 2020) to learn to grasp with incorporate unlabeled data. The third strategy is to utilize self-supervised learning methods to generate the labeled data for 6D object pose estimation Deng et al. (2020) or grasp detection Suzuki et al. (2020).

The third challenge is the generalities in grasping novel objects. The mentioned grasp estimation methods, except for methods of evaluating the 6D object pose, all have certain generalities in dealing with novel objects. But these methods mostly work well on trained dataset and show reduced performance for novel objects. Other than improving the performance of the mentioned algorithms, some strategies could be adopted. The first strategy is to utilize the category-level 6D object pose estimation. Lots of works (Wang et al. 2019; Park et al. 2020; Wang et al. 2019; Chen et al. 2020) start to deal with the 6D object pose estimation of category-level objects, since high performance have been achieved on instance-level objects. The second strategy is to involve more semantic information in the grasp detection system. With the help of various shape segmentation methods (Yu et al. 2019; Luo et al. 2020), parts of the object instead of the complete shape can be used to decrease the range of candidate grasping points. The surface material and the weight information could also be estimated to obtain more precise grasping detection results.

The fourth challenge lies in grasping transparent objects. Transparent objects are prevalent in our daily life, but capturing their 3D information is rather difficult for nowadays depth sensors. There exist some pioneering works that tackle this problem in different ways. GlassLoc Zhou et al. (2019) was proposed for grasp pose detection of transparent objects in transparent clutter using plenoptic sensing. KeyPose Liu et al. (2020) conducted multi-view 3D labeling and keypoint estimation for transparent objects in order to estimate their 6D poses. ClearGrasp Sajjan et al. (2019) estimates accurate 3D geometry of transparent objects from a single RGB-D image for robotic manipulation. This area will be further researched in order to make grasps much accurate and robust in daily life.

References

Akkaya I, Andrychowicz M, Chociej M, Litwin M, McGrew B, Petron A, Paino A, Plappert M, Powell G, Ribas R, et al (2019) Solving rubik's cube with a robot hand. Preprint arXiv:1910.07113

Aldoma A, Vincze M, Blodow N, Gossow D, Gedikli S, Rusu RB, Bradski G (2011) Cad-model recognition and 6dof pose estimation using 3d cues. In: 2011 IEEE international conference on computer vision workshops (ICCV workshops), IEEE, pp 585–592

Aoki Y, Goforth H, Srivatsan RA, Lucey S (2019) Pointnetlk: robust & efficient point cloud registration using pointnet. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7163–7172

Ardón P, Pairet È, Petrick RP, Ramamoorthy S, Lohan KS (2019) Learning grasp affordance reasoning through semantic relations. IEEE Robot Autom Lett 4(4):4571–4578

Asif U, Tang J, Harrer S (2018) Graspnet: an efficient convolutional neural network for real-time grasp detection for low-powered devices. In: IJCAI, pp 4875–4882

Bay H, Tuytelaars T, Van Gool L (2006) Surf: speeded up robust features. In: European conference on computer vision, Springer, pp 404–417

Bellekens B, Spruyt V, Berkvens R, Weyn M (2014) A survey of rigid 3d pointcloud registration algorithms. In: AMBIENT 2014: the fourth international conference on ambient computing, applications, services and technologies, August 24–28, 2014, Rome, Italy, pp 8–13

Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522

Berscheid L, Meißner P, Kröger T (2019) Robot learning of shifting objects for grasping in cluttered environments. Preprint arXiv:1907.11035

Besl PJ, McKay ND (1992) A method for registration of 3-d shapes. IEEE Trans Pattern Anal Mach Intell 14(2):239–256

Bhatia S, Chalup SK et al (2013) Segmenting salient objects in 3d point clouds of indoor scenes using geodesic distances. J Signal Inf Process 4(03):102

Billings G, Johnson-Roberson M (2018) Silhonet: An RGB method for 3d object pose estimation and grasp planning. CoRR abs/1809.06893

Blomqvist K, Breyer M, Cramariuc A, Förster J, Grinvald M, Tschopp F, Chung JJ, Ott L, Nieto J, Siegwart R (2020) Go fetch: mobile manipulation in unstructured environments. Preprint arXiv:2004.00899

Bochkovskiy A, Wang CY, Liao HYM (2020) Yolov4: optimal speed and accuracy of object detection. Preprint arXiv:2004.10934

Bohg J, Kragic D (2010) Learning grasping points with shape context. Robot Auton Syst 58(4):362–377

Bohg J, Morales A, Asfour T, Kragic D (2014) Data-driven grasp synthesis: a survey. IEEE Trans Robot 30(2):289–309

Bolya D, Zhou C, Xiao F, Lee YJ (2019) Yolact++: better real-time instance segmentation. Preprint arXiv:1912.06218

Bolya D, Zhou C, Xiao F, Lee YJ (2019) Yolact: real-time instance segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 9157–9166

Borji A, Cheng MM, Hou Q, Jiang H, Li J (2019) Salient object detection: A survey. Computational visual media 5(2):117–150

Borst C, Fischer M, Hirzinger G (2003) Grasping the dice by dicing the grasp. In: IEEE/RSJ international conference on intelligent robots and systems, IEEE, vol 4, pp 3692–3697

Bousmalis K, Irpan A, Wohlhart P, Bai Y, Kelcey M, Kalakrishnan M, Downs L, Ibarz J, Pastor P, Konolige K et al (2018) Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, pp 4243–4250

Brachmann E, Krull A, Michel F, Gumhold S, Shotton J, Rother C (2014) Learning 6d object pose estimation using 3d object coordinates. In: European conference on computer vision, Springer, pp 536–551

Brachmann E, Michel F, Krull A, Ying Yang M, Gumhold S et al (2016) Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3364–3372

Bradski G, Kaehler A (2008) Learning OpenCV: computer vision with the OpenCV library. " O'Reilly Media, Inc."

Cai J, Cheng H, Zhang Z, Su J (2019) Metagrasp: data efficient grasping by affordance interpreter network. In: 2019 international conference on robotics and automation (ICRA), IEEE, pp 4960–4966

Caldera S, Rassau A, Chai D (2018) Review of deep learning methods in robotic grasp detection. Multimodal Technol Interact 2(3):57

Castro P, Armagan A, Kim TK (2020) Accurate 6d object pose estimation by pose conditioned mesh reconstruction. In: ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE, pp 4147–4151

Chen D, Li J, Wang Z, Xu K (2020) Learning canonical shape space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11973–11982

Chen H, Li Y (2018) Progressively complementarity-aware fusion network for rgb-d salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3051–3060

Chen H, Li Y (2019) Cnn-based rgb-d salient object detection: learn, select and fuse. Preprint arXiv:1909.09309

Chen H, Li Y, Su D (2019) Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. Pattern Recogn 86:376–385

Chen H, Sun K, Tian Z, Shen C, Huang Y, Yan Y (2020) Blendmask: top-down meets bottom-up for instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8573–8581

Chen IM, Burdick JW (1993) Finding antipodal point grasps on irregularly shaped objects. IEEE Trans Robot Autom 9(4):507–512

Chen K, Pang J, Wang J, Xiong Y, Li X, Sun S, Feng W, Liu Z, Shi J, Ouyang W, et al (2019) Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4974–4983

Chen LC, Hermans A, Papandreou G, Schroff F, Wang P, Adam H (2018) Masklab: instance segmentation by refining object detection with semantic and direction features. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4013–4022

Chen W, Jia X, Chang HJ, Duan J, Leonardis A (2020) G2l-net: global to local network for real-time 6d pose estimation with embedding vector features. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4233–4242

Chen X, Girshick R, He K, Dollár P (2019) Tensormask: a foundation for dense object segmentation. In: Proceedings of the IEEE international conference on computer vision, pp 2061–2069

Chen X, Ma H, Wan J, Li B, Xia T (2017) Multi-view 3d object detection network for autonomous driving. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1907–1915

Cheng MM, Mitra NJ, Huang X, Torr PH, Hu SM (2014) Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell 37(3):569–582

Choy C, Dong W, Koltun V (2020) Deep global registration. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2514–2523

Chu FJ, Xu R, Vela PA (2018) Real-world multiobject, multigrasp detection. IEEE Robot Autom Lett 3(4):3355–3362

Chu FJ, Xu R, Vela PA (2019) Detecting robotic affordances on novel objects with regional attention and attributes. Preprint arXiv:1909.05770

Crivellaro A, Rad M, Verdie Y, Yi KM, Fua P, Lepetit V (2017) Robust 3d object tracking from monocular images using stable parts. IEEE Trans Pattern Anal Mach Intell 40(6):1465–1479

Dai A, Nießner M, Zollhöfer M, Izadi S, Theobalt C (2017) Bundlefusion: real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. ACM Trans Graph (ToG) 36(4):1

Dai J, He K, Li Y, Ren S, Sun J (2016) Instance-sensitive fully convolutional networks. In: European conference on computer vision, Springer, pp 534–549

Dai J, He K, Sun J (2016) Instance-aware semantic segmentation via multi-task network cascades. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3150–3158

Dai J, Li Y, He K, Sun J (2016) R-fcn: object detection via region-based fully convolutional networks. In: Advances in neural information processing systems, pp 379–387

Danielczuk M, Matl M, Gupta S, Li A, Lee A, Mahler J, Goldberg K (2019) Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data. In: 2019 international conference on robotics and automation (ICRA), IEEE, pp 7283–7290

Deng X, Xiang Y, Mousavian A, Eppner C, Bretl T, Fox D (2020) Self-supervised 6d object pose estimation for robot manipulation. In: International conference on robotics and automation (ICRA)

Depierre A, Dellandréa E, Chen L (2018) Jacquard: a large scale dataset for robotic grasp detection. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 3511–3516

Depierre A, Dellandréa E, Chen L (2020) Optimizing correlated graspability score and grasp regression for better grasp prediction. Preprint arXiv:2002.00872

DeTone D, Malisiewicz T, Rabinovich A (2018) Superpoint: self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 224–236

Ding D, Liu YH, Wang MY (2001) On computing immobilizing grasps of 3-d curved objects. In: IEEE international symposium on computational intelligence in robotics and automation, IEEE, pp 11–16

Do TT, Cai M, Pham T, Reid I (2018) Deep-6dpose: recovering 6d object pose from a single rgb image. Preprint arXiv:1802.10367

Do TT, Nguyen A, Reid I (2018) Affordancenet: an end-to-end deep learning approach for object affordance detection. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, pp 1–5

Domae Y, Okuda H, Taguchi Y, Sumi K, Hirai T (2014) Fast graspability evaluation on single depth maps for bin picking with general grippers. In: 2014 IEEE international conference on robotics and automation (ICRA), IEEE, pp. 1997–2004

Dong Z, Li G, Liao Y, Wang F, Ren P, Qian C (2020) Centripetalnet: pursuing high-quality keypoint pairs for object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10519–10528

Douglas DH, Peucker TK (1973) Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. Cartogr Int J Geogr Inf Geovis 10(2):112–122

Drost B, Ilic S (2012) 3d object detection and localization using multimodal point pair features. In: International conference on 3D imaging, modeling, processing, visualization transmission, pp 9–16

Drost B, Ulrich M, Navab N, Ilic S (2010) Model globally, match locally: efficient and robust 3d object recognition. In: 2010 IEEE computer society conference on computer vision and pattern recognition, pp 998–1005

Du L, Ye X, Tan X, Feng J, Xu Z, Ding E, Wen S (2020) Associate-3ddet: perceptual-to-conceptual association for 3d point cloud object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13329–13338

Duan K, Bai S, Xie L, Qi H, Huang Q, Tian Q (2019) Centernet: keypoint triplets for object detection. In: Proceedings of the IEEE international conference on computer vision, pp 6569–6578

Engelmann F, Bokeloh M, Fathi A, Leibe B, Nießner M (2020) 3d-mpa: multi-proposal aggregation for 3d semantic instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9031–9040

Erhan D, Szegedy C, Toshev A, Anguelov D (2014) Scalable object detection using deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2147–2154

Falco P, Lu S, Natale C, Pirozzi S, Lee D (2019) A transfer learning approach to cross-modal object recognition: from visual observation to robotic haptic exploration. IEEE Trans Robot 35(4):987–998

Fan Y, Tomizuka M (2019) Efficient grasp planning and execution with multifingered hands by surface fitting. IEEE Robot Autom Lett 4(4):3995–4002

Fan Z, Yu JG, Liang Z, Ou J, Gao C, Xia GS, Li Y (2020) Fgn: fully guided network for few-shot instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9172–9181

Fang HS, Wang C, Gou M, Lu C (2020) Graspnet-1billion: a large-scale benchmark for general object grasping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11444–11453

Fang K, Bai Y, Hinterstoisser S, Savarese S, Kalakrishnan M (2018) Multi-task domain adaptation for deep learning of instance grasping from simulation. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, pp 3516–3523

Fischler MA, Bolles RC (1981) Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun ACM 24(6):381–395

Fitzgibbon AW, Fisher RB et al (1996) A buyer's guide to conic fitting. Department of Artificial Intelligence, University of Edinburgh, Edinburgh

Florence PR, Manuelli L, Tedrake R (2018) Dense object nets: learning dense visual object descriptors by and for robotic manipulation. Preprint arXiv:1806.08756

Frome A, Huber D, Kolluri R, Bülow T, Malik J (2004) Recognizing objects in range data using regional point descriptors. In: European conference on computer vision, Springer, pp 224–237

Gao G, Lauri M, Wang Y, Hu X, Zhang J, Frintrop S (2020) 6d object pose regression via supervised learning on point clouds. Preprint arXiv:2001.08942

Gao W, Tedrake R (2019) kpam-sc: generalizable manipulation planning using keypoint affordance and shape completion. Preprint arXiv:1909.06980

Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE international conference on computer vision, pp 1440–1448

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE conference on computer vision and pattern recognition, CVPR '14, pp 580–587

Gojcic Z, Zhou C, Wegner JD, Wieser A (2019) The perfect match: 3d point cloud matching with smoothed densities. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5545–5554

Gonzalez M, Kacete A, Murienne A, Marchand E (2020) Yoloff: you only learn offsets for robust 6dof object pose estimation. Preprint arXiv:2002.00911

Gordo A, Almazán J, Revaud J, Larlus D (2016) Deep image retrieval: learning global representations for image search. In: European conference on computer vision, Springer, pp 241–257

Goron LC, Marton ZC, Lazea G, Beetz M (2012) Robustly segmenting cylindrical and box-like objects in cluttered scenes using depth cameras. In: ROBOTIK 2012; 7th German conference on robotics, VDE, pp 1–6

Graham B, Engelcke M, van der Maaten L (2018) 3d semantic segmentation with submanifold sparse convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9224–9232

Graham B, van der Maaten L (2017) Submanifold sparse convolutional networks. Preprint arXiv:1706.01307

Guo D, Kong T, Sun F, Liu H (2016) Object discovery and grasp detection with a shared convolutional neural network. In: IEEE international conference on robotics and automation (ICRA), IEEE, pp 2038–2043

Guo D, Sun F, Liu H, Kong T, Fang B, Xi N (2017) A hybrid deep architecture for robotic grasp detection. In: 2017 IEEE international conference on robotics and automation (ICRA), IEEE, pp 1609–1614

Guo F, Wang W, Shen J, Shao L, Yang J, Tao D, Tang YY (2017) Video saliency detection using object proposals. IEEE Trans Cybern 48(11):3159–3170

Guo Y, Bennamoun M, Sohel F, Lu M, Wan J, Kwok NM (2016) A comprehensive performance evaluation of 3d local feature descriptors. Int J Comput Vis 116(1):66–89

Guo Y, Wang H, Hu Q, Liu H, Liu L, Bennamoun M (2019) Deep learning for 3d point clouds: a survey. Preprint arXiv:1912.12033

Hafiz AM, Bhat GM (2020) A survey on instance segmentation: state of the art. Int J Multimed Inf Retr 9(3):171–189

Hagelskjær F, Buch AG (2019) Pointposenet: accurate object detection and 6 dof pose estimation in point clouds. Preprint arXiv:1912.09057

Han J, Zhang D, Cheng G, Liu N, Xu D (2018) Advanced deep-learning techniques for salient and category-specific object detection: a survey. IEEE Signal Process Mag 35(1):84–100

Han L, Zheng T, Xu L, Fang L (2020) Occuseg: occupancy-aware 3d instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2940–2949

Hariharan B, Arbeláez P, Girshick R, Malik J (2014) Simultaneous detection and segmentation. In: European conference on computer vision, Springer, pp 297–312

He K, Gkioxari G, Dollár P, Girshick RB (2017) Mask r-cnn. IEEE International conference on computer vision (ICCV), pp 2980–2988

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778

He Y, Sun W, Huang H, Liu J, Fan H, Sun J (2020) Pvn3d: a deep point-wise 3d keypoints voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11632–11641

Hinterstoisser S, Lepetit V, Ilic S, Holzer S, Bradski G, Konolige K, Navab N (2012) Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision, Springer, pp 548–562

Hinton GE, Krizhevsky A, Wang SD (2011) Transforming auto-encoders. In: International conference on artificial neural networks, Springer, pp 44–51

Hodan T, Barath D, Matas J (2020) Epos: estimating 6d pose of objects with symmetries. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11703–11712

Hodaň T, Haluza P, Obdržálek Š, Matas J, Lourakis M, Zabulis X (2017) T-LESS: an RGB-D dataset for 6D pose estimation of texture-less objects. In: IEEE winter conference on applications of computer vision (WACV)

Hodan T, Kouskouridas R, Kim T, Tombari F, Bekris KE, Drost B, Groueix T, Walas K, Lepetit V, Leonardis A, Steger C, Michel F, Sahin C, Rother C, Matas J (2018) A summary of the 4th international workshop on recovering 6d object pose. CoRR abs/1810.03758

Hodaň T, Michel F, Brachmann E, Kehl W, GlentBuch A, Kraft D, Drost B, Vidal J, Ihrke S, Zabulis X et al (2018) Bop: benchmark for 6d object pose estimation. In: Proceedings of the European conference on computer vision (ECCV), pp 19–34

Hodaň T, Zabulis X, Lourakis M, Obdržálek Š, Matas J (2015) Detection and fine 3d pose estimation of texture-less objects in rgb-d images. In: 2015 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 4421–4428

Hogan FR, Ballester J, Dong S, Rodriguez A (2020) Tactile dexterity: manipulation primitives with tactile feedback. Preprint arXiv:2002.03236

Hou J, Dai A, Nießner M (2019) 3d-sis: 3d semantic instance segmentation of rgb-d scans. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4421–4430

Hou Q, Cheng MM, Hu X, Borji A, Tu Z, Torr PH (2017) Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3203–3212

Hu Y, Fua P, Wang W, Salzmann M (2020) Single-stage 6d object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2930–2939

Hu Y, Hugonot J, Fua P, Salzmann M (2019) Segmentation-driven 6d object pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3385–3394

Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: a discriminative regional feature integration approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2083–2090

Jiang H, Xiao J (2013) A linear approach to matching cuboids in rgbd images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2171–2178

Jiang Y, Moseson S, Saxena A (2011) Efficient grasping from rgbd images: learning using a new rectangle representation. In: IEEE international conference on robotics and automation, IEEE, pp 3304–3311

Johnson AE (1997) Spin-images: a representation for 3-d surface matching

Kaiser A, Ybanez Zepeda JA, Boubekeur T (2019) A survey of simple geometric primitives detection methods for captured 3d data. In: Computer graphics forum, Wiley Online Library, vol 38, pp 167–196

Kehl W, Manhardt F, Tombari F, Ilic S, Navab N (2017) Ssd-6d: making rgb-based 3d detection and 6d pose estimation great again. In: Proceedings of the IEEE international conference on computer vision, pp 1521–1529

Khan SH, He X, Bennamoun M, Sohel F, Togneri R (2015) Separating objects and clutter in indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4603–4611

Kim G, Huber D, Hebert M (2008) Segmentation of salient regions in outdoor scenes using imagery and 3-d data. In: 2008 IEEE workshop on applications of computer vision, IEEE, pp 1–8

Kirillov A, Wu Y, He K, Girshick R (2020) Pointrend: image segmentation as rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9799–9808

Kirkpatrick D, Mishra B, Yap CK (1992) Quantitative steinitz's theorems with applications to multifingered grasping. Discrete Comput Geom 7(3):295–318

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems—volume 1, NIPS'12, pp 1097–1105

Kumra S, Joshi S, Sahin F (2019) Antipodal robotic grasping using generative residual convolutional neural network. Preprint arXiv:1909.04810

Kumra S, Kanan C (2017) Robotic grasp detection using deep convolutional neural networks. In: IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 769–776

Lang AH, Vora S, Caesar H, Zhou L, Yang J, Beijbom O (2019) Pointpillars: fast encoders for object detection from point clouds. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 12697–12705

Law H, Deng J (2018) Cornernet: detecting objects as paired keypoints. In: Proceedings of the European conference on computer vision (ECCV), pp 734–750

Lee MA, Zhu Y, Srinivasan K, Shah P, Savarese S, Fei-Fei L, Garg A, Bohg J (2019) Making sense of vision and touch: self-supervised learning of multimodal representations for contact-rich tasks. In: 2019 international conference on robotics and automation (ICRA), IEEE, pp 8943–8950

Lee Y, Park J (2020) Centermask: real-time anchor-free instance segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 13906–13915

Lenz I, Lee H, Saxena A (2015) Deep learning for detecting robotic grasps. Int J Robot Res 34(4–5):705–724

León B, Ulbrich S, Diankov R, Puche G, Przybylski M, Morales A, Asfour T, Moisio S, Bohg J, Kuffner J, Dillmann R (2010) Opengrasp: a toolkit for robot grasping simulation. In: Ando N, Balakirsky S, Hemker T, Reggiani M, von Stryk O (eds) Simulation, modeling, and programming for autonomous robots. Springer, Berlin, pp 109–120

Lepetit V, Fua P et al (2005) Monocular model-based 3d tracking of rigid objects: a survey. Found Trends® Comput Graph Vis 1(1):1–89

Lepetit V, Moreno-Noguer F, Fua P (2009) Epnp: an accurate o(n) solution to the pnp problem. IJCV 81(2):155–166

Li G, Liu Z, Ye L, Wang Y, Ling H (2020) Cross-modal weighting network for rgb-d salient object detection

Li Y, Qi H, Dai J, Ji X, Wei Y (2017) Fully convolutional instance-aware semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2359–2367

Li Y, Wang G, Ji X, Xiang Y, Fox D (2018) Deepim: deep iterative matching for 6d pose estimation. Lecture notes in computer science, pp 695–711

Li Z, Wang G, Ji X (2019) Cdpn: coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In: Proceedings of the IEEE international conference on computer vision, pp 7678–7687

Liang H, Ma X, Li S, Görner M, Tang S, Fang B, Sun F, Zhang J (2019) Pointnetgpd: detecting grasp configurations from point sets. In: 2019 international conference on robotics and automation (ICRA), IEEE, pp 3629–3635

Liang M, Yang B, Chen Y, Hu R, Urtasun R (2019) Multi-task multi-sensor fusion for 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7345–7353

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2117–2125

Lin TY, Goyal P, Girshick R, He K, Dollár P (2017) Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2980–2988

Liu C, Furukawa Y (2019) Masc: multi-scale affinity with sparse convolution for 3d instance segmentation. Preprint arXiv:1902.04478

Liu F, Fang P, Yao Z, Fan R, Pan Z, Sheng W, Yang H (2019) Recovering 6d object pose from rgb indoor image based on two-stage detection network withmulti-task loss. Neurocomputing 337:15–23

Liu L, Ouyang W, Wang X, Fieguth P, Chen J, Liu X, Pietikäinen M (2020) Deep learning for generic object detection: a survey. Int J Comput Vis 128(2):261–318

Liu M, Pan Z, Xu K, Ganguly K, Manocha D (2019) Generating grasp poses for a high-dof gripper using neural networks. Preprint arXiv:1903.00425

Liu N, Han J (2016) Dhsnet: deep hierarchical saliency network for salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 678–686

Liu N, Han J, Yang MH (2018) Picanet: learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3089–3098

Liu S, Qi L, Qin H, Shi J, Jia J (2018) Path aggregation network for instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8759–8768

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: single shot multibox detector. In: European conference on computer vision, Springer, pp 21–37

Liu X, Jonschkowski R, Angelova A, Konolige K (2020) Keypose: multi-view 3d labeling and keypoint estimation for transparent objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11602–11610

Liu Y, Zhang Q, Zhang D, Han J (2019) Employing deep part-object relationships for salient object detection. In: Proceedings of the IEEE international conference on computer vision, pp 1232–1241

Liu Z, Zhao X, Huang T, Hu R, Zhou Y, Bai X (2020) Tanet: robust 3d object detection from point clouds with triple attention. In: AAAI, pp 11677–11684

Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440

Lou X, Yang Y, Choi C (2019) Learning to generate 6-dof grasp poses with reachability awareness. Preprint arXiv:1910.06404

Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the international conference on computer vision-Volume 2, ICCV '99, p 1150

Lu W, Wan G, Zhou Y, Fu X, Yuan P, Song S (2019) Deepicp: an end-to-end deep neural network for 3d point cloud registration. Preprint arXiv:1905.04153

Lundell J, Verdoja F, Kyrki V (2019) Robust grasp planning over uncertain shape completions. Preprint arXiv:1903.00645

Luo T, Mo K, Huang Z, Xu J, Hu S, Wang L, Su H (2020) Learning to group: a bottom-up framework for 3d part discovery in unseen categories. In: International conference on learning representations

Mahajan M, Bhattacharjee T, Krishnan A, Shukla P, Nandi G (2020) Semi-supervised grasp detection by representation learning in a vector quantized latent space. Preprint arXiv:2001.08477

Mahler J, Liang J, Niyaz S, Laskey M, Doan R, Liu X, Ojea JA, Goldberg K (2017) Dex-net 2.0: seep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. CoRR arXiv:1703.09312

Malisiewicz T, Gupta A, Efros AA (2011) Ensemble of exemplar-svms for object detection and beyond. In: 2011 International conference on computer vision, IEEE, pp 89–96

Mellado N, Aiger D, Mitra NJ (2014) Super 4pcs fast global pointcloud registration via smart indexing. In: Computer graphics forum, Wiley Online Library, vol 33, pp 205–215

Van der Merwe M, Lu Q, Sundaralingam B, Matak M, Hermans T (2019) Learning continuous 3d reconstructions for geometrically aware grasping. Preprint arXiv:1910.00983

Miller AT, Allen PK (2004) Graspit! a versatile simulator for robotic grasping. IEEE Robot Autom Mag 11(4):110–122

Miller AT, Knoop S, Christensen HI, Allen PK (2003) Automatic grasp planning using shape primitives. ICRA 2:1824–1829

Minaee S, Boykov Y, Porikli F, Plaza A, Kehtarnavaz N, Terzopoulos D (2020) Image segmentation using deep learning: a survey. Preprint arXiv:2001.05566

Mirtich B, Canny J (1994) Easily computable optimum grasps in 2-d and 3-d. In: IEEE international conference on robotics and automation, IEEE, pp 739–747

Morrison D, Corke P, Leitner J (2018) Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach. Preprint arXiv:1804.05172

Morrison D, Corke P, Leitner J (2019) Multi-view picking: next-best-view reaching for improved grasping in clutter. In: 2019 international conference on robotics and automation (ICRA), IEEE, pp 8762–8768

Mousavian A, Eppner C, Fox D (2019) 6-dof graspnet: variational grasp generation for object manipulation. In: Proceedings of the IEEE international conference on computer vision, pp 2901–2910

Mur-Artal R, Montiel JMM, Tardos JD (2015) Orb-slam: a versatile and accurate monocular slam system. IEEE Trans Robot 31(5):1147–1163

Murali A, Mousavian A, Eppner C, Paxton C, Fox D (2019) 6-dof grasping for target-driven object manipulation in clutter. Preprint arXiv:1912.03628

Najibi M, Lai G, Kundu A, Lu Z, Rathod V, Funkhouser T, Pantofaru C, Ross D, Davis LS, Fathi A (2020) Dops: learning to detect 3d objects and predict their 3d shapes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11913–11922

Nguyen VD (1987) Constructing stable grasps in 3d. In: IEEE international conference on robotics and automation, IEEE, vol 4, pp 234–239

Ni P, Zhang W, Zhu X, Cao Q (2020) Pointnet++ grasping: learning an end-to-end spatial grasp generation algorithm from sparse point clouds. Preprint arXiv:2003.09644

Nikandrova E, Kyrki V (2015) Category-based task specific grasping. Robot Auton Syst 70:25–35

Oberweger M, Rad M, Lepetit V (2018) Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In: Proceedings of the European conference on computer vision (ECCV), pp 119–134

Pang Y, Zhang L, Zhao X, Lu H (2020) Hierarchical dynamic filtering network for rgb-d salient object detection. In: Proceedings of the European conference on computer vision (ECCV)

Park D, Chun SY (2018) Classification based grasp detection using spatial transformer network. Preprint arXiv:1803.01356

Park D, Seo Y, Chun SY (2018) Real-time, highly accurate robotic grasp detection using fully convolutional neural network with rotation ensemble module. Preprint arXiv:1812.07762

Park D, Seo Y, Shin D, Choi J, Chun SY (2019) A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection. Preprint arXiv:1909.07050

Park K, Mousavian A, Xiang Y, Fox D (2020) Latentfusion: end-to-end differentiable reconstruction and rendering for unseen object pose estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10710–10719

Park K, Patten T, Vincze M (2019) Pix2pose: pixel-wise coordinate regression of objects for 6d pose estimation. In: Proceedings of the IEEE international conference on computer vision, pp 7668–7677

ten Pas A, Gualtieri M, Saenko K, Platt R (2017) Grasp pose detection in point clouds. Int J Rob Res 36(13–14):1455–1473

Pas At, Platt R (2015) Using geometry to detect grasps in 3d point clouds. Preprint arXiv:1501.03100

Patil AV, Rabha P (2018) A survey on joint object detection and pose estimation using monocular vision. Preprint arXiv:1811.10216

Patten T, Park K, Vincze M (2020) Dgcm-net: dense geometrical correspondence matching network for incremental experience-based robotic grasping. Preprint arXiv:2001.05279

Peng H, Li B, Ling H, Hu W, Xiong W, Maybank SJ (2016) Salient object detection via structured matrix decomposition. IEEE Trans Pattern Anal Mach Intell 39(4):818–832

Peng H, Li B, Xiong W, Hu W, Ji R (2014) Rgbd salient object detection: a benchmark and algorithms. In: European conference on computer vision, Springer, pp 92–109

Peng S, Liu Y, Huang Q, Zhou X, Bao H (2019) Pvnet: pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4561–4570

Pereira N, Alexandre LA (2019) Maskedfusion: mask-based 6d object pose estimation. Preprint arXiv:1911.07771

Pham QH, Nguyen T, Hua BS, Roig G, Yeung SK (2019) Jsis3d: joint semantic-instance segmentation of 3d point clouds with multi-task pointwise networks and multi-value conditional random fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8827–8836

Pham QH, Uy MA, Hua BS, Nguyen DT, Roig G, Yeung SK (2020) Lcd: learned cross-domain descriptors for 2d–3d matching. In: AAAI, pp 11856–11864

Piao Y, Ji W, Li J, Zhang M, Lu H (2019) Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE international conference on computer vision, pp 7254–7263

Pinheiro PO, Collobert R, Dollár P (2015) Learning to segment object candidates. In: Advances in neural information processing systems, pp 1990–1998

Pinheiro PO, Lin TY, Collobert R, Dollár P (2016) Learning to refine object segments. In: European conference on computer vision, Springer, pp 75–91

Pinto L, Gupta A (2016) Supersizing self-supervision: learning to grasp from 50k tries and 700 robot hours. In: IEEE International conference on robotics and automation (ICRA), IEEE, pp 3406–3413

Ponce J, Sullivan S, Boissonnat JD, Merlet JP (1993) On characterizing and computing three-and four-finger force-closure grasps of polyhedral objects. In: IEEE international conference on robotics and automation, IEEE, pp 821–827

Qi CR, Chen X, Litany O, Guibas LJ (2020) Imvotenet: boosting 3d object detection in point clouds with image votes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4404–4413

Qi CR, Litany O, He K, Guibas LJ (2019) Deep hough voting for 3d object detection in point clouds. In: Proceedings of the IEEE international conference on computer vision, pp 9277–9286

Qi CR, Liu W, Wu C, Su H, Guibas LJ (2018) Frustum pointnets for 3d object detection from rgb-d data. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 918–927

Qi CR, Su H, Mo K, Guibas LJ (2017) Pointnet: deep learning on point sets for 3d classification and segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 652–660

Qi CR, Yi L, Su H, Guibas LJ (2017) Pointnet++: deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems, pp 5099–5108

Qi Q, Zhao S, Shen J, Lam KM (2019) Multi-scale capsule attention-based salient object detection with multi-crossed layer connections. In: 2019 IEEE international conference on multimedia and expo (ICME), IEEE, pp 1762–1767

Qin Y, Chen R, Zhu H, Song M, Xu J, Su H (2020) S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In: Conference on robot learning, pp 53–65

Qu L, He S, Zhang J, Tian J, Tang Y, Yang Q (2017) Rgbd salient object detection via deep fusion. IEEE Trans Image Process 26(5):2274–2285

Rabbani T, Van Den Heuvel F (2005) Efficient hough transform for automatic detection of cylinders in point clouds. Isprs Wg Iii/3, Iii/4 3:60–65

Rad M, Lepetit V (2017) Bb8: a scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In: IEEE international conference on computer vision, pp 3828–3836

Redmon J, Angelova A (2015) Real-time grasp detection using convolutional neural networks. In: 2015 IEEE international conference on robotics and automation (ICRA), IEEE, pp 1316–1322

Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 779–788

Redmon J, Farhadi A (2017) Yolo9000: better, faster, stronger. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7263–7271

Redmon J, Farhadi A (2018) Yolov3: an incremental improvement. Preprint arXiv:1804.02767

Ren J, Gong X, Yu L, Zhou W, Ying Yang M (2015) Exploiting global priors for rgb-d saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 25–32

Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: towards real-time object detection with region proposal networks. In: Advances in neural information processing systems, pp 91–99

Rennie C, Shome R, Bekris KE, De Souza AF (2016) A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. IEEE Robot Autom Lett 1(2):1179–1185

Rosten E, Drummond T (2005) Fusing points and lines for high performance tracking. In: Tenth IEEE international conference on computer vision (ICCV'05) Volume 1, IEEE, vol 2, pp 1508–1515

Rublee E, Rabaud V, Konolige K, Bradski G (2011) Orb: an efficient alternative to sift or surf. In: 2011 International conference on computer vision, IEEE, pp 2564–2571

Rusu RB, Blodow N, Beetz M (2009) Fast point feature histograms (fpfh) for 3d registration. In: IEEE international conference on robotics and automation, pp 3212–3217

Rusu RB, Blodow N, Marton ZC, Beetz M (2009) Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in domestic environments. In: 2009 IEEE/RSJ international conference on intelligent robots and systems, IEEE, pp 1–6

Sabour S, Frosst N, Hinton G (2018) Matrix capsules with em routing. In: 6th international conference on learning representations, ICLR, pp 1–15

Sabour S, Frosst N, Hinton GE (2017) Dynamic routing between capsules. In: Advances in neural information processing systems, pp 3856–3866

Sahbani A, El-Khoury S, Bidaud P (2012) An overview of 3d object grasp synthesis algorithms. Robot Auton Syst 60(3):326–336 Autonomous Grasping

Sajjan SS, Moore M, Pan M, Nagaraja G, Lee J, Zeng A, Song S (2019) Cleargrasp: 3d shape estimation of transparent objects for manipulation. Preprint arXiv:1910.02550

Salti S, Tombari F, Stefano LD (2014) Shot: Unique signatures of histograms for surface and texture description. Comput Vis Image Underst 125:251–264

Sanchez J, Corrales JA, Bouzgarrou BC, Mezouar Y (2018) Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey. Int J Robot Res 37(7):688–716

Sarode V, Li X, Goforth H, Aoki Y, Dhagat A, Srivatsan RA, Lucey S, Choset H (2019) One framework to register them all: pointnet encoding for point cloud alignment. Preprint arXiv:1912.05766

Sarode V, Li X, Goforth H, Aoki Y, Srivatsan RA, Lucey S, Choset H (2019) Pcrnet: point cloud registration network using pointnet encoding. Preprint arXiv:1908.07906

Saxena A, Driemeyer J, Kearns J, Osondu C, Ng AY (2008a) Learning to grasp novel objects using vision. In: Experimental robotics, Springer, pp 33–42

Saxena A, Driemeyer J, Ng AY (2008b) Robotic grasping of novel objects using vision. Int J Robot Res 27(2):157–173

Sermanet P, Eigen D, Zhang X, Mathieu M, Fergus R, LeCun Y (2013) Overfeat: integrated recognition, localization and detection using convolutional networks. Preprint arXiv:1312.6229

Shi J, Yan Q, Xu L, Jia J (2015) Hierarchical image saliency detection on extended cssd. IEEE Trans Pattern Anal Mach Intell 38(4):717–729

Shi S, Wang X, Li H (2019) Pointrcnn: 3d object proposal generation and detection from point cloud. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–779

Shi S, Wang Z, Shi J, Wang X, Li H (2020) From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. Preprint arXiv:1907.03670

Shi W, Rajkumar R (2020) Point-gnn: graph neural network for 3d object detection in a point cloud. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1711–1719

Simon M, Fischer K, Milz S, Witt CT, Gross HM (2020) Stickypillars: robust feature matching on point clouds using graph neural networks. Preprint arXiv:2002.03983

Song C, Song J, Huang Q (2020) Hybridpose: 6d object pose estimation under hybrid representations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 431–440

Song S, Xiao J (2014) Sliding shapes for 3d object detection in depth images. In: European conference on computer vision, Springer, pp 634–651

Song S, Xiao J (2016) Deep sliding shapes for amodal 3d object detection in rgb-d images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 808–816

Sultana F, Sufian A, Dutta P (2020) Evolution of image segmentation using deep convolutional neural network: a survey. Preprint arXiv:2001.04074

Sultana F, Sufian A, Dutta P (2020) A review of object detection models based on convolutional neural network. In: Intelligent computing: image processing based applications, Springer, pp 1–16

Sundermeyer M, Marton ZC, Durner M, Brucker M, Triebel R (2018) Implicit 3d orientation learning for 6d object detection from rgb images. In: European conference on computer vision, Springer International Publishing, pp 712–729

Suzuki K, Yokota Y, Kanazawa Y, Takebayashi T (2020) Online self-supervised learning for object picking: detecting optimum grasping position using a metric learning approach. In: 2020 IEEE/SICE international symposium on system integration (SII), IEEE, pp 205–212

Szegedy C, Reed S, Erhan D, Anguelov D, Ioffe S (2014) Scalable, high-quality object detection. Preprint arXiv:1412.1441

Tam GK, Cheng ZQ, Lai YK, Langbein FC, Liu Y, Marshall D, Martin RR, Sun XF, Rosin PL (2013) Registration of 3d point clouds and meshes: a survey from rigid to nonrigid. IEEE Trans Vis Comput Graph 19(7):1199–1217

Tejani A, Tang D, Kouskouridas R, Kim TK (2014) Latent-class hough forests for 3d object detection and pose estimation. In: European conference on computer vision, Springer, pp 462–477

Tekin B, Sinha SN, Fua P (2018) Real-time seamless single shot 6d object pose prediction. In: IEEE conference on computer vision and pattern recognition, pp 292–301

Tian H, Wang C, Manocha D, Zhang X (2019) Transferring grasp configurations using active learning and local replanning. In: 2019 international conference on robotics and automation (ICRA), IEEE, pp 1622–1628

Tian M, Pan L, Ang Jr MH, Lee G.H (2020) Robust 6d object pose estimation by learning rgb-d features. Preprint arXiv:2003.00188

Tian Z, Shen C, Chen H, He T (2019) Fcos: fully convolutional one-stage object detection. In: Proceedings of the IEEE international conference on computer vision, pp 9627–9636

Tosun T, Yang D, Eisner B, Isler V, Lee D (2020) Robotic grasping through combined image-based grasp proposal and 3d reconstruction. Preprint arXiv:2003.01649

Tremblay J, To T, Sundaralingam B, Xiang Y, Fox D, Birchfield S (2018) Deep object pose estimation for semantic robotic grasping of household objects. Preprint arXiv:1809.10790

Truong P, Apostolopoulos S, Mosinska A, Stucky S, Ciller C, Zanet SD (2019) Glampoints: greedily learned accurate match points. In: Proceedings of the IEEE international conference on computer vision, pp 10732–10741

Uijlings JR, Van De Sande KE, Gevers T, Smeulders AW (2013) Selective search for object recognition. Int J Comput Vis 104(2):154–171

Vacchetti L, Lepetit V, Fua P (2004) Stable real-time 3d tracking using online and offline information. IEEE Trans Pattern Anal Mach Intell 26(10):1385–1391

Vahrenkamp N, Westkamp L, Yamanobe N, Aksoy EE, Asfour T (2016) Part-based grasp planning for familiar objects. In: IEEE-RAS 16th international conference on humanoid robots (Humanoids), IEEE, pp 919–925

Varley J, DeChant C, Richardson A, Ruales J, Allen P (2017) Shape completion enabled robotic grasping. In: 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 2442–2447

Vidal J, Lin C, Martí R (2018) 6d pose estimation using an improved method based on point pair features. In: 4th international conference on control, automation and robotics (ICCAR), pp 405–409

Villena-Martinez V, Oprea S, Saval-Calvo M, Azorin-Lopez J, Fuster-Guillo A, Fisher RB (2020) When deep learning meets data alignment: a review on deep registration networks (drns). Preprint arXiv:2003.03167

Vohra M, Prakash R, Behera L (2019) Real-time grasp pose estimation for novel objects in densely cluttered environment. In: 2019 28th IEEE international conference on robot and human interactive communication (RO-MAN), IEEE, pp 1–6

Wada K, Sucar E, James S, Lenton D, Davison AJ (2020) Morefusion: multi-object reasoning for 6d pose estimation from volumetric fusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 14540–14549

Wang C, Martín-Martín R, Xu D, Lv J, Lu C, Fei-Fei L, Savarese S, Zhu Y (2019) 6-pack: category-level 6d pose tracker with anchor-based keypoints. Preprint arXiv:1910.10750

Wang C, Xu D, Zhu Y, Martín-Martín R, Lu C, Fei-Fei L, Savarese S (2019) Densefusion: 6d object pose estimation by iterative dense fusion. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3343–3352

Wang H, Sridhar S, Huang J, Valentin J, Song S, Guibas LJ (2019) Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2642–2651

Wang S, Jiang X, Zhao J, Wang X, Zhou W, Liu Y (2019) Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images. In: 2019 IEEE international conference on robotics and biomimetics (ROBIO), IEEE, pp 474–480

Wang S, Wu J, Sun X, Yuan W, Freeman WT, Tenenbaum JB, Adelson EH (2018) 3d shape perception from monocular vision, touch, and shape priors. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 1606–1613

Wang W, Lai Q, Fu H, Shen J, Ling H (2019) Salient object detection in the deep learning era: an in-depth survey. Preprint arXiv:1904.09146

Wang W, Shen J, Shao L, Porikli F (2016) Correspondence driven saliency transfer. IEEE Trans Image Process 25(11):5025–5034

Wang W, Yu R, Huang Q, Neumann U (2018) Sgpn: similarity group proposal network for 3d point cloud instance segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2569–2578

Wang X, Kong T, Shen C, Jiang Y, Li L (2019) Solo: segmenting objects by locations. Preprint arXiv:1912.04488

Wang X, Liu S, Shen X, Shen C, Jia J (2019) Associatively segmenting instances and semantics in point clouds. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4096–4105

Wang Y, Solomon JM (2019) Deep closest point: learning representations for point cloud registration. In: Proceedings of the IEEE international conference on computer vision, pp 3523–3532

Wang Y, Solomon JM (2019) Prnet: self-supervised learning for partial-to-partial registration. In: Advances in neural information processing systems, pp 8812–8824

Wang Z, Jia K (2019) Frustum convnet: sliding frustums to aggregate local point-wise features for amodal 3d object detection. In: 2019 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 1742–1749

Watkins-Valls D, Varley J, Allen P (2019) Multi-modal geometric learning for grasping and manipulation. In: 2019 international conference on robotics and automation (ICRA), IEEE, pp 7339–7345

Wei Y, Wen F, Zhu W, Sun J (2012) Geodesic saliency using background priors. In: European conference on computer vision, Springer, pp 29–42

Wong JM, Kee V, Le T, Wagner S, Mariottini GL, Schneider A, Hamilton L, Chipalkatty R, Hebert M, Johnson DM, et al (2017) Segicp: integrated deep semantic segmentation and pose estimation. In: IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 5784–5789

Xiang Y, Schmidt T, Narayanan V, Fox D (2018) Posecnn: a convolutional neural network for 6d object pose estimation in cluttered scenes. PreprintarXiv:1711.00199

Xie C, Xiang Y, Mousavian A, Fox D (2020) The best of both modes: separately leveraging rgb and depth for unseen object instance segmentation. In: Conference on robot learning, pp 1369–1378

Xie E, Sun P, Song X, Wang W, Liu X, Liang D, Shen C, Luo P (2020) Polarmask: single shot instance segmentation with polar representation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12193–12202

Xie Q, Lai YK, Wu J, Wang Z, Zhang Y, Xu K, Wang J (2020) Mlcvnet: multi-level context votenet for 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10447–10456

Xu D, Anguelov D, Jain A (2018) Pointfusion: deep sensor fusion for 3d bounding box estimation. In: 2018 IEEE/CVF conference on computer vision and pattern recognition

Xue Z, Kasper A, Zoellner JM, Dillmann R (2009) An automatic grasp planning system for service robots. In: 2009 international conference on advanced robotics, IEEE, pp 1–6

Yan X, Hsu J, Khansari M, Bai Y, Pathak A, Gupta A, Davidson J, Lee H (2018) Learning 6-dof grasping interaction via deep geometry-aware 3d representations. In: 2018 IEEE international conference on robotics and automation (ICRA), IEEE, pp 1–9

Yan X, Khansari M, Hsu J, Gong Y, Bai Y, Pirk S, Lee H (2019) Data-efficient learning for sim-to-real robotic grasping using deep point cloud prediction networks. Preprint arXiv:1906.08989

Yan Y, Mao Y, Li B (2018) Second: sparsely embedded convolutional detection. Sensors 18(10):3337

Yang B, Wang J, Clark R, Hu Q, Wang S, Markham A, Trigoni N (2019) Learning object bounding boxes for 3d instance segmentation on point clouds. In: Advances in neural information processing systems, pp 6737–6746

Yang C, Zhang L, Lu H, Ruan X, Yang MH (2013) Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3166–3173

Yang H, Shi J, Carlone L (2020) Teaser: fast and certifiable point cloud registration. Preprint arXiv:2001.07715

Yang J, Li H, Campbell D, Jia Y (2015) Go-icp: a globally optimal solution to 3d icp point-set registration. IEEE Trans Pattern Anal Mach Intell 38(11):2241–2254

Yang S, Zhang W, Lu W, Wang H, Li Y (2019) Learning actions from human demonstration video for robotic manipulation. Preprint arXiv:1909.04312

Yang Z, Sun Y, Liu S, Jia J (2020) 3dssd: point-based 3d single stage object detector. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 11040–11048

Yang Z, Sun Y, Liu S, Shen X, Jia J (2019) Std: sparse-to-dense 3d object detector for point cloud. In: Proceedings of the IEEE international conference on computer vision, pp 1951–1960

Ye M, Xu S, Cao T (2020) Hvnet: hybrid voxel network for lidar based 3d object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 1631–1640

Yew ZJ, Lee GH (2018) 3dfeat-net: weakly supervised local 3d features for point cloud registration. In: European conference on computer vision, Springer, pp 630–646

Yi KM, Trulls E, Lepetit V, Fua P (2016) Lift: learned invariant feature transform. In: European conference on computer vision, Springer, pp 467–483

Yi L, Zhao W, Wang H, Sung M, Guibas LJ (2019) Gspn: generative shape proposal network for 3d instance segmentation in point cloud. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3947–3956

Yokota Y, Suzuki K, Kanazawa Y, Takebayashi T (2020) A multi-task learning framework for grasping-position detection and few-shot classification. In: 2020 IEEE/SICE international symposium on system integration (SII), IEEE, pp 1033–1039

Yu F, Liu K, Zhang Y, Zhu C, Xu K (2019) Partnet: a recursive part decomposition network for fine-grained and hierarchical shape segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 9491–9500

Yu P, Rao Y, Lu J, Zhou J (2019) P^{2}gnet: pose-guided point cloud generating networks for 6-dof object pose estimation. Preprint arXiv:1912.09316 (2019)

Yu X, Zhuang Z, Koniusz P, Li H (2020) 6dof object pose estimation via differentiable proxy voting loss. Preprint arXiv:2002.03923

Yuan Y, Hou J, Nüchter A, Schwertfeger S (2020) Self-supervised point set local descriptors for point cloud registration. Preprint arXiv:2003.05199

Zakharov S, Shugurov I, Ilic S (2019) Dpod: 6d pose object detector and refiner. In: Proceedings of the IEEE international conference on computer vision, pp 1941–1950

Zapata-Impata BS, Gil P, Pomares J, Torres F (2019) Fast geometry-based computation of grasping points on three-dimensional point clouds. Int J Adv Robot Syst 16(1):1729881419831846

Zapata-Impata BS, Mateo Agulló C, Gil P, Pomares J (2017) Using geometry to detect grasping points on 3d unknown point cloud

Zeng A, Song S, Nießner M, Fisher M, Xiao J, Funkhouser T (2017a) 3dmatch: learning local geometric descriptors from rgb-d reconstructions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1802–1811

Zeng A, Yu KT, Song S, Suo D, Walker E, Rodriguez A, Xiao J (2017b) Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In: IEEE international conference on robotics and automation (ICRA), IEEE, pp 1386–1383

Zeng A, Song S, Yu KT, Donlon E, Hogan FR, Bauza M, Ma D, Taylor O, Liu M, Romo E, et al (2018) Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In: IEEE international conference on robotics and automation (ICRA), IEEE, pp 1–8

Zhang F, Guan C, Fang J, Bai S, Yang R, Torr P, Prisacariu V (2020) Instance segmentation of lidar point clouds. ICRA, Cited by **4**(1)

Zhang H, Lan X, Bai S, Wan L, Yang C, Zheng N (2018) A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes. Preprint arXiv:1809.07081

Zhang H, Lan X, Bai S, Zhou X, Tian Z, Zheng N (2018) Roi-based robotic grasp detection for object overlapping scenes. Preprint arXiv:1808.10313

Zhang J, Sclaroff S, Lin Z, Shen X, Price B, Mech R (2016) Unconstrained salient object detection via proposal subset optimization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5733–5742

Zhang Q, Qu D, Xu F, Zou F (2017) Robust robot grasp detection in multimodal fusion. In: MATEC web of conferences, EDP Sciences, vol 139, p 00060

Zhang Z, Sun B, Yang H, Huang Q (2020) H3dnet: 3d object detection using hybrid geometric primitives. In: Proceedings of the European conference on computer vision (ECCV)

Zhao L, Tao W (2020) Jsnet: Joint instance and semantic segmentation of 3d point clouds. In: Thirty-Fourth AAAI conference on artificial intelligence

Zhao R, Ouyang W, Li H, Wang X (2015) Saliency detection by multi-context deep learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1265–1274

Zhao S, Li B, Xu P, Keutzer K (2020) Multi-source domain adaptation in the deep learning era: a systematic survey. Preprint arXiv:2002.12169

Zhao ZQ, Zheng P, Xu S, Wu X (2019) Object detection with deep learning: a review. IEEE Trans Neural Netw Learn Syst 30(11):3212–3232

Zhao B, Zhang H, Lan X, Wang H, Tian Z, Zheng N (2020) Regnet: region-based grasp network for single-shot grasp detection in point clouds. Preprint arXiv:2002.12647

Zheng T, Chen C, Yuan J, Li B, Ren K (2019) Pointcloud saliency maps. In: Proceedings of the IEEE international conference on computer vision, pp 1598–1606

Zhou QY, Park J, Koltun V (2016) Fast global registration. In: European conference on computer vision, Springer, pp 766–782

Zhou X, Lan X, Zhang H, Tian Z, Zhang Y, Zheng N (2018) Fully convolutional grasp detection network with oriented anchor box. In: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, pp 7223–7230

Zhou X, Wang D, Krähenbühl P (2019) Objects as points. Preprint arXiv:1904.07850

Zhou X, Zhuo J, Krahenbuhl P (2019) Bottom-up object detection by grouping extreme and center points. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 850–859

Zhou Y, Tuzel O (2018) Voxelnet: end-to-end learning for point cloud based 3d object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4490–4499

Zhou Z, Pan T, Wu S, Chang H, Jenkins OC (2019) Glassloc: plenoptic grasp pose detection in transparent clutter. Preprint arXiv:1909.04269

Zhu A, Yang J, Zhao C, Xian K, Cao Z, Li X (2020) Lrf-net: learning local reference frames for 3d local shape description and matching. Preprint arXiv:2001.07832

Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2814–2821

Zou Z, Shi Z, Guo Y, Ye J (2019) Object detection in 20 years: a survey. Preprint arXiv:1905.05055

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.