3D Grasp Pose Generation from 2D Anchors and Local Surface

Yangchang Sun*, Minghao Yang*, Jialing Li, Baohua Qiang, Jinlong Chen

***Abstract*—This work proposes a three dimensional (3D) robot grasp pose generation method for robot manipulator from the predicted two dimensional (2D) anchors and the depth information of local surface. Compared to the traditional image based grasp area detection methods in which the grasp pose are only presented by two contacts, the proposed method is able to generate more accurate 3D grasp pose. Furthermore, different from the 6-DoF object pose regression methods in which the point cloud of the whole objects is considered, the proposed method is very lightweight, since the 3D computation is only processed on the depth information of the local grasp surface. The method consists of three steps: (1) detecting the 2D grasp anchor and extracting the local grasp surface from image; (2) obtaining the normal vector of the objects' local grasp surface from the objects' local point cloud; (3) generating the 3D grasp pose from 2D grasp anchor based on the normal vector of local grasp surface. The experiments are carried on the Cornell and Jacquard grasp datasets. It is found that the proposed method yields improvement on the grasp accuracy compared to the state-of-art 2D anchor methods. And the proposed method is also validated on the practical grasp tasks deployed on a UR5 arm with Robotiq Grippers F85. It outperforms the state-of-art 2D anchor methods on the grasp success rate for dozens of piratical grasp tasks.**

I. Introduction

With the rapid development of robotic grasp detection techniques, the deep architecture based grasp detection methods are able to predict the objects' grasp areas accurately in real-time from images [1-4]. These traditional image based grasp area detection methods output the 2D grasp contact points for robot gripper. In this way, the gripper is able to grasp the objects along the directions orthogonal to the image planes. However, in practical grasp task, the appropriate grasp pose in 3D space is not always orthogonal to the image planes because of the various object pose in the scenes. Fig. 1 presents an example from Jacquard, an open grasp dataset [5]. The box in Fig. 1(a) presents the predicted 2D grasp anchor, where the line segments in red present the range of two contacts of the gripper. The red lines in Fig. 1(b) present the grasp pose of a simulated gripper whose grasp direction is orthogonal to the 2D image planes. In Fig. 1(c), the lines in blue present the generated 3D grasp pose of the simulated gripper, whose grasp direction is parallel to the normal vector of the local grasp surface around the anchor area. Quite a few researchers have discussed that the appropriate grasp pose near to the objects should be given by the normal distribution of the objects' surfaces [6-9]. In Fig.1(c), the big green arrow presents the normal vector of the local grasp surface. It is reasonable that the grasp pose presented in Fig.1(c) is better than the one given in Fig.1(b).

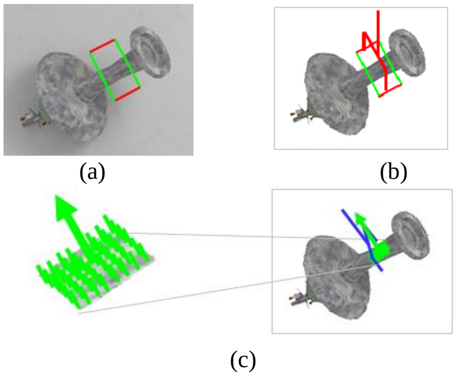

Figure. 1. An grasp example from Jacquard dataset, where the object inclines towards table.

The case presented in Fig.1 indicates it is necessary to generate more accurate 3D grasp pose from images. In term of this, various 6-DoF object pose estimation methods have been proposed [7, 9-13]. However, these methods focus on estimating the objects' pose without grasp details, at the same time, these 6-DoF methods are usually rather time cost since the whole object point cloud is considered in the 6-DoF parameters estimation.

We can see that despite the various 2D grasp anchor detection methods and 6-DoF objects pose regression methods, a faster and more efficient 3D grasp pose generation strategy from image is still needed. This work proposes a 3D grasp pose generation method from the 2D grasp area. The contributions of the proposed method include:

- Methodologically, it is a novel strategy to generate robust 3D grasp pose from the candidate grasp anchors in images and the depth information of local grasp surfaces.
- In the terms of effect, the proposed method yields accuracy improvement on grasp compared to the state-of-art 2D grasp area detection methods. It also outperforms the state-of-art methods on the grasp success rate on dozens of real objects grasp tasks in the practical grasp scenes.
- For experimental aspect, the discussions presented in this work indicate that the 3D grasp pose orthogonal to the object local grasp surface is better than the 2D grasp pose presented only by two contacts.

*These authors made equal contributions to this work.

Yangchang Sun is with the University of Chinese Academy of Sciences, and the Research Center for Brain-inspired Intelligence (BII), Institute of Automation, Chinese Academy of Sciences (CASIA), China (e-mail: yangchang.sun@ia.ac.cn);.

Minhao Yang is with the Research Center for Brain-inspired Intelligence (BII), Institute of Automation, Chinese Academy of Sciences (CASIA), (e-mail: mhyang@nlpr.ia.ac.cn);

Jialin Li, Baohua Qiang, Jinlong Chen are with the Guilin University of Electronic Science and technology, China (e-mail: Yuanhao.Qu@guet.edu.cn, jinlong.chen@guet.edu.cn).

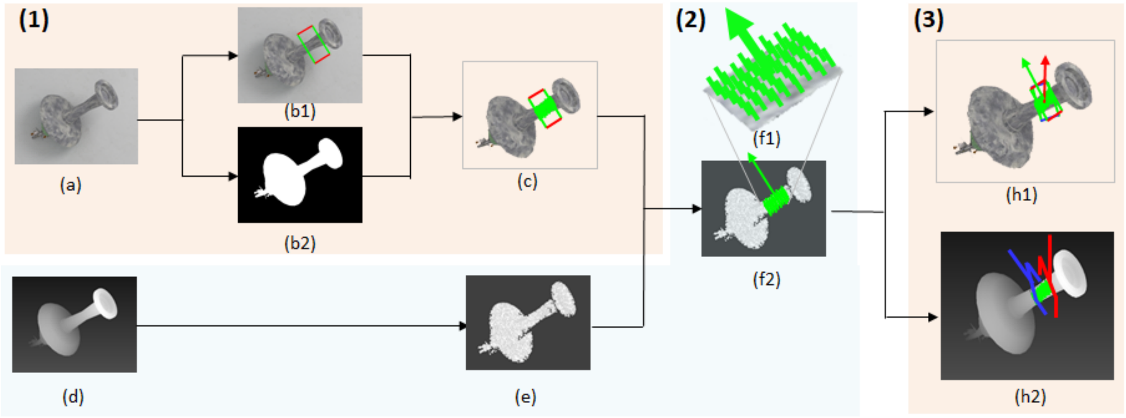

Figure. 2. The outline of the proposed method.

The reminders of this work are organized as following: related works are introduced in section II; the idea and the details of the proposed method are presented in section III; and the experiments, discussion and conclusions are given in section IV and V respectively.

II. Related Work

Accurate and robust robotic grasping is a challenging task in robotic manipulator operation. Before the robot grasps objects, it is necessary to know where the target object located and how to catch them. In early period, some researchers used shape matching methods to detect the objects in the scenes, such as Hu moments [14-16], Histogram of oriented gradient (HOG) [17, 18], Shape contex (SC) [19, 20], Iterative Closest Point (ICP) [21], Point Distribution Model (PMP) [22], etc. Once the objects are located, the gripper touch areas are also obtained from the object templates [23]. In this period, these methods focus more on extracting the shape well from the background and matching the candidate shapes accurately according to the shape templates. How to generate a more robust grasp pose at the grasp area was not well discussed.

Shape based grasp detection is challenged by the self-occlusion situations. Texture or feature points based object detection methods are less constrained to occlusions scenes, even occlusions happen among the different objects. The texture or feature points based object detection, such as Harr & Adaboosting [24], histograms of oriented gradients (HOG) [17], etc., are able to extract objects in complex scene. Recently, with the development of deep learning, deep neural network (DNN) based object detection methods, such as Yolo v1 - Yolo v4 [25-27], SSD [28] contribute to accurate object detections. Inspired by deep learning methods, deep architectures have been widely adopt to predict the grasp areas from image presentations [3, 29, 30] since 2013 when the deep neural network firstly used in grasping areas prediction [1]. In these methods, the grasp areas are presented by anchors, which are used to lable the ranges of object grasp areas such as the width, height and the center point. Recently, deep architecture has been expanded in grasp multi-task learning, in which not only the grasping anchors, but also the object categories, semantic segmentations and grasp styles (pick or suction) are obtained simultaneously [4, 31, 32]. These methods focus on improving the anchors' accuracy [30, 33], while how to generate detailed 3D pose from 2D grasp anchor was not discussed.

There are some researchers who aim to predict the 6-DoF object pose according to the objects' appearances and geometries [34]. This kind of methods provide a different approach to obtain the grasp pose through 6-DoF objects pose (the objects' 3D translation and 3D rotation to camera). With the pioneer work of SSD-6D [35] and PoseCNN [12], a series of deep learning based 6-DoF pose prediction methods, such as [13] DenseFusion [36], etc., were able to predict the objects' distances and rotations to the camera accurately as well as the objects' occupied areas in images. In spite of high performance of the whole objects' pose prediction, most of these methods did not provide detailed grasp in the images. In addition, since the objects appearances and the geometries (or point cloud) were needed to estimate the 6-DoF parameters simultaneously, the time cost of these methods is considerable, for example, the pose estimation and ICP for a single object in the scene cost about 0.1 and 0.3 seconds per frame on NVIDIA Titan X GPU [13].

Our idea is that it is no need to estimate all the 6 parameters from object to camera, since the image detection methods are able to obtain very accurate grasp areas on images. Based on the 2D grasp areas, we can obtain the detailed 3D grasp pose (the two grasp contacts and the grasp forward direction in manipulator's 3D operation space) by estimating the normal vector of the local grasp surface. The local grasp surface normal calculation is simple straightforward, which could be easily obtained from the depth information of the grasp areas using optimal neighborhood from local information [37], or the principal component analysis based (PCA) normal components computing [38, 39].

III. The Proposed Method

The total framework of the proposed method is presented in Fig. 2, which consists of three main steps:

- 2D grasp area detection. The first step aims to find the 2D grasp anchors on objects and extract the local grasp surface according to the object mask. The box labeled with the red-green line segments in Fig.2(b1) is a possible anchor for the object in the image of Fig.1(a). Fig. 2(b2) is the object mask and Fig. 2(c)

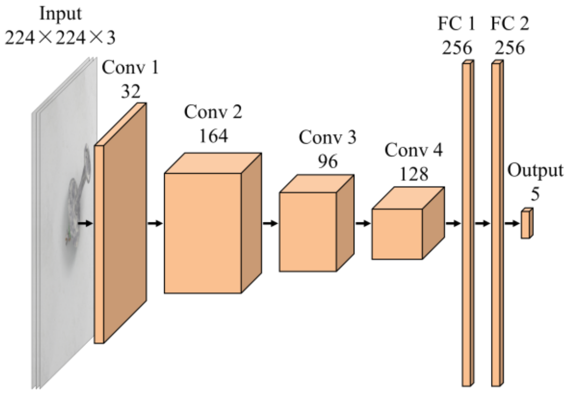

Figure. 3. The backbone strcuture of ENet used for 2D grasp anchor detection.

presents the grasp area on the extracted object from image background.

- Obtaining the normal vector of the local grasp surface. The second step aims to obtain the normal vector of the objects' local surfaces from the local point cloud around the grasp anchor. Fig.2(d) and Fig.2(e) present the depth information and the point cloud of the object respectively. The big green arrow in Fig. 2(f2) presents the normal vector of the points in the range of grasp areas obtained in the first step. And the Fig. 2(f1) presents the detailed information of the normal vectors around the grasp area.

- 3D grasp pose generation. The third step aims to generate the 3D grasp pose according to the normal vector of the local grasp surface. Fig. 2(h2) presents the final 3D grasp pose where the lines in blue and red are the 3D grasp pose of the final 3D simulated gripper and the initial one.

A. *2D Grasp Anchor Detection*

In this step, we adopt the convolutional neural networks (CNN) based architecture described in [40] to predict the 2D grasp anchors. We call the CNN architecture as ENet in this work. It receives RGB image in the size of 224 × 224 × 3 as input and generates encoded grasp {x, y, w, h, θ}. Fig. 3 presents an outline of ENet, which contains only 4 convolutional layers and 2 FC layers. With its simple structure, the ENet is a lightweight network for 2D grasp anchor prediction.

The predicted 2D grasp areas usually contain the image backgrounds which should be excluded. This could be processed by the intersection of the 2D grasp anchors and the object masks. The green area in Fig. 2(c) presents the local surface obtained by the intersection of grasp anchor (Fig. 2(b1)) and the object mask (Fig.2(b2)). Although most of the grasp datasets provide the object mask information, in practical grasp tasks, it is necessary to extract the object from the image background. In this work, we use the a real-time instance segmentation technique YOLACT [41] to extract the objects from image background. The segmentation process of YOLACT is broken into two parts: geanerating a set of prototype masks and predicting pre-instance mask coefficients.

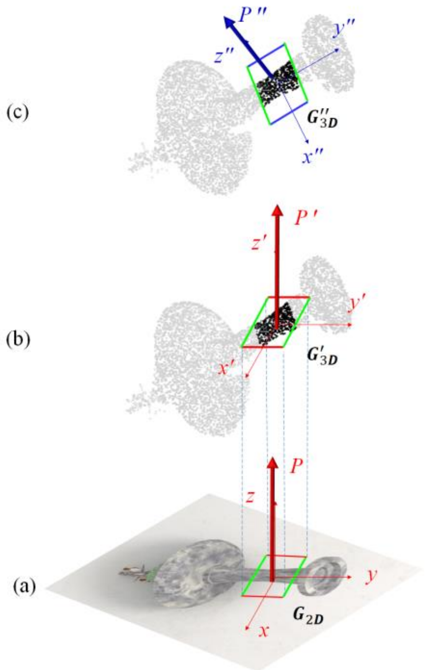

Figure. 4. The process of 3D grasp pose generation.

As the two parts can be computed independently and in parallel style, YOLACT runs faster than previous one-stage instance segmentation approaches [42].

B. *Obtaining the Grasp Local Surface Normal*

The 2D grasp anchors obtained from ENet are not good enough for 3D grasping. To generate the 3D grasp pose, this work calculates the normal vector from the point cloud in the local grasp area. As described in [38, 39], the typical data dimensionality reduction method PCA (Principal Component Analysis) could be applied to calculate the normal vector of the point cloud.

Based on the 2D grasp anchor and the local surface extracted by the intersection of grasp anchor and object mask, the corresponding point cloud can be straightforward obtained from the whole object point cloud. Fig. 2(d) and Fig. 2(e) present the depth information and point cloud for the object present in Fig. 2(a). The points in the green area in Fig. 2(f2) are the local grasp point cloud, which are used for local surface normal vector calculation. We use Principal Component Analysis (PCA) [38] method to calculate the normal vector of the local grasp point cloud. The local points can be stacked in a in a matrix $\mathrm{X} = [x_1, x_2, \dots, x_k] \in \mathbb{R}^{3\times K}$. P, P', P'' in Fig. 4(a), Fig. 4(b) and Fig. 4(c) are the normal vectors orthogonal to the image plane, the horizontal plane of manipulator work space and the local grasp surface respectively. Then the normal vector P'' of the local grasp surface can be represented with the eigenvector of covariance matrix $S = \sum_{i=1}^{K}(x_i - m)(x_i - m)^T = YY^T$ with the maximum eigenvalue, where $m = \frac{1}{K}\sum_{i=1}^{K} x_i$ is the centroid of local point cloud. S is a 3×3 matrix, its eigenvectors can be computed by SVD (singular value decomposition). The normal vector illustrates the direction orthogonal to the local point cloud surface. The normal vector illustrates the direction orthogonal to the local grasp surface.

C. 3D Grasp Pose Generation

After the normal vector P'' is obtained, the 3D grasp pose can be estimated from 2D grasp anchor. Supposing $G_{2D} = (x, y, z, \alpha, \beta, \gamma, w, h)$ presents the 2D grasp anchor predicted by ENet [40], which is expanded from the traditional 2D grasp anchor (x, y, α, w, h), where the x, y, z are the coordinates of anchor center, w, h are the width and height of grasp anchor, and α, β, γ are the yaw, pitch and roll angle of anchor rotation. Since the point (x, y, z) is located in image plane, the values of z, β and γ are zero. Let the anchors along the normal vector P' and P'' be presented by $G'_{3D} = (\mathrm{x}', \mathrm{y}', \mathrm{z}', \alpha', \beta', \gamma', \mathrm{w}', \mathrm{h}')$ and G''_{3D}=$(x'', y'', z'', \alpha'', \beta'', \gamma'', \mathrm{w}'', \mathrm{h}'')$ respectively, where the parameters in G'_{3D} is equal to those in G_{2D} since the image plane is the horizontal plane of manipulator work space. In term of the conversion from G'_{3D} to G''_{3D}, the values of $x'', y'', z'', \mathrm{w}'', \mathrm{h}''$ are equal to $\mathrm{x}', \mathrm{y}', \mathrm{z}', \mathrm{w}', \mathrm{h}'$. Then α'', β'' and γ'' could be obtained by (1)~(3) according to the conversion from axis-angle to Euler [43].

$$\alpha'' = \ arctan\frac{y_u \sin\omega - x_u z_u (1-\cos\omega)}{1-(y_u^2+z_u^2)(1-\cos\omega)} \tag{1}$$

$$\beta''= arcsin(x_u y_u(1-\cos\omega)) + z_u sin\omega \tag{2}$$

$$\gamma'' = arctan\frac{x_u sin\omega - y_u z_u (1-\cos\omega)}{1-(x_u^2+z_u^2)(1-\cos\omega)} \tag{3}$$

In (1)~(3), ω and U =$(x_u, y_u\ z_u)$ are the rotation angle and rotation axis from G'_{3D} to G''_{3D}, which can be obtained using (4) and (5) [44].

$$\omega = P' \cdot P'' \tag{4}$$

$$U = P' \times P'' \tag{5}$$

IV. Experiments

We first evaluate the proposed method on two widely-recognized open grasp datasets and then on dozens of practical grasp tasks. These two datasets are Cornell [45] and Jacquard [46] datasets. On these two datasets, we compare the proposed methods with some state-of-art methods both on the 3D grasp accuracy and time-cost. In practical grasp tasks, we mainly compare the grasp success rate between the proposed method and some state-of-art 2D grasp detection methods. The proposed method is deployed on a PC with CPU core at 3.60GHz, 32G RAM. The 2D anchor prediction network ENet runs on the MindSpore AI framework [47] deployed on Ascend 910.

A. Grasp Pose Prediction

Cornell grasp dataset contains 885 images of 240 grasp objects, and Jacquard contains 54k images of 11k objects. The labels are presented with 2D oriented rectangles with metric $G_{2D} = (x, y, 0, \alpha, 0, 0, w, h)$. In 3D space, the 2D anchor is lack of one dimension along the grasp forward direction to the touch areas of objects. Therefore, in this section, we adopt a 3D box based grasp accuracy estimation method (G-index for short) to estimate the performances of the proposed method on 3D grasp accuracy.

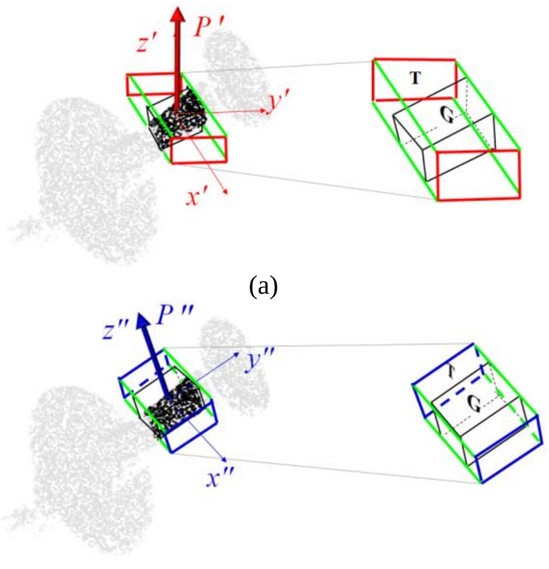

Figure. 5. Demonstrations of G-index measurement.

Let B_l, B_h, B_t be the length, width and thickness of the grasp box in 3D space, where B_l, B_h are equal to the values of $\mathrm{w}'', \mathrm{h}''$in G''_{3D}. The direction of box thickness is parallel to the gripper forward direction. B_t =$(\max(\forall_i d_i) - \min(\forall_i d_i))$ is the differences between the maximal and minimal depth values of all local surface points along $P'(G'_{3D})$, $P''(G''_{3D})$, where d_i is the depth value of point $p_i (1 \le i \le I)$. I is the number of the points in the local surface. We call the box with size of $\mathrm{B}_l \times \mathrm{B}_h \times \mathrm{B}_t$ as the target box (TBox). In addition, let x_{max}, x_{min}, y_{max}, y_{min}, z_{max}, z_{min} be the maximal and minimal values of the surface points along the generated normal vector P'' in TBox, and an cube under the volume restrict of these parameters is defined as ideal grasp box (GBox). Fig. 5(a) and Fig. 5(b) present the examples of the 3D configuration of the TBox and GBox around the objects' surface. In Fig. 5, the GBox is labeled with black lines, and the box labeled with blue and green lines in Fig. 5(b) are the TBoxes generated by the proposed method, while red and green one in Fig. 5(a) for that before generated.

G-index could be obtained by (6), where Ω_G is the volume of GBox and Ω_T is that of TBox. The value of G-index (δ) is in the range of (0, 1), larger value mean better grasping.

$$\delta = \Omega_G / \Omega_T \tag{6}$$

Table 1 lists the experiment results of grasp accuracy between the proposed method and state-of-art methods on Cornell and Jacquard datasets. The datasets were divided into train-set, validation-set and test-set in the ratio of 6:2:2. We can see from Table 1 that [2], [40], [48], [49] achieve image-wise accuracy of 42.5% to 51.6% on Cornell dataset and 51.7% to 58.9% on Jacquard dataset under the measurement of G-index. While the proposed method yields image-wise accuracy of 69.5% and 72.1% on these two datasets respectively. As for object-wise accuracy, the proposed method reaches 61.5% on Cornell dataset and 68.2% on Jacquard dataset, which are also obviously higher than other methods such as ENet [40], which gets 49.2 % and 53.6% on Cornell and Jacquard.

TABLE 1. THE RESULTS OF THE PROPOSED COMPARED WITH SOME STATE-OF-ART METHODS ON THE GRASP ACCURACY MEASURED BY G-INDEX.

Approach	**Dataset**	**Accuracy (G-index)**		**Speed (fps)**
		Image-wise (%)	**Object-wise (%)**	
Redmon (2015) [2]	Cornell	42.5	39.3	13.15
Zhou (2018) [48]		48.7	46.7	8.51
ENet (2021) [40]		51.6	49.2	74.07
Proposed		**69.5**	**61.5**	**88.42**
Redmon (2015) [2]	Jacquard	51.7	45.4	13.15
Zhang (2019) [49]		54.3	51.9	25.16
ENet (2021) [40]		58.9	53.6	74.07
Proposed		**72.1**	**68.2**	**85.79**

TABLE 2. THE GRASP SUCCESS RATE IN PRACTICAL GRASP TASKS.

Approach	Redmon (2015) [2]	ENet (2021) [40]	Mousavian (2019) [50]	Proposed
$\tilde{\boldsymbol{R}}$	61.5%	76.5%	81.3%	**87.2%**

B. Grasp Success Rate in Practical Tasks

We use a UR5 arm with Robotiq Grippers F85 in practical grasp tasks. The system performance is explored with 50 common household objects in different sizes and shapes. Some samples of these objects are listed in Fig. 6. The robot executed 80 grasp trails on each object for various placements. A Kinect 2 horizontally fixed on the top of work space is used to obtain RGB and depth images simultaneously.

The grasp success rate R_i for each object is calculated by (7). In (7), the G_i^s and $G_i^u$are the number of successful and unsuccessful grasps for the i^{th}(1≤ i ≤50) object, where (G_i^s + G_i^u)= 80. And $\tilde{R}$=($\sum_{i=1}^{80} R_i$)$/80$ is the average success rate for all grasps.

$$R_i= \; G_i^s/(G_i^s+G_i^u) \tag{7}$$

Table 2 lists the success rate of the proposed method and [2], [40], [50] on practical grasp tasks. All the methods were trained with the same dataset mixed with all Cornell data and Jacquard data. In the grasps of [2], [40], [50] methods, the gripper grasps the objects according to the detected 2D grasp anchors in image and along the direction orthogonal to the image planes. While the gripper driven by the proposed method grasp the object along the directions orthogonal to the 3D local grasp surface. We can see from Table 2 that [2], [40], [50] obtain 61.5%, 76.5% and 81.3% on practical grasp tasks. The proposed method achieves a success rate of 87.2%, which is obviously higher than those of previous methods.

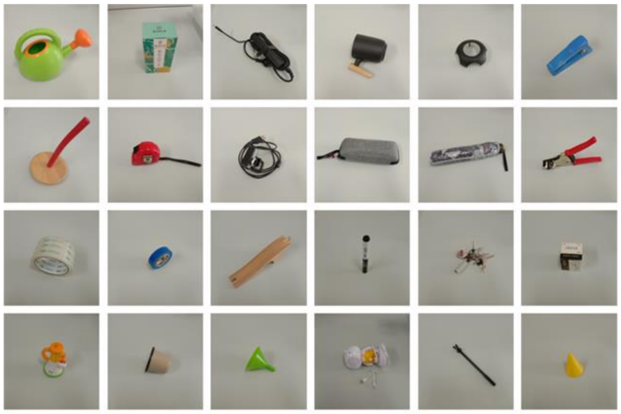

Figure. 6. Some objects in practical grasp tasks.

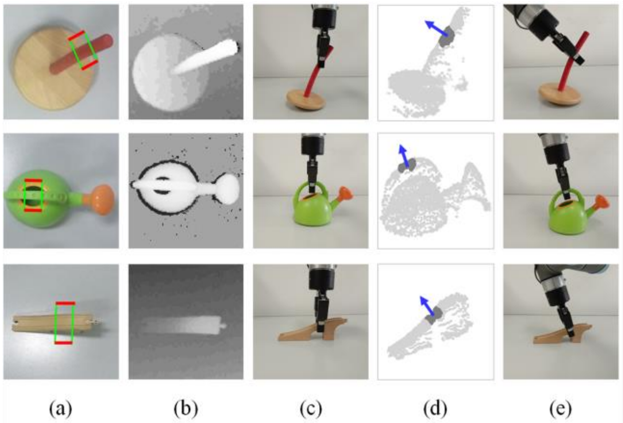

Figure. 7. Grasping examples in practical grasp tasks. (a) RGB images of the grasp objects and the 2D grasp anchors; (b) depth images of grasp objects; (c) grasp before 3D grasp pose generated; (d) local grasp point cloud (dark gray) with normal vector (blue); (e) grasp using the 3D grasp pose generated by the proposed method.

Fig. 7(a) lists some objects whose main axes at the grasp anchors are not parallel or vertical to the work plane. The images in Fig. 7(b) present their depth information. And Fig. 7(c) lists the successful grasp situations guided by the 2D grasp pose G_{2D}. Fig. 7(d) presents the point cloud generated from depth information, which is viewed along horizontal plane. The blue arrows in Fig. 7(d) indicate the normal vectors of the local grasp surfaces. Fig. 7(e) presents the successful grasp situations guided by the 3D grasp pose G''_{3D} generated from the proposed method. We can see from Fig. 7(e) that the Robotiq F85 gripper hold the objects more stablely than those in Fig. 7(c). It indicates that the proposed method is effective to generate 3D grasp pose orthogonal to local grasp surfaces.

Fig. 8 lists the average success rate of each object among 80 trails for each object. The success rates of all the objects are more than 65.0%, which illustrates the effectiveness of the proposed method. It is worth mentioning that the efficacy of the proposed method relies on the accuracy of the captured point cloud. In terms of the objects with more than 90% success rate such as pot, ink box and can, these objects have two characters: (1) obvious contours or shapes for accurate 2D

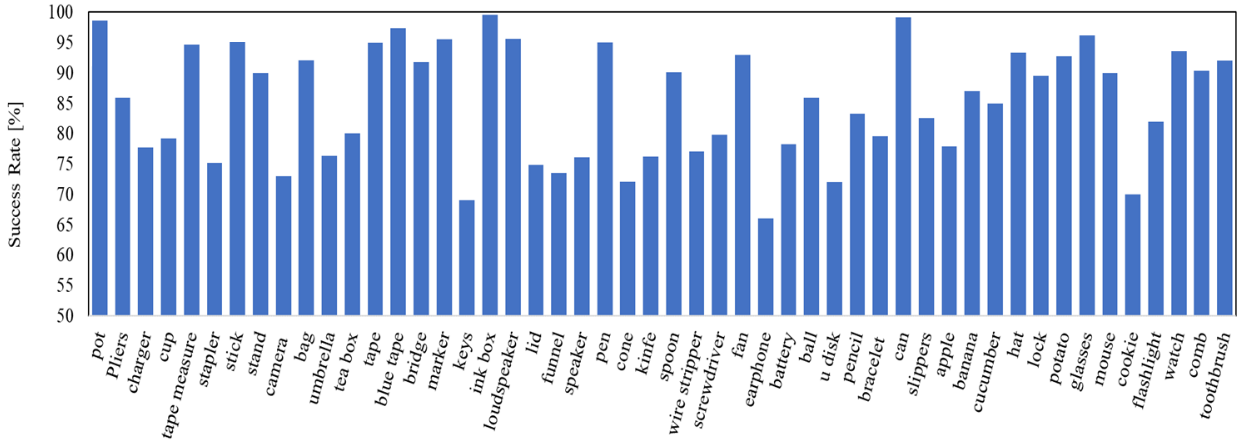

Figure. 8. Success rate of all 50 objects in practical grasp tasks from 80 trials for each object.

grasp anchors prediction; (2) enough depth along the camera sightline for good touch in local grasp surface. The objects with small size or thin thickness structure like earphone, u disk or cookie tend to produce low quality depth information, which make it difficult to construct enough depth information and lead to failure grasp.

C. Time Cost

Since the object's mask and depth information are provided in Cornell and Jacquard datasets, two steps remain in the proposed method: detecting the 2D grasp anchors and calculating the normal vector of local grasp surface. For the first step, we use the backbone similar to the structure of ENet [40]. ENet costs 13.6ms each frame on NVIDIA GeForce GTX 1050 Ti. The proposed method runs on Ascend 910 with MindSpore AI framework, and costs 7.3ms on 2D anchor prediction per frame. For the second step, 200-300 points sampled from the local grasp point cloud are used to calculate the surface normal using the PCA. This step costs about 4.0ms each frame. Then the time cost for the proposed method is about 11.3ms each frame or 88.42fps. The last column of Table 1 presents the time cost for the methods in detail. We can see that the proposed method is obviously faster.

D. Discussions

In experiments, we first compared the proposed method with some state-of-art methods on the grasp box accuracy. In the proposed method, the vertical direction of the grasp box, or the grasp forward direction, is parallel to the normal vector of the local grasp surface. While for the other methods, the grasp directions are orthogonal to the 2D image planes. With the depth information provided by Cornell and Jacquard datasets, the GBox volume given by the bounding cube of local grasp point cloud in the TBox could be obtained, which is used as 3D grasp accuracy in this work. The comparison results presented in Table 1 indicate that the proposed method obviously outperforms other methods.

We also compared those methods in practical grasp tasks. With dozens of grasps on each one in 50 common household objects, the success rate of the proposed method is higher than other state-of-art methods.

V. Conclusions

This work proposes a 3D grasp pose generation method from 2D grasp areas and local grasp surfaces. Compared to the traditional 2D image-based grasp detection methods, where the grasp direction is orthogonal to the image planes, the generated gasp pose is orthogonal to the local grasp surface and therefore more suitable to gripper grasping. In addition, compared to the traditional 6-DoF object pose estimation methods, the proposed method is a very lightweight method since its cost time is rather less than the related methods.

References

[1] H. L. Ian Lenz, Ashutosh Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research,* vol. 34, no. 4-5, 2013.

[2] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015: IEEE, pp. 1316-1322.

[3] F.-J. C. R. X. P. A. Vela, "Detecting Robotic Affordances on Novel Objects with Regional Attention and Attributes," *IEEE Robotics & Automation Letters*, 2019.

[4] T. O. Ryosuke Araki, Tsubasa Hirakawa, Takayoshi Yamashita, Hironobu Fujiyosh, "MT-DSSD Deconvolutional Single Shot Detector Using Multi Task Learning for Object Detection, Segmentation, and Grasping Detection," presented at the IEEE International Conference on Robotics and Automation (ICRA 2020), 2020.

[5] E. D. Amaury Depierre, Liming Chen, "Jacquard: A Large Scale Dataset for Robotic Grasp Detection," presented at the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, October, 1-5, 2018, 2018.

[6] Y. S. Zhe Cao, and Natasha Kholgade Banerjee, "Real-time scalable 6DOF pose estimation for textureless objects," presented at the In IEEE International Conference on Robotics and Automation (ICRA), 2016.

[7] X. Z. Georgios Pavlakos, Aaron Chan, KonstantinosG Derpanis, and Kostas Daniilidis, "6-DOF object pose from semantic keypoints " presented at the IEEE International Conference on Robotics and Automation (ICRA), 2017.

[8] Z. Li, G. Wang, and X. Ji, "CDPN Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation," presented at the International Conference on Computer Vision (ICCV 2019), 2019.

[9] A. M. Adithyavairavan Murali, Clemens Eppner, Chris Paxton, Dieter Fox, "6-DOF Grasping for Target-driven Object Manipulation in Clutter

" presented at the the International Conference on Robotics and Automation (ICRA 2020), 2020.

[10] F. M. Eric Brachmann, Alexander Krull, Michael Ying Yang, Stefan Gumhold, and Carsten Rother, "Uncertainty-driven 6D pose estimation of objects and scenes from a single RGB image " presented at the In IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), 2016.

[11] F. M. Wadim Kehl, Federico Tombari, Slobodan Ilic, and Nassir Navab, "Deep learning of local RGB-D patches for 3D object detection and 6D pose estimation," presented at the In European Conference on Computer Vision (ECCV), 2016.

[12] T. S. Yu Xiang, Venkatraman Narayanan and Dieter Fox, "PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes," presented at the Science and Systems Conference (RSS), 2018.

[13] M. L. Ge Gao, Yulong Wang, Xiaolin Hu, Jianwei Zhang, Simone Frintrop, "6D Object Pose Regression via Supervised Learning on Point Clouds," presented at the 2020 IEEE International Conference on Robotics and Automation (ICRA 2020), Paris (Virtual Conference), 2020.

[14] K. H. Joviša unic, Paul L. Rosin, "A Hu moment invariant as a shape circularity measure," *Pattern Recognition,* vol. 43, no. 1, pp. 47-57, 2010.

[15] K. H. Joviša Žunić, Dragan Dukić, Mehmet Ali Akta, "On a 3d analogue of the first hu moment invariant and a family of shape," *Mach Vision Appl,* vol. 27, pp. 129–144, 2016.

[16] O. J. CarlosMartinez, Žunić Errata "A 3D polar-radius-moment invariant as a shape circularity measure," *Neurocomputing* vol. 325, no. 24, pp. 303-304, 2019.

[17] B. T. Navneet Dalal, "Histograms of Oriented Gradients for Human Detection," presented at the IEEE Computer Society Conference on Computer Vision & Pattern Recognition (CVPR 2005), San Diego, CA, USA, 2005.

[18] B. F. Chunfang Liu, Fuchun Sun, Xiaoli Li, Wenbing Huang, "Learning to Grasp Familiar Objects Based on Experience and Objects' Shape Affordance," *IEEE Transactions on Systems, Man, and Cybernetics: Systems,* vol. 49, no. 12, pp. 2710 - 2723, 2019.

[19] D. K. Jeannette Bohg, "Learning grasping points with shape context," *Robotics and Autonomous Systems,* vol. 58, pp. 362-377, 2010.

[20] S. B. Xiang Bai, Zhuotun Zhu, Longin Jan Latecki, "3D Shape Matching via Two Layer Coding," *Ieee T Pattern Anal,* vol. 37, no. 12, pp. 2361 - 2373, 2015.

[21] T. C. David Liu, "Soft shape context for iterative closest point registration," presented at the 2004 International Conference on Image Processing (ICIP 2004), Singapore, 24-27 Oct. 2004, 2004.

[22] M. S. Alexandros Bouganis, "Flexible Object Recognition in Cluttered Scenes Using Relative Point Distribution Models," presented at the 2008 19th International Conference on Pattern Recognition (ICPR 2008), Tampa, FL, USA, 8-11 Dec. 2008, 2008.

[23] N. Guo, B. Zhang, J. Zhou, K. Zhan, S. J. C. Lai, and E. i. Agriculture, "Pose estimation and adaptable grasp configuration with point cloud registration and geometry understanding for fruit grasp planning," vol. 179, p. 105818, 2020.

[24] M. J. J. Paul Viola, "Rapid Object Detection using a Boosted Cascade of Simple Features," presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2001.

[25] R. Joseph, D. Santosh, G. Ross, and F. Ali, "You Only Look Once: Unified, Real-Time Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition, 2015.

[26] A. F. Joseph Redmon, "YOLOv3: An Incremental Improvement," presented at the IEEE on Computer Vision and Pattern Recognition 2018 (CVPR) Salt Lake City, UT, USA, June 18-22, 2018, 2018.

[27] A. Bochkovskiy, Wang, Chien-Yao , Liao, Hong-Yuan Mark, "YOLOv4: Optimal Speed and Accuracy of Object Detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2020), Seattle, WA, USA, June 13-19, 2020, 2020.

[28] W. Liu *et al.*, "SSD: Single Shot MultiBox Detector," presented at the ECCV 2016 - 14th European Conference on Computer Vision, Amsterdam, The Netherlands, October 11-14, 2016, 2016.

[29] Y. F. Jin Xie, Fan Zhu, Edward Wong, "Deepshape: Deep learned shape descriptor for 3D shape matching and retrieval," presented at the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7-12 June 2015, 2015.

[30] A. A. Joseph Redmon, "Real-Time Grasp Detection Using Convolutional Neural Networks," presented at the Proceedings of 2015 IEEE International Conference on Robotics and Automation (ICRA), Washington State Convention CenterSeattle, Washington May 26-30, 2015, 2015.

[31] S. S. Andy Zeng, Kuan-Ting Yu, Elliott Donlon, Francois R. Hogan, "Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching," presented at the 018 IEEE International Conference on Robotics and Automation (ICRA), Brisbane, QLD, Australia, 21-25 May 2018, 2018.

[32] Y. S. Dongwon Park, Dongju Shin, Jaesik Choi，Se Young Chun, "A Single Multi-Task Deep Neural Network with Post-Processing for Object Detection," presented at the 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May - 31 August, 2020, 2020.

[33] N. T. Hamid Rezatofighi, JunYoung Gwak, Amir Sadeghian, Ian Reid, Silvio Savarese, "Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression," presented at the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15-20 June 2019, 2019.

[34] S. H. Chavdar Papazov, Sven Parusel, "Rigid 3d geometry matching for grasping of known objects in cluttered scenes," *The International Journal of Robotics Research,* vol. 31, no. 4, pp. 538-553, 2012.

[35] F. M. Wadim Kehl, Federico Tombari, Slobodan Ilic, and Nassir Navab, "SSD-6D: Making RGB based 3D detection and 6D pose estimation great again," presented at the In IEEE International Conference on Computer Vision (ICCV), 2017.

[36] D. X. Chen Wang, Yuke Zhu, Roberto Mart, Cewu Lu, Li Fei-Fei, Silvio Savarese, "DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion," presented at the 2020 IEEE International Conference on Robotics and Automation (ICRA 2020), Paris (Virtual Conference), 2020.

[37] A. N. Niloy J. Mitra and N. n. Vol. 14, pp. 261-276 (2004), "Estimating surface normals in noisy point cloud data," *International Journal of Computational Geometry & Applications,* vol. 14, no. 5, pp. 261-276, 2004.

[38] K. Klasing, D. Althoff, D. Wollherr, and M. Buss, "Comparison of surface normal estimation methods for range sensing applications," in *2009 IEEE International Conference on Robotics and Automation*, 2009: IEEE, pp. 3206-3211.

[39] P. S. S. Zafeiriou, "Kernel-PCA Analysis of Surface Normals for Shape-from-Shading," presented at the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 23-28 June 2014, 2014.

[40] R. d. Q. M. Eduardo Godinho Ribeiro, Valdir Grassi Jr, "Real-time deep learning approach to visual servo control and grasp detection for autonomous robotic manipulation," *Robotics and Autonomous Systems,* vol. 139, p. 103757, 2021.

[41] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9157-9166.

[42] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961-2969.

[43] A.-A. t. Euler. (2021, 14. Sep). *http://www.euclideanspace.com/maths/geometry/rotations/conversions/angleToEuler/index.htm*.

[44] euler-angles-between-two-3d-vectors. (2021, 14 Sep). *https://stackoverflow.com/questions/15101103/euler-angles-between-two-3d-vectors*

[45] S. M. Yun Jiang, Ashutosh Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," presented at the IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China 9-13 May 2011, 2011.

[46] E. D. Amaury Depierre, Liming Chen, "Jacquard: A Large Scale Dataset for Robotic Grasp Detection " presented at the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1-5 Oct. 2018, 2018.

[47] MindSpore. (2021, 13 Sep). *https://www.mindspore.cn/doc/note/en/r1.1/network_list_ms.html*.

[48] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018: IEEE, pp. 7223-7230.

[49] L. X. Zhang Hanbo, Bai Site, Zhou Xinwen, Tian Zhiqiang, Zheng Nanning, "Roi-based robotic grasp detection for object overlapping scenes," presented at the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019.

[50] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901-2910.