Vision-based Robotic Grasping From Object Localization, Object Pose Estimation to Grasp Estimation for Parallel Grippers: A Review

Guoguang Du , Kai Wang , Shiguo Lian , Kaiyong Zhao
CloudMinds Technologies
george.du@cloudminds.com

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.

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 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.

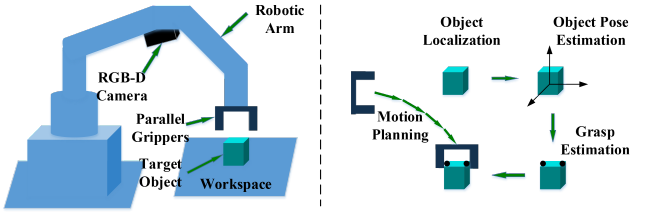

Figure 1: The grasp detection system. (Left)The robotic arm, equipped with one RGB-D camera and one parallel gripper, is to grasp the target object placed on a planar work surface. (Right)The grasp detection system involves target object localization, object pose estimation, and grasp estimation.

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; Zeng *et al.*, 2017b], which is a relatively simple situation. With the struggle of academia, dexterous grippers [Liu *et al.*, 2019b; 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

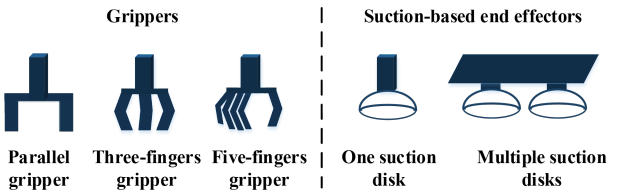

Figure 2: Different kinds of end effectors. (Left)Grippers. (Right)Suction-based end effectors. In this paper, we mainly consider parallel grippers.

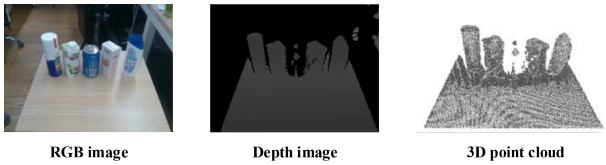

Figure 3: A RGB-D image. The depth image is transformed into 3D point cloud.

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. [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.*, 2019b; 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.*, 2019a; Mousavian *et al.*, 2019; Qin *et al.*, 2020; Zhao *et al.*, 2020a] 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.

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

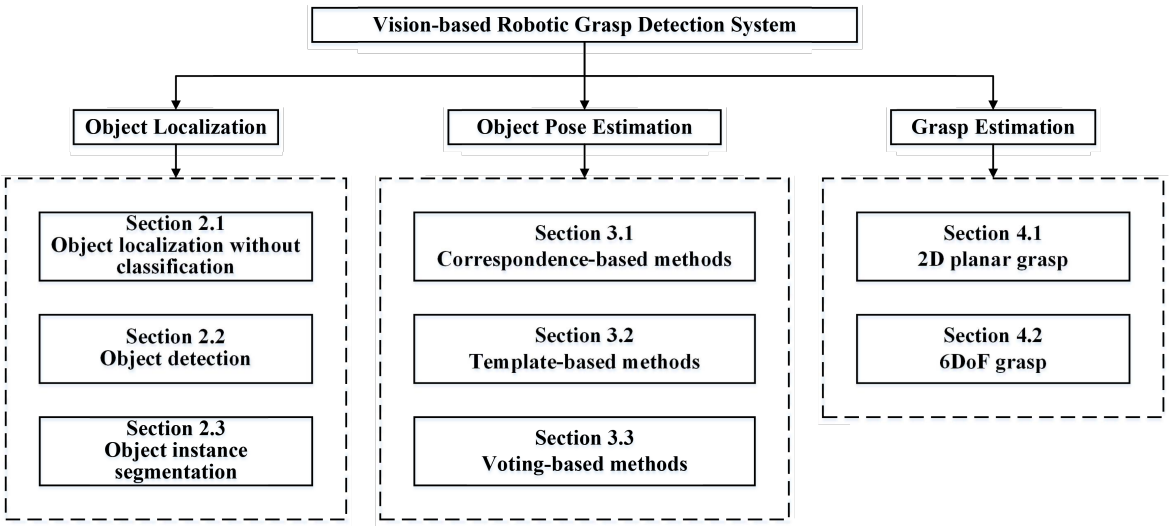

Figure 4: A taxonomy of tasks in vision-based robotic grasp detection system.

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.

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 Section 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.

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.

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 *et al.*, 1996]

Table 1: Methods of object localization without classification.

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

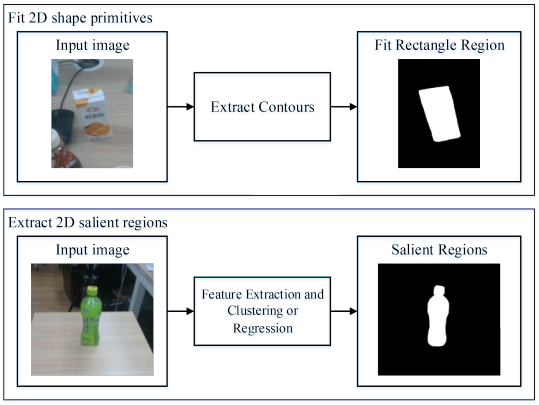

Figure 5: Typical functional flow-chart of 2D object localization without classification.

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.*, 2017b] 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. [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. [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 [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. [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. [Liu *et al.*, 2018a] 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; Sabour *et al.*, 2018], some capsule-based methods are proposed. Liu et al. [Liu *et al.*, 2019c] incorporated the part-object relationships in salient object detection, which is implemented by the Capsule Network. Qi et al. [Qi *et al.*, 2019b] 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.*, 2019e] 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.

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.

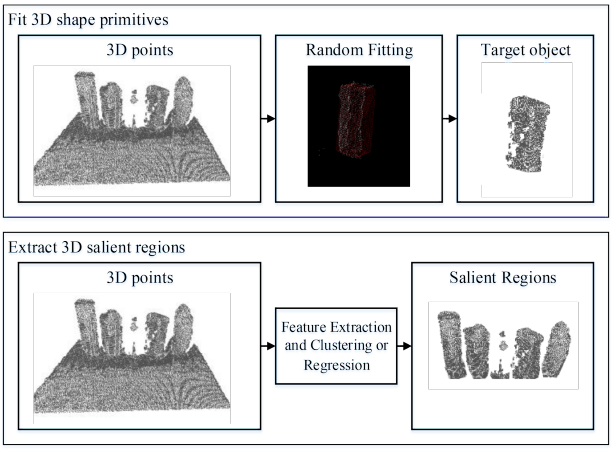

Figure 6: Typical functional flow-chart of 3D object localization without classification.

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.*, 2009b; 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 et al. [Jiang and Xiao, 2013] and Khan et al. [Khan *et al.*, 2015] explored the 3D structures in an indoor scene and estimated their geometry using cuboids. Rabbani et al. [Rabbani and Van Den Heuvel, 2005] presented an efficient Hough transform for automatic detection of cylinders in point clouds. Some methods [Rusu *et al.*, 2009b; Goron *et al.*, 2012] conduct primitive fitting after segmenting the scene. Rusu et al. [Rusu *et al.*, 2009b] 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. [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. [Peng *et al.*, 2014] proposed a simple fusion strategy which extends RGB-based saliency models by incorporating depth-induced saliency. Ren et al. [Ren *et al.*, 2015a] exploited the normalized depth prior and the global-context surface orientation prior for salient object detection. Qu et al. [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.*, 2019a; Chen and Li, 2019] utilized two-stream CNNs-based models with different fusion structures. Chen and Li [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. [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. [Pang *et al.*, 2020] proposed a hierarchical dynamic filtering network (HDFNet) and a hybrid enhanced loss. Li et al. [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. [Kim *et al.*, 2008] described a segmentation method for extracting salient regions in outdoor scenes using both 3D point clouds and RGB image. Bhatia et al. [Bhatia *et al.*, 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. [Rusu *et al.*, 2009b] and Goron et al. [Goron *et al.*, 2012], fitting 3D shape primitives has been successfully applied into robotic grasping tasks. In Zapata-Impata et al. [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.

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

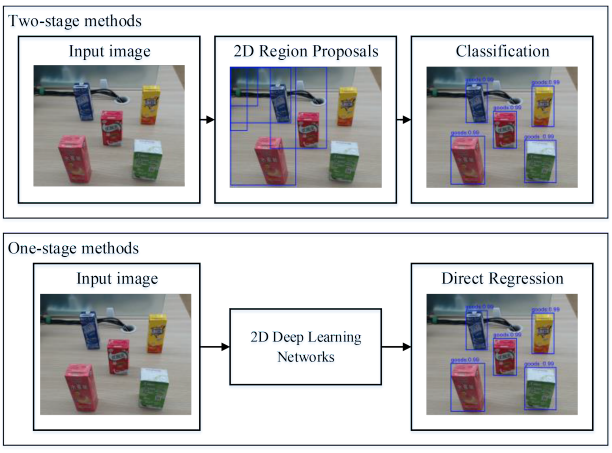

Figure 7: Typical functional flow-chart of 2D object detection.

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. [Erhan *et al.*, 2014] and Szegedy et al. [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.*, 2015b] is further proposed by replacing SS with the Region Proposal Network (RPN), 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.*, 2016c], FPN [Lin *et al.*, 2017a]. 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.*, 2017b] 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

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. [Erhan *et al.*, 2014], Szegedy et al. [Szegedy *et al.*, 2014], RCNN [Girshick *et al.*, 2014], Fast R-CNN [Girshick, 2015], Faster RCNN [Ren *et al.*, 2015b], R-FCN [Dai *et al.*, 2016c], FPN [Lin *et al.*, 2017a]	YOLO [Redmon *et al.*, 2016], SSD [Liu *et al.*, 2016], YOLOv2 [Redmon and Farhadi, 2017], RetinaNet [Lin *et al.*, 2017b], YOLOv3 [Redmon and Farhadi, 2018], FCOS [Tian *et al.*, 2019b], CornerNet [Law and Deng, 2018], ExtremeNet [Zhou *et al.*, 2019b], CenterNet [Zhou *et al.*, 2019a; 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.*, 2009a], 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.*, 2019b], Part-A^2 [Shi *et al.*, 2020b], PV-RCNN [Shi *et al.*, 2020a], PointRCNN [Shi *et al.*, 2019], STD [Yang *et al.*, 2019c], VoteNet [Qi *et al.*, 2019a], MLCVNet [Xie *et al.*, 2020c], H3DNet [Zhang *et al.*, 2020b], ImVoteNet [Qi *et al.*, 2020]	VoxelNet [Zhou and Tuzel, 2018], SECOND [Yan *et al.*, 2018b], PointPillars [Lang *et al.*, 2019], TANet [Liu *et al.*, 2020c], HVNet [Ye *et al.*, 2020], 3DSSD [Yang *et al.*, 2020b], Point-GNN [Shi and Rajkumar, 2020], DOPS [Najibi *et al.*, 2020], Associate-3Ddet [Du *et al.*, 2020]

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.*, 2019b], CornerNet [Law and Deng, 2018], ExtremeNet [Zhou *et al.*, 2019b], CenterNet [Zhou *et al.*, 2019a; 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.*, 2020a; Sultana *et al.*, 2020b].

Discussions The 2D object detection methods are widely used in 2D planar robotic grasping tasks. This part can refer to Section 4.1.

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.*, 2009a], 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 Section 3.1. 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, 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 fea-

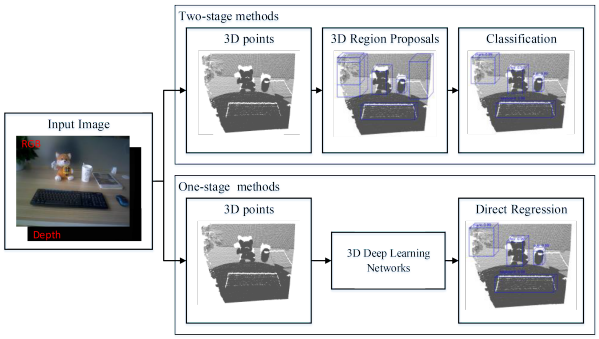

Figure 8: Typical functional flow-chart of 3D object detection.

tures 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.*, 2019b] 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.*, 2017a]. PointFusion [Xu *et al.*, 2018] utilized Faster R-CNN [Ren *et al.*, 2015b] 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.*, 2019b] 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.*, 2020b] 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.*, 2020a] 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.*, 2019c] 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.*, 2019a] 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.*, 2020c] proposed Multi-level Context VoteNet which considers the contextual information between the objects. H3DNet [Zhang *et al.*, 2020b] 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.*, 2019a], 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.*, 2018b] 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.*, 2020c] 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. [Yang *et al.*, 2020b] 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. Comprehen-

sive review about 3D object detection could refer to the survey [Guo *et al.*, 2020].

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.

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.

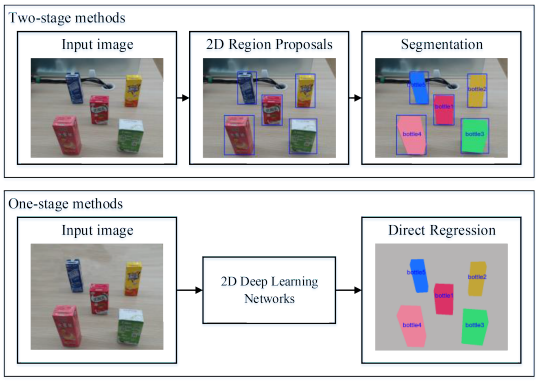

Figure 9: Typical functional flow-chart of 2D object instance segmentation.

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.*, 2016b] conducts instance segmentation via three networks, respectively differentiating instances, estimating masks, and categorizing objects. Path Aggregation Network (PANet) [Liu *et al.*, 2018b] 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.*, 2015b] 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.*, 2015b] and additionally produces semantic and instance center direction outputs. Chen et al. [Chen *et al.*, 2019b] 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 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.*, 2016a] 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.*, 2019c] uses structured 4D tensors to represent masks over a spatial domain and presents a framework to predict dense masks. YOLACT [Bolya *et al.*, 2019b] 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.*, 2019a]. PolarMask [Xie *et al.*, 2020b] 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.*, 2019f] 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.*, 2019b]) in the same vein with Mask R-CNN [He *et al.*, 2017]. BlendMask [Chen *et al.*, 2020b] also builds upon the FCOS [Tian *et al.*, 2019b] 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 [Sultana *et al.*, 2020a; 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. [Xie

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.*, 2016b], PANet [Liu *et al.*, 2018b], Mask R-CNN [He *et al.*, 2017], MaskLab [Chen *et al.*, 2018], HTC [Chen *et al.*, 2019b], PointRend [Kirillov *et al.*, 2020], FGN [Fan *et al.*, 2020]	DeepMask [Pinheiro *et al.*, 2015], SharpMask [Pinheiro *et al.*, 2016], InstanceFCN [Dai *et al.*, 2016a], FCIS [Li *et al.*, 2017], TensorMask [Chen *et al.*, 2019c], YOLACT [Bolya *et al.*, 2019b], YOLACT++ [Bolya *et al.*, 2019a], PolarMask [Xie *et al.*, 2020b], SOLO [Wang *et al.*, 2019f], CenterMask [Lee and Park, 2020], BlendMask [Chen *et al.*, 2020b]
3D instance segmentation	GSPN [Yi *et al.*, 2019], 3D-SIS [Hou *et al.*, 2019], 3D-MPA [Engelmann *et al.*, 2020]	SGPN [Wang *et al.*, 2018b], MASC [Liu and Furukawa, 2019], ASIS [Wang *et al.*, 2019g], JSIS3D [Pham *et al.*, 2019], JSNet [Zhao and Tao, 2020], 3D-BoNet [Yang *et al.*, 2019a], LiDARSeg [Zhang *et al.*, 2020a], OccuSeg [Han *et al.*, 2020]

et al., 2020a] separately leverage RGB and Depth for unseen object instance segmentation. Danielczuk et al. [Danielczuk *et al.*, 2019] segments unknown 3d objects from real depth images using Mask R-CNN [He *et al.*, 2017] trained on synthetic data.

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.

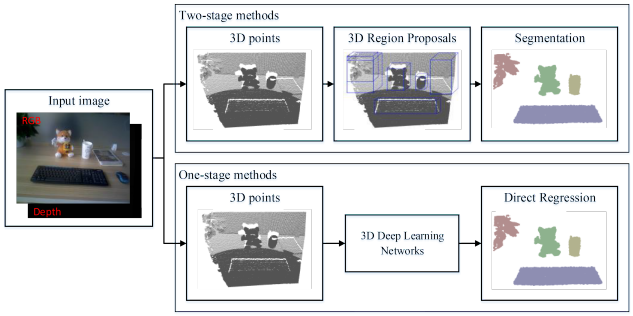

Figure 10: Typical functional flow-chart of 3D object instance segmentation.

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.*, 2018b] proposed the Similarity Group Proposal Network (SGPN) to 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 submanifold 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.*, 2019g] 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.*, 2019a] 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.*, 2020a] 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.

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.

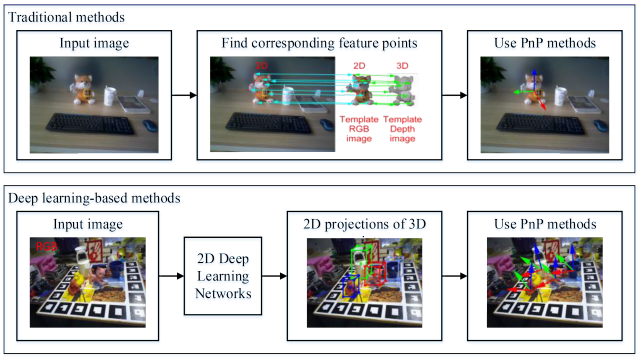

Figure 11: Typical functional flow-chart of 2D correspondence-based 6D object pose estimation methods. Data from the lineMod dataset [Hinterstoisser *et al.*, 2012].

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. [Yi *et al.*, 2016] presented a SIFT-like feature descriptor. Truong et al. [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 [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. [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. [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.*, 2020b] predicts object poses using 3D keypoints from stereo input, and is suitable for transparent objects. Hu et al. [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, including keypoints, edge vectors, and symmetry correspondences. Some other

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. [Tekin *et al.*, 2018], Crivellaro et al. [Crivellaro *et al.*, 2017], KeyPose [Liu *et al.*, 2020b], Hu et al. [Hu *et al.*, 2020], HybridPose [Song *et al.*, 2020], Hu et al. [Hu *et al.*, 2019], DPOD [Zakharov *et al.*, 2019], Pix2pose [Park *et al.*, 2019b], 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.*, 2009a], CVFH [Aldoma *et al.*, 2011], SHOT [Salti *et al.*, 2014]	3DMatch [Zeng *et al.*, 2017a], 3DFeat-Net [Yew and Lee, 2018], Gojcic et al. [Gojcic *et al.*, 2019], Yuan et al. [Yuan *et al.*, 2020], StickyPillars [Simon *et al.*, 2020]

methods predict 3D positions for all the pixels of the object. Hu et al. [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. Pix2pose [Park *et al.*, 2019b] 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.

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.*, 2009a], 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.*, 2016b]. 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. 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. [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. [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.

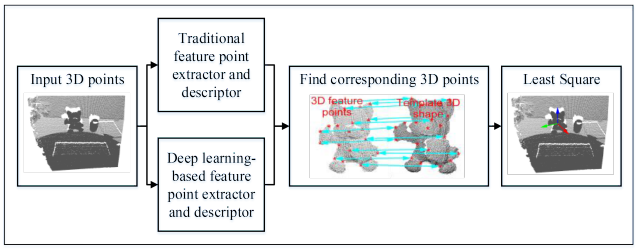

Figure 12: Typical functional flow-chart of 3D correspondence-based 6D object pose estimation methods.

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.

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. [Hodaň *et al.*, 2015]	AAE [Sundermeyer *et al.*, 2018], PoseCNN [Xiang *et al.*, 2018], SSD6D [Kehl *et al.*, 2017], Deep-6DPose [Do *et al.*, 2018a], Liu et al. [Liu *et al.*, 2019a], CDPN [Li *et al.*, 2019], Tian et al. [Tian *et al.*, 2020], NOCS [Wang *et al.*, 2019c], LatentFusion [Park *et al.*, 2020], Chen et al. [Chen *et al.*, 2020a]
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.*, 2019b], DCP [Wang and Solomon, 2019a], PointNetLK [Aoki *et al.*, 2019], PRNet [Wang and Solomon, 2019b], DeepICP [Lu *et al.*, 2019], Sarode et al. [Sarode *et al.*, 2019a], TEASER [Yang *et al.*, 2020a], DGR [Choy *et al.*, 2020], G2L-Net [Chen *et al.*, 2020c], Gao et al. [Gao *et al.*, 2020]

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 are regarded as the templates. Hinterstoisser et al. [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. [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. [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.*, 2018a; Liu *et al.*, 2019a] 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]. Yu et al. [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. [Kehl *et al.*, 2017] presented a similar method by making use of the SSD network. Do et al. [Do *et al.*, 2018a] 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

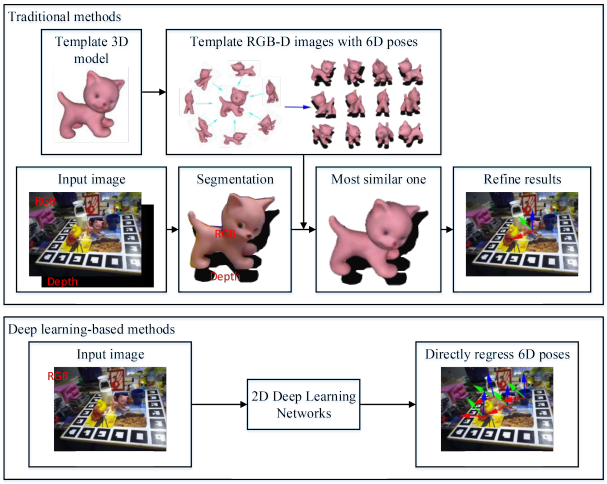

Figure 13: Typical functional flow-chart of 2D template-based 6D object pose estimation methods. Data from the lineMod dataset [Hinterstoisser *et al.*, 2012].

object poses without any post-refinements. Liu et al. [Liu *et al.*, 2019a] 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. [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.*, 2019c], LatentFusion [Park *et al.*, 2020] and Chen et al. [Chen *et al.*, 2020a] are the representative methods.

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 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].

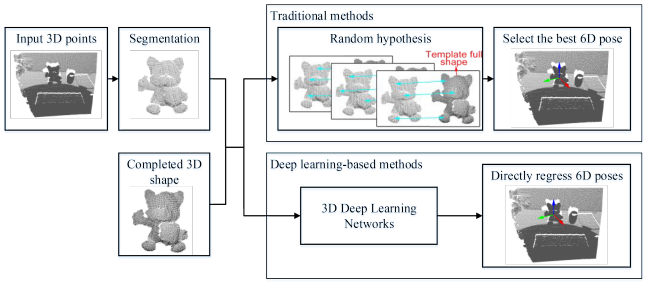

Figure 14: Typical functional flow-chart of 3D template-based 6D object pose estimation methods.

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.*, 2019b], DCP [Wang and Solomon, 2019a], PointNetLK [Aoki *et al.*, 2019], PRNet [Wang and Solomon, 2019b], DeepICP [Lu *et al.*, 2019], Sarode et al. [Sarode *et al.*, 2019a], TEASER [Yang *et al.*, 2020a] 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.*, 2020c; Gao *et al.*, 2020] that directly regress the 6D object pose from the partial point cloud. G2L-Net [Chen *et al.*, 2020c] 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. [Gao *et al.*, 2020] conducts 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.

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.

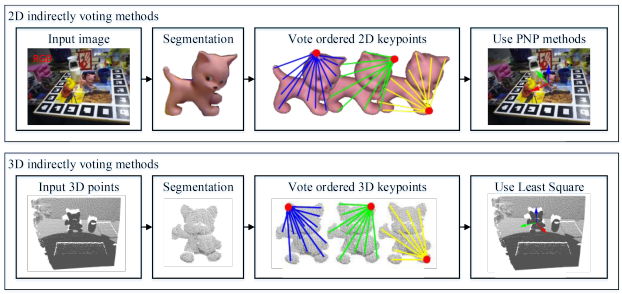

Figure 15: Typical functional flow-chart of indirect voting-based object pose estimation methods.

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. [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

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. [Yu *et al.*, 2020]	PVN3D [He *et al.*, 2020], YOLOff [Gonzalez *et al.*, 2020], 6-PACK [Wang *et al.*, 2019a]
Direct voting methods	Voting for template-based methods	Brachmann et al. [Brachmann *et al.*, 2014], Tejani et al. [Tejani *et al.*, 2014], Crivellaro et al. [Crivellaro *et al.*, 2017], PPF [Drost and Ilic, 2012]	DenseFusion [Wang *et al.*, 2019b], MoreFusion [Wada *et al.*, 2020]

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.*, 2019a] predicts 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.

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.

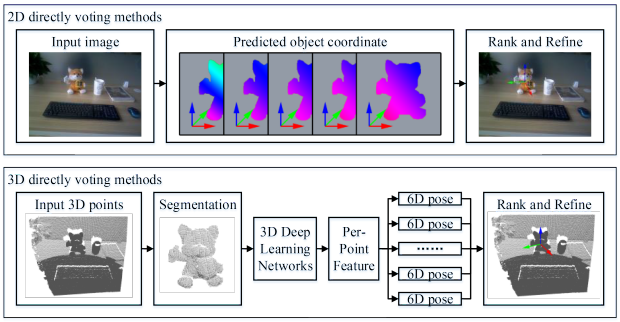

Figure 16: Typical functional flow-chart of direct voting-based 6D object pose estimation methods.

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 about the 6D object pose. Brachmann et al. [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. [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. [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. [Hodan *et al.*, 2018a] 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.*, 2019b] 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.

Datasets and evaluation metrics

There exist various benchmarks [Hodaň *et al.*, 2018b] 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.

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

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 $[\hat{\mathrm{R}}|\hat{\mathrm{T}}]$ and the ground truth pose $[\mathrm{R}|\mathrm{T}]$, ADD-S calculates the mean distance from each 3D model point transformed by $[\hat{\mathrm{R}}|\hat{\mathrm{T}}]$ to its closest point on the target model transformed by $[\mathrm{R}|\mathrm{T}]$.

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.

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 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.*, 2019c; Park *et al.*, 2020; Chen *et al.*, 2020a] 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.

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. [Domae *et al.*, 2014] presented a method that estimates graspability 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

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. [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.*, 2019b](per-pixel)	91.2	95.3
	DenseFusion [Wang *et al.*, 2019b](iterative)	93.1	96.8

Table 8: Accuracies of methods using ADD(-S) metric on LineMOD and Occlusion LineMOD dataset. Refine means methods such as ICP or DeepIM. IR is short for iterative refinement.

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. [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. [Hu *et al.*, 2019]	-	26.1
	Pix2pose [Park *et al.*, 2019b]	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. [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. [Tian *et al.*, 2020]	92.87	-
	MaskedFusion [Pereira and Alexandre, 2019]	97.3	-
Voting-based methods	Brachmann et al. [Brachmann *et al.*, 2016]	32.3	-
	Brachmann et al. [Brachmann *et al.*, 2016]+Refine	50.2	-
	PVNet [Peng *et al.*, 2019]	86.27	40.8
	DenseFusion [Wang *et al.*, 2019b](per-pixel)	86.2	
	DenseFusion [Wang *et al.*, 2019b](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^2GNet [Yu *et al.*, 2019b]	96.2	-
	P^2GNet [Yu *et al.*, 2019b]+Refine	97.4	-
	PointPoseNet [Hagelskjær and Buch, 2019]	96.3	52.6
	PointPoseNet [Hagelskjær and Buch, 2019]+Refine	-	75.1

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. [Domae *et al.*, 2014]	Zeng et al. [Zeng *et al.*, 2018], Mahler et al. [Mahler *et al.*, 2017], Cai et al. [Cai *et al.*, 2019], GG-CNN [Morrison *et al.*, 2018], MVP [Morrison *et al.*, 2019], Wang et al. [Wang *et al.*, 2019d]
Methods of evaluating oriented rectangles	Jiang et al. [Jiang *et al.*, 2011], Vohra et al. [Vohra *et al.*, 2019]	Lenz et al. [Lenz *et al.*, 2015], Pinto and Gupta [Pinto and Gupta, 2016], Park and Chun [Park and Chun, 2018], Redmon and Angelova [Redmon and Angelova, 2015], Zhang et al. [Zhang *et al.*, 2017], Kanan [Kumra and Kanan, 2017], Kumra et al. [Kumra *et al.*, 2019], Zhang et al. [Zhang *et al.*, 2018b], Guo et al. [Guo *et al.*, 2017a], Chu et al. [Chu *et al.*, 2018], Park et al. [Park *et al.*, 2018], Zhou et al. [Zhou *et al.*, 2018], Depierre et al. [Depierre *et al.*, 2020]

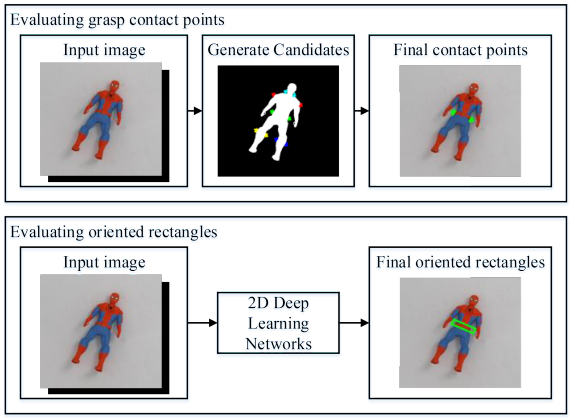

Figure 17: Typical functional flow-chart of 2D planar grasp methods. Data from the JACQUARD dataset [Depierre *et al.*, 2018].

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. [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 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.*, 2018b; 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. [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. [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. [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. [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. [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. [Wang *et al.*, 2019d] 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.

Methods of evaluating oriented rectangles

Jiang et al. [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. [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 rectan-

gles 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. [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 [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 [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 [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. [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 [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. [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. [Zhang *et al.*, 2018b] 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.*, 2015b; 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. [Guo *et al.*, 2017a] 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. [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. [Park *et al.*, 2018] proposed a rotation ensemble module (REM) for robotic grasp detection using convolutions that rotates network weights. Zhou et al. [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.*, 2017a; 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. [Depierre *et al.*, 2020] further extends Zhou et al. [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. [Guo *et al.*, 2016a] presented a shared convolutional neural network to conduct object discovery and grasp detection. Zhang et al. [Zhang *et al.*, 2018a] 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. [Zhang *et al.*, 2018a] 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. [Park *et al.*, 2019a] 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.

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

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.*, 2008b; Saxena *et al.*, 2008a]	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. [Jiang *et al.*, 2011]	227 x 227	60.50	58.30	50sec
Lenz et al. [Lenz *et al.*, 2015]	227 x 227	73.90	75.60	13.5sec
Morrison et al. [Morrison *et al.*, 2018]	300 x 300	78.56	-	7ms
Redmon et al. [Redmon and Angelova, 2015]	224 x 224	88.00	87.1	76ms
Zhang et al. [Zhang *et al.*, 2017]	224 x 224	88.90	88.20	117ms
Kumra et al. [Kumra and Kanan, 2017]	224 x 224	89.21	88.96	103ms
Chun et al. [Park and Chun, 2018]	400 x 400	89.60	-	23ms
Asif et al. [Asif *et al.*, 2018]	224 x 224	90.60	90.20	24ms
Wang et al. [Wang *et al.*, 2019d]	400 x 400	94.42	91.02	8ms
Chu et al. [Chu *et al.*, 2018]	227 x 227	96.00	96.10	120ms
Chun et al. [Park *et al.*, 2018]	360 x 360	96.60	95.40	20ms
Zhou et al. [Zhou *et al.*, 2018]	320 x 320	97.74	96.61	118ms
Park et al. [Park *et al.*, 2019a]	360 x 360	98.6	97.2	16ms

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 $J(A, B) = |A \cap B|/|A \cup B|$ 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.

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.*, 2019a] 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 *et al.*, 2020a; 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.

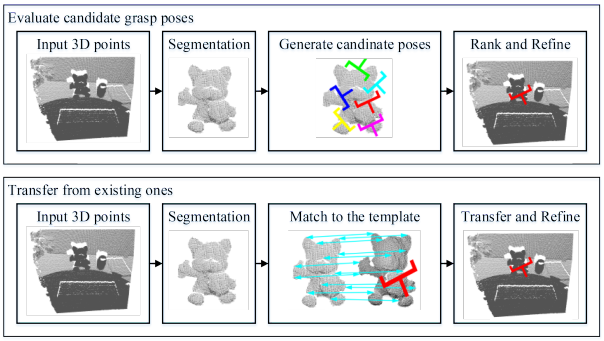

Figure 18: Typical functional flow-chart of 6DoF grasp methods based on the partial point cloud.

Bohg and Kragic [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 et al. [Pas and Platt, 2015] first used a geometrically necessary condition to sample a large

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 [Bohg and Kragic, 2010], Pas et al. [Pas and Platt, 2015], Zapata-Impata et al. [Zapata-Impata *et al.*, 2019]	GPD [ten Pas *et al.*, 2017], PointnetGPD [Liang *et al.*, 2019a], 6-DoF GraspNet [Mousavian *et al.*, 2019], S^4G [Qin *et al.*, 2020], REGNet [Zhao *et al.*, 2020a]
	Transfer grasps from existing ones	Andrew et al. [Miller *et al.*, 2003], Nikandrova and Kyrki [Nikandrova and Kyrki, 2015], Vahrenkamp et al. [Vahrenkamp *et al.*, 2016]	Tian et al. [Tian *et al.*, 2019a], 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. [Zeng *et al.*, 2017b]	Billings and Roberson [Billings and Johnson-Roberson, 2018]
	Conduct shape completion	Miller et al. [Miller *et al.*, 2003]	Varley et al. [Varley *et al.*, 2017], Lundell et al. [Lundell *et al.*, 2019], Watkins-Valls et al. [Watkins-Valls *et al.*, 2019], Merwe et al. [Van der Merwe *et al.*, 2019], Wang et al. [Wang *et al.*, 2018a], Yan et al. [Yan *et al.*, 2018a], Yan et al. [Yan *et al.*, 2019], Tosun et al. [Tosun *et al.*, 2020], kPAM-SC [Gao and Tedrake, 2019], ClearGrasp [Sajjan *et al.*, 2019]

set of high quality grasp hypotheses, which will be classified using the notion of an antipodal grasp. Zapata-Impata et al. [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 image. Each candidate is evaluated to obtain a score using a four-layer convolutional neural network finally. Lou et al. [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.*, 2019a] randomly samples candidate grasps, and evaluates the grasp quality by direct point cloud analysis with the 3D deep neural network PointNet [Qi *et al.*, 2017a]. 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. [Mousavian *et al.*, 2019] proposed an algorithm called 6-DoF GraspNet, which samples grasp proposals using a variational autoencoder and refines the sampled grasps using a grasp evaluator model. Pointnet++ [Qi *et al.*, 2017b] is used to generate and evaluate grasps. Murali et al. [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. [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.*, 2017b] 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. [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. [Zhao *et al.*, 2020a] 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.*, 2019a] and S^4G [Qin *et al.*, 2020]. Fang et al. [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 ap-

proaching 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. Andrew et al. [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 [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. [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. [Tian *et al.*, 2019a] 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. [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. [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 *et al.*, 2020a]. 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.

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.*, 2019a]	81.48	1965.60
S^4G [Qin *et al.*, 2020]	77.63	679.04
REGNet [Zhao *et al.*, 2020a]	92.47	686.31

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.*, 2019b] that learn grasps from demonstration usually utilize this kind of methods.

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 Section 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

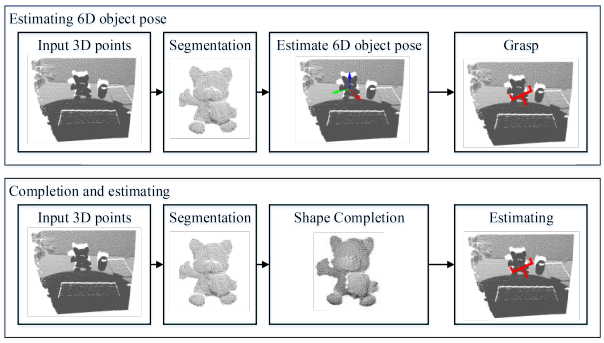

Figure 19: Typical functional flow-chart of 6DoF grasp methods based on the complete shape.

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 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. Andrew and Peter [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. [Xue *et al.*, 2009] implemented a grasping planning system based on GraspIt! to plan high-quality grasps. León et al. [León *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. [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 [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. [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 $< 5^{\circ}$ 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.*, 2019b] 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.*, 2018a; Yan *et al.*, 2018a; Yan *et al.*, 2019; Gao and Tedrake, 2019; Sajjan *et al.*, 2019] utilize the RGB-D images. Many of them [Wang *et al.*, 2018a; Watkins-Valls *et al.*, 2019] also combine tactile information for better prediction.

Varley et al. [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. [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. Merwe et al. [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. [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. [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. [Wang *et al.*, 2018a] 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. [Yan *et al.*, 2018a] 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. [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 [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. [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 [Zeng *et al.*, 2017b; 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.*, 2020b] 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.*, 2019c; Park *et al.*, 2020; Wang *et al.*, 2019a; Chen *et al.*, 2020a] 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.*, 2019a; 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.*, 2019c] was proposed for grasp pose detection of transparent objects in transparent clutter using plenoptic sensing. KeyPose [Liu *et al.*, 2020b] 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 *et al.*, 2019] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[Aldoma *et al.*, 2011] Aitor Aldoma, Markus Vincze, Nico Blodow, David Gossow, Suat Gedikli, Radu Bogdan Rusu, and Gary Bradski. Cad-model recognition and 6dof pose estimation using 3d cues. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 585–592. IEEE, 2011.

[Aoki *et al.*, 2019] Yasuhiro Aoki, Hunter Goforth, Rangaprasad Arun Srivatsan, and Simon Lucey. Pointnetlk: Robust & efficient point cloud registration using pointnet. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7163–7172, 2019.

[Ardón *et al.*, 2019] Paola Ardón, Èric Pairet, Ronald PA Petrick, Subramanian Ramamoorthy, and Katrin S Lohan. Learning grasp affordance reasoning through semantic relations. *IEEE Robotics and Automation Letters*, 4(4):4571–4578, 2019.

[Asif *et al.*, 2018] Umar Asif, Jianbin Tang, and Stefan Harrer. Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices. In *IJCAI*, pages 4875–4882, 2018.

[Bay *et al.*, 2006] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[Bellekens *et al.*, 2014] Ben Bellekens, Vincent Spruyt, Rafael Berkvens, and Maarten Weyn. 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*, pages 8–13, 2014.

[Belongie *et al.*, 2002] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522, 2002.

[Berscheid *et al.*, 2019] Lars Berscheid, Pascal Meißner, and Torsten Kröger. Robot learning of shifting objects for grasping in cluttered environments. *arXiv preprint arXiv:1907.11035*, 2019.

[Besl and McKay, 1992] Paul J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 14(2):239–256, February 1992.

[Bhatia *et al.*, 2013] Shashank Bhatia, Stephan K Chalup, et al. Segmenting salient objects in 3d point clouds of indoor scenes using geodesic distances. *Journal of Signal and Information Processing*, 4(03):102, 2013.

[Billings and Johnson-Roberson, 2018] Gideon Billings and Matthew Johnson-Roberson. Silhonet: An RGB method for 3d object pose estimation and grasp planning. *CoRR*, abs/1809.06893, 2018.

[Blomqvist *et al.*, 2020] Kenneth Blomqvist, Michel Breyer, Andrei Cramariuc, Julian Förster, Margarita Grinvald, Florian Tschopp, Jen Jen Chung, Lionel Ott, Juan Nieto, and Roland Siegwart. Go fetch: Mobile manipulation in unstructured environments. *arXiv preprint arXiv:2004.00899*, 2020.

[Bochkovskiy *et al.*, 2020] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao. Yolov4: Optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934*, 2020.

[Bohg and Kragic, 2010] Jeannette Bohg and Danica Kragic. Learning grasping points with shape context. *Robotics and Autonomous Systems*, 58(4):362–377, 2010.

[Bohg *et al.*, 2014] J. Bohg, A. Morales, T. Asfour, and D. Kragic. Data-driven grasp synthesis: A survey. *IEEE Transactions on Robotics*, 30(2):289–309, April 2014.

[Bolya *et al.*, 2019a] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact++: Better real-time in-

stance segmentation. *arXiv preprint arXiv:1912.06218*, 2019.

[Bolya *et al.*, 2019b] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: real-time instance segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9157–9166, 2019.

[Borji *et al.*, 2019] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. Salient object detection: A survey. *Computational Visual Media*, 5(2):117–150, 2019.

[Borst *et al.*, 2003] Christoph Borst, Max Fischer, and Gerd Hirzinger. Grasping the dice by dicing the grasp. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 4, pages 3692–3697. IEEE, 2003.

[Bousmalis *et al.*, 2018] Konstantinos Bousmalis, Alex Irpan, Paul Wohlhart, Yunfei Bai, Matthew Kelcey, Mrinal Kalakrishnan, Laura Downs, Julian Ibarz, Peter Pastor, Kurt Konolige, et al. Using simulation and domain adaptation to improve efficiency of deep robotic grasping. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4243–4250. IEEE, 2018.

[Brachmann *et al.*, 2014] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014.

[Brachmann *et al.*, 2016] Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. 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*, pages 3364–3372, 2016.

[Bradski and Kaehler, 2008] Gary Bradski and Adrian Kaehler. *Learning OpenCV: Computer vision with the OpenCV library*. " O'Reilly Media, Inc.", 2008.

[Cai *et al.*, 2019] Junhao Cai, Hui Cheng, Zhanpeng Zhang, and Jingcheng Su. Metagrasp: Data efficient grasping by affordance interpreter network. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4960–4966. IEEE, 2019.

[Caldera *et al.*, 2018] Shehan Caldera, Alexander Rassau, and Douglas Chai. Review of deep learning methods in robotic grasp detection. *Multimodal Technologies and Interaction*, 2(3):57, 2018.

[Castro *et al.*, 2020] Pedro Castro, Anil Armagan, and Tae-Kyun Kim. Accurate 6d object pose estimation by pose conditioned mesh reconstruction. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4147–4151. IEEE, 2020.

[Chen and Burdick, 1993] I-Ming Chen and Joel W Burdick. Finding antipodal point grasps on irregularly shaped objects. *IEEE transactions on Robotics and Automation*, 9(4):507–512, 1993.

[Chen and Li, 2018] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for rgb-d salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3051–3060, 2018.

[Chen and Li, 2019] Hao Chen and Youfu Li. Cnn-based rgb-d salient object detection: Learn, select and fuse. *arXiv preprint arXiv:1909.09309*, 2019.

[Chen *et al.*, 2017] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[Chen *et al.*, 2018] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4013–4022, 2018.

[Chen *et al.*, 2019a] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. *Pattern Recognition*, 86:376–385, 2019.

[Chen *et al.*, 2019b] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4974–4983, 2019.

[Chen *et al.*, 2019c] Xinlei Chen, Ross Girshick, Kaiming He, and Piotr Dollár. Tensormask: A foundation for dense object segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2061–2069, 2019.

[Chen *et al.*, 2020a] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. 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*, pages 11973–11982, 2020.

[Chen *et al.*, 2020b] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8573–8581, 2020.

[Chen *et al.*, 2020c] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, and Ales Leonardis. 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*, pages 4233–4242, 2020.

[Cheng *et al.*, 2014] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2014.

[Choy *et al.*, 2020] Christopher Choy, Wei Dong, and Vladlen Koltun. Deep global registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2514–2523, 2020.

[Chu *et al.*, 2018] Fu-Jen Chu, Ruinian Xu, and Patricio A Vela. Real-world multiobject, multigrasp detection. *IEEE Robotics and Automation Letters*, 3(4):3355–3362, 2018.

[Chu *et al.*, 2019] Fu-Jen Chu, Ruinian Xu, and Patricio A Vela. Detecting robotic affordances on novel objects with regional attention and attributes. *arXiv preprint arXiv:1909.05770*, 2019.

[Crivellaro *et al.*, 2017] Alberto Crivellaro, Mahdi Rad, Yannick Verdie, Kwang Moo Yi, Pascal Fua, and Vincent Lepetit. Robust 3d object tracking from monocular images using stable parts. *IEEE transactions on pattern analysis and machine intelligence*, 40(6):1465–1479, 2017.

[Dai *et al.*, 2016a] Jifeng Dai, Kaiming He, Yi Li, Shaoqing Ren, and Jian Sun. Instance-sensitive fully convolutional networks. In *European Conference on Computer Vision*, pages 534–549. Springer, 2016.

[Dai *et al.*, 2016b] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3150–3158, 2016.

[Dai *et al.*, 2016c] Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems*, pages 379–387, 2016.

[Dai *et al.*, 2017] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017.

[Danielczuk *et al.*, 2019] Michael Danielczuk, Matthew Matl, Saurabh Gupta, Andrew Li, Andrew Lee, Jeffrey Mahler, and Ken Goldberg. 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)*, pages 7283–7290. IEEE, 2019.

[Deng *et al.*, 2020] Xinke Deng, Yu Xiang, Arsalan Mousavian, Clemens Eppner, Timothy Bretl, and Dieter Fox. Self-supervised 6d object pose estimation for robot manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[Depierre *et al.*, 2018] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Jacquard: A large scale dataset for robotic grasp detection. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3511–3516. IEEE, 2018.

[Depierre *et al.*, 2020] Amaury Depierre, Emmanuel Dellandréa, and Liming Chen. Optimizing correlated graspability score and grasp regression for better grasp prediction. *arXiv preprint arXiv:2002.00872*, 2020.

[DeTone *et al.*, 2018] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 224–236, 2018.

[Ding *et al.*, 2001] Dan Ding, Yun-Hui Liu, and Michael Yu Wang. On computing immobilizing grasps of 3-d curved objects. In *IEEE International Symposium on Computational Intelligence in Robotics and Automation*, pages 11–16. IEEE, 2001.

[Do *et al.*, 2018a] Thanh-Toan Do, Ming Cai, Trung Pham, and Ian Reid. Deep-6dpose: recovering 6d object pose from a single rgb image. *arXiv preprint arXiv:1802.10367*, 2018.

[Do *et al.*, 2018b] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1–5. IEEE, 2018.

[Domae *et al.*, 2014] Yukiyasu Domae, Haruhisa Okuda, Yuichi Taguchi, Kazuhiko Sumi, and Takashi Hirai. Fast graspability evaluation on single depth maps for bin picking with general grippers. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1997–2004. IEEE, 2014.

[Dong *et al.*, 2020] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10519–10528, 2020.

[Douglas and Peucker, 1973] David H Douglas and Thomas K Peucker. Algorithms for the reduction of the number of points required to represent a digitized line or its caricature. *Cartographica: the international journal for geographic information and geovisualization*, 10(2):112–122, 1973.

[Drost and Ilic, 2012] B. Drost and S. Ilic. 3d object detection and localization using multimodal point pair features. In *International Conference on 3D Imaging, Modeling, Processing, Visualization Transmission*, pages 9–16, Oct 2012.

[Drost *et al.*, 2010] B. Drost, M. Ulrich, N. Navab, and S. Ilic. Model globally, match locally: Efficient and robust 3d object recognition. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 998–1005, June 2010.

[Du *et al.*, 2020] Liang Du, Xiaoqing Ye, Xiao Tan, Jianfeng Feng, Zhenbo Xu, Errui Ding, and Shilei Wen. 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*, pages 13329–13338, 2020.

[Duan *et al.*, 2019] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet:

Keypoint triplets for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6569–6578, 2019.

[Engelmann *et al.*, 2020] Francis Engelmann, Martin Bokeloh, Alireza Fathi, Bastian Leibe, and Matthias Nießner. 3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9031–9040, 2020.

[Erhan *et al.*, 2014] Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.

[Falco *et al.*, 2019] Pietro Falco, Shuang Lu, Ciro Natale, Salvatore Pirozzi, and Dongheui Lee. A transfer learning approach to cross-modal object recognition: From visual observation to robotic haptic exploration. *IEEE Transactions on Robotics*, 35(4):987–998, 2019.

[Fan and Tomizuka, 2019] Yongxiang Fan and Masayoshi Tomizuka. Efficient grasp planning and execution with multifingered hands by surface fitting. *IEEE Robotics and Automation Letters*, 4(4):3995–4002, 2019.

[Fan *et al.*, 2020] Zhibo Fan, Jin-Gang Yu, Zhihao Liang, Jiarong Ou, Changxin Gao, Gui-Song Xia, and Yuanqing Li. Fgn: Fully guided network for few-shot instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9172–9181, 2020.

[Fang *et al.*, 2018] Kuan Fang, Yunfei Bai, Stefan Hinterstoisser, Silvio Savarese, and Mrinal Kalakrishnan. Multi-task domain adaptation for deep learning of instance grasping from simulation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3516–3523. IEEE, 2018.

[Fang *et al.*, 2020] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020.

[Fischler and Bolles, 1981] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.

[Fitzgibbon *et al.*, 1996] Andrew W Fitzgibbon, Robert B Fisher, et al. *A buyer's guide to conic fitting*. University of Edinburgh, Department of Artificial Intelligence, 1996.

[Florence *et al.*, 2018] Peter R Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *arXiv preprint arXiv:1806.08756*, 2018.

[Frome *et al.*, 2004] Andrea Frome, Daniel Huber, Ravi Kolluri, Thomas Bülow, and Jitendra Malik. Recognizing objects in range data using regional point descriptors. In *European conference on computer vision*, pages 224–237. Springer, 2004.

[Gao and Tedrake, 2019] Wei Gao and Russ Tedrake. kpam-sc: Generalizable manipulation planning using keypoint affordance and shape completion. *arXiv preprint arXiv:1909.06980*, 2019.

[Gao *et al.*, 2020] Ge Gao, Mikko Lauri, Yulong Wang, Xiaolin Hu, Jianwei Zhang, and Simone Frintrop. 6d object pose regression via supervised learning on point clouds. *arXiv preprint arXiv:2001.08942*, 2020.

[Girshick *et al.*, 2014] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '14, pages 580–587, 2014.

[Girshick, 2015] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[Gojcic *et al.*, 2019] Zan Gojcic, Caifa Zhou, Jan D Wegner, and Andreas Wieser. The perfect match: 3d point cloud matching with smoothed densities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5545–5554, 2019.

[Gonzalez *et al.*, 2020] Mathieu Gonzalez, Amine Kacete, Albert Murienne, and Eric Marchand. Yoloff: You only learn offsets for robust 6dof object pose estimation. *arXiv preprint arXiv:2002.00911*, 2020.

[Gordo *et al.*, 2016] Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pages 241–257. Springer, 2016.

[Goron *et al.*, 2012] Lucian Cosmin Goron, Zoltan-Csaba Marton, Gheorghe Lazea, and Michael Beetz. Robustly segmenting cylindrical and box-like objects in cluttered scenes using depth cameras. In *ROBOTIK 2012; 7th German Conference on Robotics*, pages 1–6. VDE, 2012.

[Graham and van der Maaten, 2017] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017.

[Graham *et al.*, 2018] Benjamin Graham, Martin Engelcke, and Laurens van der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9224–9232, 2018.

[Guo *et al.*, 2016a] Di Guo, Tao Kong, Fuchun Sun, and Huaping Liu. Object discovery and grasp detection with a shared convolutional neural network. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 2038–2043. IEEE, 2016.

[Guo *et al.*, 2016b] Yulan Guo, Mohammed Bennamoun, Ferdous Sohel, Min Lu, Jianwei Wan, and Ngai Ming Kwok. A comprehensive performance evaluation of 3d local feature descriptors. *International Journal of Computer Vision*, 116(1):66–89, 2016.

[Guo *et al.*, 2017a] Di Guo, Fuchun Sun, Huaping Liu, Tao Kong, Bin Fang, and Ning Xi. A hybrid deep architecture for robotic grasp detection. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1609–1614. IEEE, 2017.

[Guo *et al.*, 2017b] Fang Guo, Wenguan Wang, Jianbing Shen, Ling Shao, Jian Yang, Dacheng Tao, and Yuan Yan Tang. Video saliency detection using object proposals. *IEEE transactions on cybernetics*, 48(11):3159–3170, 2017.

[Guo *et al.*, 2020] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Hafiz and Bhat, 2020] Abdul Mueed Hafiz and Ghulam Mohiuddin Bhat. A survey on instance segmentation: state of the art. *International Journal of Multimedia Information Retrieval*, 9(3):171–189, 2020.

[Hagelskjær and Buch, 2019] Frederik Hagelskjær and Anders Glent Buch. Pointposenet: Accurate object detection and 6 dof pose estimation in point clouds. *arXiv preprint arXiv:1912.09057*, 2019.

[Han *et al.*, 2018] Junwei Han, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1):84–100, 2018.

[Han *et al.*, 2020] Lei Han, Tian Zheng, Lan Xu, and Lu Fang. Occuseg: Occupancy-aware 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2940–2949, 2020.

[Hariharan *et al.*, 2014] Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. Simultaneous detection and segmentation. In *European Conference on Computer Vision*, pages 297–312. Springer, 2014.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask r-cnn. *IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.

[He *et al.*, 2020] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. 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*, pages 11632–11641, 2020.

[Hinterstoisser *et al.*, 2012] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012.

[Hinton *et al.*, 2011] Geoffrey E Hinton, Alex Krizhevsky, and Sida D Wang. Transforming auto-encoders. In *International conference on artificial neural networks*, pages 44–51. Springer, 2011.

[Hodaň *et al.*, 2015] Tomáš Hodaň, Xenophon Zabulis, Manolis Lourakis, Štěpán Obdržálek, and Jiří Matas. 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)*, pages 4421–4428. IEEE, 2015.

[Hodaň *et al.*, 2017] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

[Hodan *et al.*, 2018a] Tomas Hodan, Rigas Kouskouridas, Tae-Kyun Kim, Federico Tombari, Kostas E. Bekris, Bertram Drost, Thibault Groueix, Krzysztof Walas, Vincent Lepetit, Ales Leonardis, Carsten Steger, Frank Michel, Caner Sahin, Carsten Rother, and Jiri Matas. A summary of the 4th international workshop on recovering 6d object pose. *CoRR*, abs/1810.03758, 2018.

[Hodaň *et al.*, 2018b] Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: benchmark for 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018.

[Hodan *et al.*, 2020] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11703–11712, 2020.

[Hogan *et al.*, 2020] Francois R Hogan, Jose Ballester, Siyuan Dong, and Alberto Rodriguez. Tactile dexterity: Manipulation primitives with tactile feedback. *arXiv preprint arXiv:2002.03236*, 2020.

[Hou *et al.*, 2017] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017.

[Hou *et al.*, 2019] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4421–4430, 2019.

[Hu *et al.*, 2019] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3385–3394, 2019.

[Hu *et al.*, 2020] Yinlin Hu, Pascal Fua, Wei Wang, and Mathieu Salzmann. Single-stage 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2930–2939, 2020.

[Jiang and Xiao, 2013] Hao Jiang and Jianxiong Xiao. A linear approach to matching cuboids in rgbd images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2171–2178, 2013.

[Jiang *et al.*, 2011] Yun Jiang, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *IEEE International Conference on Robotics and Automation*, pages 3304–3311. IEEE, 2011.

[Jiang *et al.*, 2013] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2083–2090, 2013.

[Johnson, 1997] Andrew E Johnson. Spin-images: a representation for 3-d surface matching. 1997.

[Kaiser *et al.*, 2019] Adrien Kaiser, Jose Alonso Ybanez Zepeda, and Tamy Boubekeur. A survey of simple geometric primitives detection methods for captured 3d data. In *Computer Graphics Forum*, volume 38, pages 167–196. Wiley Online Library, 2019.

[Kehl *et al.*, 2017] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1529, 2017.

[Khan *et al.*, 2015] Salman H Khan, Xuming He, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. Separating objects and clutter in indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4603–4611, 2015.

[Kim *et al.*, 2008] Gunhee Kim, Daniel Huber, and Martial Hebert. Segmentation of salient regions in outdoor scenes using imagery and 3-d data. In *2008 IEEE Workshop on Applications of Computer Vision*, pages 1–8. IEEE, 2008.

[Kirillov *et al.*, 2020] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9799–9808, 2020.

[Kirkpatrick *et al.*, 1992] David Kirkpatrick, Bhubaneswar Mishra, and Chee-Keng Yap. Quantitative steinitz's theorems with applications to multifingered grasping. *Discrete & Computational Geometry*, 7(3):295–318, 1992.

[Krizhevsky *et al.*, 2012] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, 2012.

[Kumra and Kanan, 2017] Sulabh Kumra and Christopher Kanan. Robotic grasp detection using deep convolutional neural networks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 769–776. IEEE, 2017.

[Kumra *et al.*, 2019] Sulabh Kumra, Shirin Joshi, and Ferat Sahin. Antipodal robotic grasping using generative residual convolutional neural network. *arXiv preprint arXiv:1909.04810*, 2019.

[Lang *et al.*, 2019] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12697–12705, 2019.

[Law and Deng, 2018] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 734–750, 2018.

[Lee and Park, 2020] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13906–13915, 2020.

[Lee *et al.*, 2019] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. 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)*, pages 8943–8950. IEEE, 2019.

[Lenz *et al.*, 2015] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[León *et al.*, 2010] Beatriz León, Stefan Ulbrich, Rosen Diankov, Gustavo Puche, Markus Przybylski, Antonio Morales, Tamim Asfour, Sami Moisio, Jeannette Bohg, James Kuffner, and Rüdiger Dillmann. Opengrasp: A toolkit for robot grasping simulation. In Noriaki Ando, Stephen Balakirsky, Thomas Hemker, Monica Reggiani, and Oskar von Stryk, editors, *Simulation, Modeling, and Programming for Autonomous Robots*, pages 109–120, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[Lepetit *et al.*, 2005] Vincent Lepetit, Pascal Fua, et al. Monocular model-based 3d tracking of rigid objects: A survey. *Foundations and Trends® in Computer Graphics and Vision*, 1(1):1–89, 2005.

[Lepetit *et al.*, 2009] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Epnp: An accurate o(n) solution to the pnp problem. *IJCV*, 81(2):155–166, February 2009.

[Li *et al.*, 2017] Yi Li, Haozhi Qi, Jifeng Dai, Xiangyang Ji, and Yichen Wei. Fully convolutional instance-aware semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2359–2367, 2017.

[Li *et al.*, 2019] Zhigang Li, Gu Wang, and Xiangyang Ji. 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*, pages 7678–7687, 2019.

[Li *et al.*, 2020] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for rgb-d salient object detection. 2020.

[Liang *et al.*, 2019a] Hongzhuo Liang, Xiaojian Ma, Shuang Li, Michael Görner, Song Tang, Bin Fang, Fuchun Sun, and Jianwei Zhang. Pointnetgpd: Detecting grasp configurations from point sets. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3629–3635. IEEE, 2019.

[Liang *et al.*, 2019b] Ming Liang, Bin Yang, Yun Chen, Rui Hu, and Raquel Urtasun. Multi-task multi-sensor fusion for 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7345–7353, 2019.

[Lin *et al.*, 2017a] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[Lin *et al.*, 2017b] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[Liu and Furukawa, 2019] Chen Liu and Yasutaka Furukawa. Masc: multi-scale affinity with sparse convolution for 3d instance segmentation. *arXiv preprint arXiv:1902.04478*, 2019.

[Liu and Han, 2016] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016.

[Liu *et al.*, 2016] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.

[Liu *et al.*, 2018a] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018.

[Liu *et al.*, 2018b] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8759–8768, 2018.

[Liu *et al.*, 2019a] Fuchang Liu, Pengfei Fang, Zhengwei Yao, Ran Fan, Zhigeng Pan, Weiguo Sheng, and Huansong Yang. Recovering 6d object pose from rgb indoor image based on two-stage detection network with multi-task loss. *Neurocomputing*, 337:15–23, 2019.

[Liu *et al.*, 2019b] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. *arXiv preprint arXiv:1903.00425*, 2019.

[Liu *et al.*, 2019c] Yi Liu, Qiang Zhang, Dingwen Zhang, and Jungong Han. Employing deep part-object relationships for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1232–1241, 2019.

[Liu *et al.*, 2020a] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128(2):261–318, 2020.

[Liu *et al.*, 2020b] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11602–11610, 2020.

[Liu *et al.*, 2020c] Zhe Liu, Xin Zhao, Tengteng Huang, Ruolan Hu, Yu Zhou, and Xiang Bai. Tanet: Robust 3d object detection from point clouds with triple attention. In *AAAI*, pages 11677–11684, 2020.

[Long *et al.*, 2015] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[Lou *et al.*, 2019] Xibai Lou, Yang Yang, and Changhyun Choi. Learning to generate 6-dof grasp poses with reachability awareness. *arXiv preprint arXiv:1910.06404*, 2019.

[Lowe, 1999] David G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ICCV '99, pages 1150–, 1999.

[Lu *et al.*, 2019] Weixin Lu, Guowei Wan, Yao Zhou, Xiangyu Fu, Pengfei Yuan, and Shiyu Song. Deepicp: An end-to-end deep neural network for 3d point cloud registration. *arXiv preprint arXiv:1905.04153*, 2019.

[Lundell *et al.*, 2019] Jens Lundell, Francesco Verdoja, and Ville Kyrki. Robust grasp planning over uncertain shape completions. *arXiv preprint arXiv:1903.00645*, 2019.

[Luo *et al.*, 2020] Tiange Luo, Kaichun Mo, Zhiao Huang, Jiarui Xu, Siyu Hu, Liwei Wang, and Hao Su. Learning to group: A bottom-up framework for 3d part discovery in unseen categories. In *International Conference on Learning Representations*, 2020.

[Mahajan *et al.*, 2020] Mridul Mahajan, Tryambak Bhattacharjee, Arya Krishnan, Priya Shukla, and GC Nandi. Semi-supervised grasp detection by representation learning in a vector quantized latent space. *arXiv preprint arXiv:2001.08477*, 2020.

[Mahler *et al.*, 2017] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *CoRR*, abs/1703.09312, 2017.

[Malisiewicz *et al.*, 2011] Tomasz Malisiewicz, Abhinav Gupta, and Alexei A Efros. Ensemble of exemplar-svms for object detection and beyond. In *2011 International conference on computer vision*, pages 89–96. IEEE, 2011.

[Mellado *et al.*, 2014] Nicolas Mellado, Dror Aiger, and Niloy J Mitra. Super 4pcs fast global pointcloud registration via smart indexing. In *Computer Graphics Forum*, volume 33, pages 205–215. Wiley Online Library, 2014.

[Miller and Allen, 2004] A. T. Miller and P. K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics Automation Magazine*, 11(4):110–122, 2004.

[Miller *et al.*, 2003] Andrew T. Miller, Steffen Knoop, Henrik I. Christensen, and Peter K. Allen. Automatic grasp planning using shape primitives. In *ICRA*, volume 2, pages 1824–1829, Sep 2003.

[Mirtich and Canny, 1994] Brian Mirtich and John Canny. Easily computable optimum grasps in 2-d and 3-d. In *IEEE International Conference on Robotics and Automation*, pages 739–747. IEEE, 1994.

[Morrison *et al.*, 2018] Douglas Morrison, Peter Corke, and Jürgen Leitner. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. *arXiv preprint arXiv:1804.05172*, 2018.

[Morrison *et al.*, 2019] Douglas Morrison, Peter Corke, and Jürgen Leitner. Multi-view picking: Next-best-view reaching for improved grasping in clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8762–8768. IEEE, 2019.

[Mousavian *et al.*, 2019] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2901–2910, 2019.

[Mur-Artal *et al.*, 2015] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE transactions on robotics*, 31(5):1147–1163, 2015.

[Murali *et al.*, 2019] Adithyavairavan Murali, Arsalan Mousavian, Clemens Eppner, Chris Paxton, and Dieter Fox. 6-dof grasping for target-driven object manipulation in clutter. *arXiv preprint arXiv:1912.03628*, 2019.

[Najibi *et al.*, 2020] Mahyar Najibi, Guangda Lai, Abhijit Kundu, Zhichao Lu, Vivek Rathod, Thomas Funkhouser, Caroline Pantofaru, David Ross, Larry S Davis, and Alireza Fathi. Dops: Learning to detect 3d objects and predict their 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11913–11922, 2020.

[Nguyen, 1987] V-D Nguyen. Constructing stable grasps in 3d. In *IEEE International Conference on Robotics and Automation*, volume 4, pages 234–239. IEEE, 1987.

[Ni *et al.*, 2020] Peiyuan Ni, Wenguang Zhang, Xiaoxiao Zhu, and Qixin Cao. Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds. *arXiv preprint arXiv:2003.09644*, 2020.

[Nikandrova and Kyrki, 2015] Ekaterina Nikandrova and Ville Kyrki. Category-based task specific grasping. *Robotics and Autonomous Systems*, 70:25–35, 2015.

[Oberweger *et al.*, 2018] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018.

[Pang *et al.*, 2020] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for rgb-d salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[Park and Chun, 2018] Dongwon Park and Se Young Chun. Classification based grasp detection using spatial transformer network. *arXiv preprint arXiv:1803.01356*, 2018.

[Park *et al.*, 2018] Dongwon Park, Yonghyeok Seo, and Se Young Chun. Real-time, highly accurate robotic grasp detection using fully convolutional neural network with rotation ensemble module. *arXiv preprint arXiv:1812.07762*, 2018.

[Park *et al.*, 2019a] Dongwon Park, Yonghyeok Seo, Dongju Shin, Jaesik Choi, and Se Young Chun. A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection. *arXiv preprint arXiv:1909.07050*, 2019.

[Park *et al.*, 2019b] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7668–7677, 2019.

[Park *et al.*, 2020] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. 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*, pages 10710–10719, 2020.

[Pas and Platt, 2015] Andreas ten Pas and Robert Platt. Using geometry to detect grasps in 3d point clouds. *arXiv preprint arXiv:1501.03100*, 2015.

[Patil and Rabha, 2018] Aniruddha V Patil and Pankaj Rabha. A survey on joint object detection and pose estimation using monocular vision. *arXiv preprint arXiv:1811.10216*, 2018.

[Patten *et al.*, 2020] Timothy Patten, Kiru Park, and Markus Vincze. Dgcm-net: Dense geometrical correspondence matching network for incremental experience-based robotic grasping. *arXiv preprint arXiv:2001.05279*, 2020.

[Peng *et al.*, 2014] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. Rgbd salient object detection: a benchmark and algorithms. In *European conference on computer vision*, pages 92–109. Springer, 2014.

[Peng *et al.*, 2016] Houwen Peng, Bing Li, Haibin Ling, Weiming Hu, Weihua Xiong, and Stephen J Maybank.

Salient object detection via structured matrix decomposition. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):818–832, 2016.

[Peng *et al.*, 2019] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[Pereira and Alexandre, 2019] Nuno Pereira and Luís A Alexandre. Maskedfusion: Mask-based 6d object pose estimation. *arXiv preprint arXiv:1911.07771*, 2019.

[Pham *et al.*, 2019] Quang-Hieu Pham, Thanh Nguyen, Binh-Son Hua, Gemma Roig, and Sai-Kit Yeung. 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*, pages 8827–8836, 2019.

[Pham *et al.*, 2020] Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung. Lcd: Learned cross-domain descriptors for 2d-3d matching. In *AAAI*, pages 11856–11864, 2020.

[Piao *et al.*, 2019] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7254–7263, 2019.

[Pinheiro *et al.*, 2015] Pedro OO Pinheiro, Ronan Collobert, and Piotr Dollár. Learning to segment object candidates. In *Advances in Neural Information Processing Systems*, pages 1990–1998, 2015.

[Pinheiro *et al.*, 2016] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision*, pages 75–91. Springer, 2016.

[Pinto and Gupta, 2016] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3406–3413. IEEE, 2016.

[Ponce *et al.*, 1993] Jean Ponce, Steve Sullivan, J-D Boissonnat, and J-P Merlet. On characterizing and computing three-and four-finger force-closure grasps of polyhedral objects. In *IEEE International Conference on Robotics and Automation*, pages 821–827. IEEE, 1993.

[Qi *et al.*, 2017a] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017.

[Qi *et al.*, 2017b] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[Qi *et al.*, 2018] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 918–927, 2018.

[Qi *et al.*, 2019a] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9277–9286, 2019.

[Qi *et al.*, 2019b] Qi Qi, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Multi-scale capsule attention-based salient object detection with multi-crossed layer connections. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1762–1767. IEEE, 2019.

[Qi *et al.*, 2020] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4404–4413, 2020.

[Qin *et al.*, 2020] Yuzhe Qin, Rui Chen, Hao Zhu, Meng Song, Jing Xu, and Hao Su. S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes. In *Conference on Robot Learning*, pages 53–65, 2020.

[Qu *et al.*, 2017] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017.

[Rabbani and Van Den Heuvel, 2005] Tahir Rabbani and Frank Van Den Heuvel. Efficient hough transform for automatic detection of cylinders in point clouds. *Isprs Wg Iii/3, Iii/4*, 3:60–65, 2005.

[Rad and Lepetit, 2017] Mahdi Rad and Vincent Lepetit. 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*, pages 3828–3836, 2017.

[Redmon and Angelova, 2015] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1316–1322. IEEE, 2015.

[Redmon and Farhadi, 2017] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[Redmon and Farhadi, 2018] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[Redmon *et al.*, 2016] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[Ren *et al.*, 2015a] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for rgb-d saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 25–32, 2015.

[Ren *et al.*, 2015b] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[Rennie *et al.*, 2016] Colin Rennie, Rahul Shome, Kostas E Bekris, and Alberto F De Souza. A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters*, 1(2):1179–1185, 2016.

[Rosten and Drummond, 2005] Edward Rosten and Tom Drummond. Fusing points and lines for high performance tracking. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, pages 1508–1515. Ieee, 2005.

[Rublee *et al.*, 2011] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[Rusu *et al.*, 2009a] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *IEEE International Conference on Robotics and Automation*, pages 3212–3217, May 2009.

[Rusu *et al.*, 2009b] Radu Bogdan Rusu, Nico Blodow, Zoltan Csaba Marton, and Michael Beetz. 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*, pages 1–6. IEEE, 2009.

[Sabour *et al.*, 2017] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in neural information processing systems*, pages 3856–3866, 2017.

[Sabour *et al.*, 2018] Sara Sabour, Nicholas Frosst, and Geoffrey Hinton. Matrix capsules with em routing. In *6th international conference on learning representations, ICLR*, pages 1–15, 2018.

[Sahbani *et al.*, 2012] A. Sahbani, S. El-Khoury, and P. Bidaud. An overview of 3d object grasp synthesis algorithms. *Robotics and Autonomous Systems*, 60(3):326 – 336, 2012. Autonomous Grasping.

[Sajjan *et al.*, 2019] Shreeyak S Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Cleargrasp: 3d shape estimation of transparent objects for manipulation. *arXiv preprint arXiv:1910.02550*, 2019.

[Salti *et al.*, 2014] Samuele Salti, Federico Tombari, and Luigi Di Stefano. Shot: Unique signatures of histograms for surface and texture description. *Computer Vision and Image Understanding*, 125:251 – 264, 2014.

[Sanchez *et al.*, 2018] Jose Sanchez, Juan-Antonio Corrales, Belhassen-Chedli Bouzgarrou, and Youcef Mezouar. Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey. *The International Journal of Robotics Research*, 37(7):688–716, 2018.

[Sarode *et al.*, 2019a] Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Animesh Dhagat, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. One framework to register them all: Pointnet encoding for point cloud alignment. *arXiv preprint arXiv:1912.05766*, 2019.

[Sarode *et al.*, 2019b] Vinit Sarode, Xueqian Li, Hunter Goforth, Yasuhiro Aoki, Rangaprasad Arun Srivatsan, Simon Lucey, and Howie Choset. Pcrnet: Point cloud registration network using pointnet encoding. *arXiv preprint arXiv:1908.07906*, 2019.

[Saxena *et al.*, 2008a] Ashutosh Saxena, Justin Driemeyer, Justin Kearns, Chioma Osondu, and Andrew Y Ng. Learning to grasp novel objects using vision. In *Experimental Robotics*, pages 33–42. Springer, 2008.

[Saxena *et al.*, 2008b] Ashutosh Saxena, Justin Driemeyer, and Andrew Y Ng. Robotic grasping of novel objects using vision. *The International Journal of Robotics Research*, 27(2):157–173, 2008.

[Sermanet *et al.*, 2013] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[Shi and Rajkumar, 2020] Weijing Shi and Raj Rajkumar. 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*, pages 1711–1719, 2020.

[Shi *et al.*, 2015] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *IEEE transactions on pattern analysis and machine intelligence*, 38(4):717–729, 2015.

[Shi *et al.*, 2019] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–779, 2019.

[Shi *et al.*, 2020a] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10529–10538, 2020.

[Shi *et al.*, 2020b] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[Simon *et al.*, 2020] Martin Simon, Kai Fischer, Stefan Milz, Christian Tobias Witt, and Horst-Michael Gross. Stickypillars: Robust feature matching on point clouds using graph neural networks. *arXiv preprint arXiv:2002.03983*, 2020.

[Song and Xiao, 2014] Shuran Song and Jianxiong Xiao. Sliding shapes for 3d object detection in depth images. In *European conference on computer vision*, pages 634–651. Springer, 2014.

[Song and Xiao, 2016] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgb-d images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 808–816, 2016.

[Song *et al.*, 2020] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 431–440, 2020.

[Sultana *et al.*, 2020a] Farhana Sultana, Abu Sufian, and Paramartha Dutta. Evolution of image segmentation using deep convolutional neural network: A survey. *arXiv preprint arXiv:2001.04074*, 2020.

[Sultana *et al.*, 2020b] Farhana Sultana, Abu Sufian, and Paramartha Dutta. A review of object detection models based on convolutional neural network. In *Intelligent Computing: Image Processing Based Applications*, pages 1–16. Springer, 2020.

[Sundermeyer *et al.*, 2018] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *European Conference on Computer Vision.*, pages 712–729. Springer International Publishing, 2018.

[Suzuki *et al.*, 2020] Kanata Suzuki, Yasuto Yokota, Yuzi Kanazawa, and Tomoyoshi Takebayashi. 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)*, pages 205–212. IEEE, 2020.

[Szegedy *et al.*, 2014] Christian Szegedy, Scott Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.

[Tam *et al.*, 2013] Gary KL Tam, Zhi-Quan Cheng, Yu-Kun Lai, Frank C Langbein, Yonghuai Liu, David Marshall, Ralph R Martin, Xian-Fang Sun, and Paul L Rosin. Registration of 3d point clouds and meshes: A survey from rigid to nonrigid. *IEEE Transactions on Visualization and Computer Graphics*, 19(7):1199–1217, 2013.

[Tejani *et al.*, 2014] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. In *European Conference on Computer Vision*, pages 462–477. Springer, 2014.

[Tekin *et al.*, 2018] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 292–301, 2018.

[ten Pas *et al.*, 2017] Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. *Int. J. Rob. Res.*, 36(13-14):1455–1473, December 2017.

[Tian *et al.*, 2019a] Hao Tian, Changbo Wang, Dinesh Manocha, and Xinyu Zhang. Transferring grasp configurations using active learning and local replanning. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1622–1628. IEEE, 2019.

[Tian *et al.*, 2019b] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9627–9636, 2019.

[Tian *et al.*, 2020] Meng Tian, Liang Pan, Marcelo H Ang Jr, and Gim Hee Lee. Robust 6d object pose estimation by learning rgb-d features. *arXiv preprint arXiv:2003.00188*, 2020.

[Tosun *et al.*, 2020] Tarik Tosun, Daniel Yang, Ben Eisner, Volkan Isler, and Daniel Lee. Robotic grasping through combined image-based grasp proposal and 3d reconstruction. *arXiv preprint arXiv:2003.01649*, 2020.

[Tremblay *et al.*, 2018] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *arXiv preprint arXiv:1809.10790*, 2018.

[Truong *et al.*, 2019] Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, and Sandro De Zanet. Glampoints: Greedily learned accurate match points. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10732–10741, 2019.

[Uijlings *et al.*, 2013] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[Vacchetti *et al.*, 2004] Luca Vacchetti, Vincent Lepetit, and Pascal Fua. Stable real-time 3d tracking using online and offline information. *IEEE transactions on pattern analysis and machine intelligence*, 26(10):1385–1391, 2004.

[Vahrenkamp *et al.*, 2016] Nikolaus Vahrenkamp, Leonard Westkamp, Natsuki Yamanobe, Eren E Aksoy, and Tamim Asfour. Part-based grasp planning for familiar objects. In *IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 919–925. IEEE, 2016.

[Van der Merwe *et al.*, 2019] Mark Van der Merwe, Qingkai Lu, Balakumar Sundaralingam, Martin Matak, and Tucker Hermans. Learning continuous 3d reconstructions for geometrically aware grasping. *arXiv preprint arXiv:1910.00983*, 2019.

[Varley *et al.*, 2017] Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2442–2447. IEEE, 2017.

[Vidal *et al.*, 2018] J. Vidal, C. Lin, and R. Martí. 6d pose estimation using an improved method based on point pair features. In *4th International Conference on Control, Automation and Robotics (ICCAR)*, pages 405–409, April 2018.

[Villena-Martinez *et al.*, 2020] Victor Villena-Martinez, Sergiu Oprea, Marcelo Saval-Calvo, Jorge Azorin-Lopez, Andres Fuster-Guillo, and Robert B Fisher. When deep learning meets data alignment: A review on deep registration networks (drns). *arXiv preprint arXiv:2003.03167*, 2020.

[Vohra *et al.*, 2019] Mohit Vohra, Ravi Prakash, and Laxmidhar Behera. 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)*, pages 1–6. IEEE, 2019.

[Wada *et al.*, 2020] Kentaro Wada, Edgar Sucar, Stephen James, Daniel Lenton, and Andrew J Davison. Morefusion: Multi-object reasoning for 6d pose estimation from volumetric fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14540–14549, 2020.

[Wang and Jia, 2019] Zhixin Wang and Kui Jia. 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)*, pages 1742–1749. IEEE, 2019.

[Wang and Solomon, 2019a] Yue Wang and Justin M Solomon. Deep closest point: Learning representations for point cloud registration. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3523–3532, 2019.

[Wang and Solomon, 2019b] Yue Wang and Justin M Solomon. Prnet: Self-supervised learning for partial-to-partial registration. In *Advances in Neural Information Processing Systems*, pages 8812–8824, 2019.

[Wang *et al.*, 2016] Wenguan Wang, Jianbing Shen, Ling Shao, and Fatih Porikli. Correspondence driven saliency transfer. *IEEE Transactions on Image Processing*, 25(11):5025–5034, 2016.

[Wang *et al.*, 2018a] Shaoxiong Wang, Jiajun Wu, Xingyuan Sun, Wenzhen Yuan, William T Freeman, Joshua B Tenenbaum, and Edward H Adelson. 3d shape perception from monocular vision, touch, and shape priors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1606–1613. IEEE, 2018.

[Wang *et al.*, 2018b] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2569–2578, 2018.

[Wang *et al.*, 2019a] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. *arXiv preprint arXiv:1910.10750*, 2019.

[Wang *et al.*, 2019b] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3343–3352, 2019.

[Wang *et al.*, 2019c] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. 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*, pages 2642–2651, 2019.

[Wang *et al.*, 2019d] Shengfan Wang, Xin Jiang, Jie Zhao, Xiaoman Wang, Weiguo Zhou, and Yunhui Liu. 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)*, pages 474–480. IEEE, 2019.

[Wang *et al.*, 2019e] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, and Haibin Ling. Salient object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019.

[Wang *et al.*, 2019f] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. *arXiv preprint arXiv:1912.04488*, 2019.

[Wang *et al.*, 2019g] Xinlong Wang, Shu Liu, Xiaoyong Shen, Chunhua Shen, and Jiaya Jia. Associatively segmenting instances and semantics in point clouds. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4096–4105, 2019.

[Watkins-Valls *et al.*, 2019] David Watkins-Valls, Jacob Varley, and Peter Allen. Multi-modal geometric learning for grasping and manipulation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7339–7345. IEEE, 2019.

[Wei *et al.*, 2012] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *European conference on computer vision*, pages 29–42. Springer, 2012.

[Wong *et al.*, 2017] Jay M Wong, Vincent Kee, Tiffany Le, Syler Wagner, Gian-Luca Mariottini, Abraham Schneider, Lei Hamilton, Rahul Chipalkatty, Mitchell Hebert, David MS Johnson, et al. Segicp: Integrated deep semantic segmentation and pose estimation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5784–5789. IEEE, 2017.

[Xiang *et al.*, 2018] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolu-

tional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems (RSS)*, 2018.

[Xie *et al.*, 2020a] Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox. The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation. In *Conference on Robot Learning*, pages 1369–1378, 2020.

[Xie *et al.*, 2020b] Enze Xie, Peize Sun, Xiaoge Song, Wenhai Wang, Xuebo Liu, Ding Liang, Chunhua Shen, and Ping Luo. Polarmask: Single shot instance segmentation with polar representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12193–12202, 2020.

[Xie *et al.*, 2020c] Qian Xie, Yu-Kun Lai, Jing Wu, Zhoutao Wang, Yiming Zhang, Kai Xu, and Jun Wang. Mlcvnet: Multi-level context votenet for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10447–10456, 2020.

[Xu *et al.*, 2018] Danfei Xu, Dragomir Anguelov, and Ashesh Jain. Pointfusion: Deep sensor fusion for 3d bounding box estimation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Jun 2018.

[Xue *et al.*, 2009] Zhixing Xue, Alexander Kasper, J Marius Zoellner, and Ruediger Dillmann. An automatic grasp planning system for service robots. In *2009 International Conference on Advanced Robotics*, pages 1–6. IEEE, 2009.

[Yan *et al.*, 2018a] Xinchen Yan, Jasmined Hsu, Mohammad Khansari, Yunfei Bai, Arkanath Pathak, Abhinav Gupta, James Davidson, and Honglak Lee. Learning 6-dof grasping interaction via deep geometry-aware 3d representations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–9. IEEE, 2018.

[Yan *et al.*, 2018b] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

[Yan *et al.*, 2019] Xinchen Yan, Mohi Khansari, Jasmine Hsu, Yuanzheng Gong, Yunfei Bai, Sören Pirk, and Honglak Lee. Data-efficient learning for sim-to-real robotic grasping using deep point cloud prediction networks. *arXiv preprint arXiv:1906.08989*, 2019.

[Yang *et al.*, 2013] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[Yang *et al.*, 2015] Jiaolong Yang, Hongdong Li, Dylan Campbell, and Yunde Jia. Go-icp: A globally optimal solution to 3d icp point-set registration. *IEEE transactions on pattern analysis and machine intelligence*, 38(11):2241–2254, 2015.

[Yang *et al.*, 2019a] Bo Yang, Jianan Wang, Ronald Clark, Qingyong Hu, Sen Wang, Andrew Markham, and Niki Trigoni. Learning object bounding boxes for 3d instance segmentation on point clouds. In *Advances in Neural Information Processing Systems*, pages 6737–6746, 2019.

[Yang *et al.*, 2019b] Shuo Yang, Wei Zhang, Weizhi Lu, Hesheng Wang, and Yibin Li. Learning actions from human demonstration video for robotic manipulation. *arXiv preprint arXiv:1909.04312*, 2019.

[Yang *et al.*, 2019c] Zetong Yang, Yanan Sun, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Std: Sparse-to-dense 3d object detector for point cloud. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1951–1960, 2019.

[Yang *et al.*, 2020a] Heng Yang, Jingnan Shi, and Luca Carlone. Teaser: Fast and certifiable point cloud registration. *arXiv preprint arXiv:2001.07715*, 2020.

[Yang *et al.*, 2020b] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11040–11048, 2020.

[Ye *et al.*, 2020] Maosheng Ye, Shuangjie Xu, and Tongyi Cao. Hvnet: Hybrid voxel network for lidar based 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1631–1640, 2020.

[Yew and Lee, 2018] Zi Jian Yew and Gim Hee Lee. 3dfeat-net: Weakly supervised local 3d features for point cloud registration. In *European Conference on Computer Vision*, pages 630–646. Springer, 2018.

[Yi *et al.*, 2016] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. In *European Conference on Computer Vision*, pages 467–483. Springer, 2016.

[Yi *et al.*, 2019] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019.

[Yokota *et al.*, 2020] Yasuto Yokota, Kanata Suzuki, Yuzi Kanazawa, and Tomoyoshi Takebayashi. A multi-task learning framework for grasping-position detection and few-shot classification. In *2020 IEEE/SICE International Symposium on System Integration (SII)*, pages 1033–1039. IEEE, 2020.

[Yu *et al.*, 2019a] Fenggen Yu, Kun Liu, Yan Zhang, Chenyang Zhu, and Kai Xu. 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*, pages 9491–9500, 2019.

[Yu *et al.*, 2019b] Peiyu Yu, Yongming Rao, Jiwen Lu, and Jie Zhou. P^2gnet: Pose-guided point cloud generating networks for 6-dof object pose estimation. *arXiv preprint arXiv:1912.09316*, 2019.

[Yu *et al.*, 2020] Xin Yu, Zheyu Zhuang, Piotr Koniusz, and Hongdong Li. 6dof object pose estimation via differentiable proxy voting loss. *arXiv preprint arXiv:2002.03923*, 2020.

[Yuan *et al.*, 2020] Yijun Yuan, Jiawei Hou, Andreas Nüchter, and Sören Schwertfeger. Self-supervised point set local descriptors for point cloud registration. *arXiv preprint arXiv:2003.05199*, 2020.

[Zakharov *et al.*, 2019] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1941–1950, 2019.

[Zapata-Impata *et al.*, 2017] Brayan S Zapata-Impata, Carlos Mateo Agulló, Pablo Gil, and Jorge Pomares. Using geometry to detect grasping points on 3d unknown point cloud. 2017.

[Zapata-Impata *et al.*, 2019] Brayan S Zapata-Impata, Pablo Gil, Jorge Pomares, and Fernando Torres. Fast geometry-based computation of grasping points on three-dimensional point clouds. *International Journal of Advanced Robotic Systems*, 16(1):1729881419831846, 2019.

[Zeng *et al.*, 2017a] Andy Zeng, Shuran Song, Matthias Nießner, Matthew Fisher, Jianxiong Xiao, and Thomas Funkhouser. 3dmatch: Learning local geometric descriptors from rgb-d reconstructions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1802–1811, 2017.

[Zeng *et al.*, 2017b] Andy Zeng, Kuan-Ting Yu, Shuran Song, Daniel Suo, Ed Walker, Alberto Rodriguez, and Jianxiong Xiao. Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 1386–1383. IEEE, 2017.

[Zeng *et al.*, 2018] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. 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)*, pages 1–8. IEEE, 2018.

[Zhang *et al.*, 2016] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5733–5742, 2016.

[Zhang *et al.*, 2017] Qiang Zhang, Daokui Qu, Fang Xu, and Fengshan Zou. Robust robot grasp detection in multimodal fusion. In *MATEC Web of Conferences*, volume 139, page 00060. EDP Sciences, 2017.

[Zhang *et al.*, 2018a] Hanbo Zhang, Xuguang Lan, Site Bai, Lipeng Wan, Chenjie Yang, and Nanning Zheng. A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes. *arXiv preprint arXiv:1809.07081*, 2018.

[Zhang *et al.*, 2018b] Hanbo Zhang, Xuguang Lan, Site Bai, Xinwen Zhou, Zhiqiang Tian, and Nanning Zheng. Roi-based robotic grasp detection for object overlapping scenes. *arXiv preprint arXiv:1808.10313*, 2018.

[Zhang *et al.*, 2020a] Feihu Zhang, Chenye Guan, Jin Fang, Song Bai, Ruigang Yang, Philip Torr, and Victor Prisacariu. Instance segmentation of lidar point clouds. *ICRA, Cited by*, 4(1), 2020.

[Zhang *et al.*, 2020b] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3dnet: 3d object detection using hybrid geometric primitives. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[Zhao and Tao, 2020] Lin Zhao and Wenbing Tao. Jsnet: Joint instance and semantic segmentation of 3d point clouds. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2020.

[Zhao *et al.*, 2015] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1265–1274, 2015.

[Zhao *et al.*, 2019] Zhong-Qiu Zhao, Peng Zheng, Shou-tao Xu, and Xindong Wu. Object detection with deep learning: A review. *IEEE transactions on neural networks and learning systems*, 30(11):3212–3232, 2019.

[Zhao *et al.*, 2020a] Binglei Zhao, Hanbo Zhang, Xuguang Lan, Haoyu Wang, Zhiqiang Tian, and Nanning Zheng. Regnet: Region-based grasp network for single-shot grasp detection in point clouds. *arXiv preprint arXiv:2002.12647*, 2020.

[Zhao *et al.*, 2020b] Sicheng Zhao, Bo Li, Pengfei Xu, and Kurt Keutzer. Multi-source domain adaptation in the deep learning era: A systematic survey. *arXiv preprint arXiv:2002.12169*, 2020.

[Zheng *et al.*, 2019] Tianhang Zheng, Changyou Chen, Junsong Yuan, Bo Li, and Kui Ren. Pointcloud saliency maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1598–1606, 2019.

[Zhou and Tuzel, 2018] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4490–4499, 2018.

[Zhou *et al.*, 2016] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Fast global registration. In *European Conference on Computer Vision*, pages 766–782. Springer, 2016.

[Zhou *et al.*, 2018] Xinwen Zhou, Xuguang Lan, Hanbo Zhang, Zhiqiang Tian, Yang Zhang, and Narming Zheng. Fully convolutional grasp detection network with oriented anchor box. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7223–7230. IEEE, 2018.

[Zhou *et al.*, 2019a] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.

[Zhou *et al.*, 2019b] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krahenbuhl. Bottom-up object detection by grouping extreme and center points. In *Proceedings of the IEEE*

Conference on Computer Vision and Pattern Recognition, pages 850–859, 2019.

[Zhou *et al.*, 2019c] Zheming Zhou, Tianyang Pan, Shiyu Wu, Haonan Chang, and Odest Chadwicke Jenkins. Glassloc: Plenoptic grasp pose detection in transparent clutter. *arXiv preprint arXiv:1909.04269*, 2019.

[Zhu *et al.*, 2014] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821, 2014.

[Zhu *et al.*, 2020] Angfan Zhu, Jiaqi Yang, Chen Zhao, Ke Xian, Zhiguo Cao, and Xin Li. Lrf-net: Learning local reference frames for 3d local shape description and matching. *arXiv preprint arXiv:2001.07832*, 2020.

[Zou *et al.*, 2019] Zhengxia Zou, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *arXiv preprint arXiv:1905.05055*, 2019.