Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Attention Grasping Network: A Real-time Approach to Generating Grasp Synthesis

Qipeng Gu[1,2], Jianhua Su[1*] and Xueting Bi[3]
[1]*the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*
[2]*the school of Artificial Intelligence, University of Chinese Academy of sciences*
[3]*Key Laboratory of Electromagnetic Radiation and Sensing Technology, Institute of Electronics, Chinese Academy of Sciences, Beijing 100190, China*
{guqipeng2018,jianhua.su}@ia.ac.cn,bixueting18@mails.ucas.edu.cn

***Abstract*—This paper presents a real-time, pixelwise method to generate grasp synthesis based on fully convolutional netural networks (FCN). Our proposed Attention Grasping Network (AGN) applies a novel attention mechanism to robotic grasp detection, which automatically learns to focus on salient features of the input image. The model with attention mechnisms can compensate for the loss of detail information in standard FCN, which increases the sensitivity of the model and accuracy of prediction. In addition,in order to ensure a real-time grasp and save computing resources, the light-weight AGN model predicts the position and angle of grasping point. Our method only takes 22ms to execute the grasp detection pipeline on a GPU-equipped computer, and achieves 97.8% accuracy on Cornell Grasping Dataset.**

I. Introduction

A grasp is an important skill for robots to interact with humans and environment. In order to complete the grasping task smoothly in complex environment, a robot must be able to calculate the positon of the object it. It is a challenging task for robots because it involves perception, planning and controls. A successful grasp is robust to noise and errors in perception, inaccuracies in control and perturbations in the robotic system.

In recent years, the development of deep netural network [22,23,24] has made a big progress on detection, classification and regression tasks. Deep netural network could establish maping from a image to grasp, and realize end-to-end learning. Current methods of deep learning for robotic grasp detection can be roughly divided into three categories.

- Based on sliding windows: The sliding window uses a classifier to slide on the whole image.The classifier needs to determine whether the small patches of an image constitute good grasps for an object in that image [6].
- Based on bounding boxes: The bounding boxes are proposed for object detection. Using bounding boxes for grasp detetion, the region proposal is extracted by the selective search or anchor box. Furthermore, The region proposal is bounded by the bounding boxes to get a grasping box close to the ground-truth grasp rectangle[1,11,21].

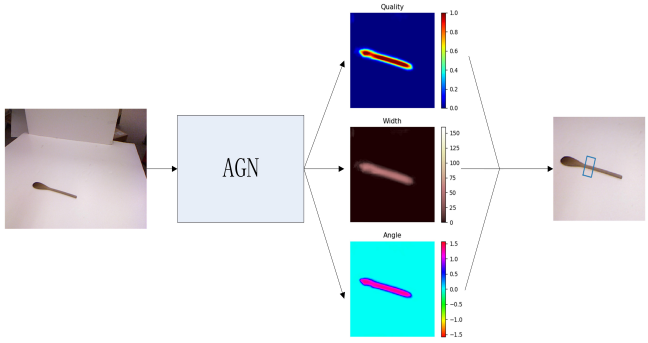

Fig. 1. Given a RGB image, our AGN model generates a best grasp in the grasp map. Our model learns a grasp for each pixel, and these grasps form a grasp map constained the maps of grasp quality, rotation angle around Z-axis and the gripper width. The best grasp can be calculated by the grasp map.

- Based on pixels: Fully convolution network [12] has made great achievements in the field of image segmentation. These network models are based on pixelwise detection, which predict a grasp for each pixel.[2]

The method of the sliding windows performs multiple iterations, and the speed of response is slow, making it difficult to meet real-time requirements [1,4,6]. Bounding boxes have a positive effect in the object detection. when used in grasp detection, a big problem is that the groud-truth rectangle is oriented. The orientation matters much more to robotic grasp detection [6]. Thus the oriented anchor box is proposed [11], But the architecture of these model is complex and a large number of model parameters need to be learned.

The development of the fully convolutional neural network provides a new perspective for robotic grasp detection [2]. The FCN models operate on the pixels directly, predicting a grasp for each pixel. Compared with other networks, the FCN models can directly generate grasps instead of sampling and classifying grasp candidates. We propose Attention Grasping Network (AGN) based on a fully convolutional network to predict a grasp for each pixel. As shown in Figure 1, a best grasp is defined as the highest grasp quality in the grasp map.

In summary, our contributions in this paper are as follow:
(1) We propose a light weight Attention Grasping Network

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

(AGN) model based on fully convolutional network. Our experiments show that the FCN will lose some detail features of the input image, which has a negative impact on generating grasp synthesis.
(2) Our proposed model integrates attention mechanisms into a fully convolutional network for generating grasp synthesis. The attention mechanisms make our model focused on salient features of the input image, which increase the accuracy of prediction and the sensitivity of the model.
(3) To evaluate our model,we validate the performance of our model in Cornell Grasping dataset[8]. Our model reaches 97.8% accuracy and the speed of 22ms at a frame.

II. Related Work

Grasp synthesis refers to the problem of finding a stable grasp for given an object [13]. The methodologies about grasp synthesis can be divided into analytic and empirical [14]. Analytic formulations towards grasp synthesis are based on models of geometry, kinematics and dynamics. Empirical methods are based on some exsting grasp experience. Contray to analytic methods, the empirical approaches place more weight on the object representation,pose estimation,feature extract,e.g. However, for grasping unkown objects, the empirical methods,especially techniques like deep learning, are more advanced [8,15].

Lenz el at. [6] first use a sliding window detection framwork to grasp detection. They use netural network as a classifier to determine whether the small patches have good grasps for the object in that image. Moreover,the robotic grasp presented a five-dimension configuration can be prejected to three-dimension grasp vector. The accuracy of their models reaches 75.6% on object-wise split and 73.9% on Image-wise. The method of sliding windows is repeated calculation, so the model runs slowly and can not be used in real-time grasp detection.

Redmon el at. [1] apply a single network once to an image and directly predict grasp coordinates. They dividing the image into grid cells and regressing a grasp rectangle in every grid cell. This method also support to predict several grasp rectangles at the same time. Their model achieves 88% accuracy and runs at 13 frames per second, which promotes the progress of grasp detection technologies. Guo et al. [9] associate each grid cell with serveral horizontal reference rectangles with multiple scales, instead of only one grasp prediction in each grid cell. Thereby, rotation angle is seperated from location regression terms and is predicted by classification. They reach an accuracy of 93.2% on image-wise split of Cornell Grasping Dataset.

Chu et al. [3] propose to incorpoate a grasp region proposal network to generate candidate regions for feature extraction. They reset oriented ground-truth grasp rectangle to have horizontal height and vertical weight. Thus,the object detection such as Faster-RCNN,YOLO and so on could be applied to robotic grasp detection. Their model reaches the accuracy of 96%, and the running speed is faster than its previous models

With the development of robotic grasp detection, it is a very significant problem that large parameters need to be learned in the network model. Morrison et al. [2] propose a light weight based on U-net[16,25]. Their model is an object-independent grasp sythesis model. Grasp pose can be generated from a deep image on a pixelwise basis. however, compared with previous models, their model's accuracy is not higher.

In this paper, inspired by Morrison et al. [2], we employ light weight network to deal with the problem of generating grasp synthesis. Light weight network is easy to train and fast to respond. In order to improve the sensitivity of the model and accuracy of prediction, we integrate attention mechanisms to our models.

III. Grasp Detection Method

A. Problem Definition

In parctise, we define a grasping configuration by $g = \{p, \theta, w, q\}$ as shown in Figure 2 (a), where p is the center cartesian coordinates position of the gripper (x, y, z), θ is the grasp rotation angle around the Z-axis, w is the width of the gripper, and q is the quality of the grasp, which represents the opportunity to successfully grasp.

Assuming that the input is a RGB image, we need to calculate the grasp $\widetilde{g}$ in the image. Then, according to the camera's intrinsic parameters and the parameters of hand-eye calibration, the actual grasp g can be calculated by the following equation.

$$g = T_{RC}(T_{CO}(\widetilde{g})) \tag{1}$$

Where T_{CO} represents the transformation of the object's pixel coordinates into camera coordinates,and T_{RC} transforms from the camera coordinates to robot coordinates.

This paper mainly considers how to use netural network to generate the grasp $\widetilde{g}$ in the image. $\widetilde{g}$ is simailar to g and can be defined as $\widetilde{g} = \left\{\widetilde{p}, \widetilde{\theta}, \widetilde{w}, q\right\}$,where $\widetilde{p} = (u, v)$ is the center point of the gripper's pixel coordinates, $\widetilde{\theta}$ is the rotation angle in image coordinates, and $\widetilde{w}$ is the grasp width in image coordinates.

(a) (b)

Fig. 2. (a): A successful grasp g requires cartesian coordinate (x, y, z), rotation angle θ around Z-axis and gripper width w. (b): Grasp pose $\widetilde{g}$ in RGB images is determined by the coordinate of pixel (u, v), rotation angle $\widetilde{\theta}$ and perceived width $\widetilde{w}$.

B. Attention Grasping Network Framework

We need to use our AGN model to predict a grasp for each pixel in the image, then all of the predicted grasps form a grasp map. We define a function F from a image i to the grasp map D: $D = F(i)$. A best grasp can be calculated by $\widetilde{g}^* = \underset{q}{maxD}$

Compared with the traditional method of robotic grasp detection using CNNs, the FCN has two advantages: one is that it can input any size of input images, without requiring all training images and test images to have the same size. The other is that FCN is more efficient, because it avoids the problem of repeated storage and computational convolution caused by the use of pixel blocks. At the same time, the shortcomings of FCN are obvious: for one thing, the output results are still not precise enough. The up-sampling results are blurred and smooth, and it is insensitive to the details of the image. For another, each pixel is executed without considering the relationship between the pixels,which causes lacking spatial consistency. These have a negetive influence on the generation of the grasp map, as shown in Figure 3.

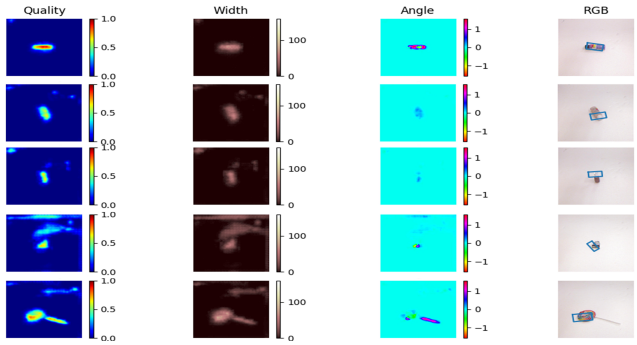

Fig. 3. Grasp map and result learned by general fully convolutional network

Therefore, there are two main ways to solve this problem. One is to use complex CNN structures such as Resnet and VGG in fully convolutional neural network, but this will cause the network required learning a lot of parameters, and it is difficult to satisfy the real-time requirements of grasping. So we adopt the other way, integrating the attention mechanism [17,19] into our model. In order to ensure that the salient features of the image are not lost, our model should focus on learning the features related to robotic grasp detection.

As shown in Figure 4 of our network model, the attention mechanism is incorporated into the fully convolution network. We input a RGB image for end-to-end learning, then we can get the grasp pose $\widetilde{g}$ directly. The process can be as follows: firstly, the image is extracted by the forward convolution network, so as to capture a large enough field of perception. After that the attention mechanism is integrated to suppress the corresponding irrelevant background area in the deconvolution layer, which enlarges the saliency feature and improves the performance of the network.

The attention gate passes through skipping connection to highlight salient features, as shown in Figure 4. The input features f are scaled by attention coefficient α calculated in

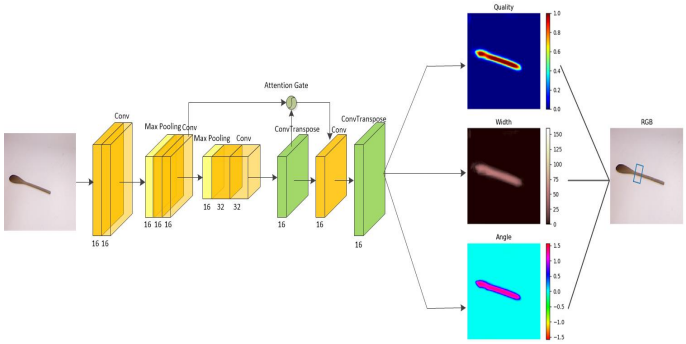

Fig. 4. The full architecture of our AGN model

the attention gate (see Figure 5). The attetion coefficient can be computed by Equation 2. The output of the attention model can be obtained by the element-wise mutiplication of the input feature maps and attention coefficients (see Equation 3)

$$\alpha = \sigma_2(\phi(\sigma_1(W_g(g) + W_x(f))) + b_\phi) \tag{2}$$

$$\widehat{f} = f * \alpha \tag{3}$$

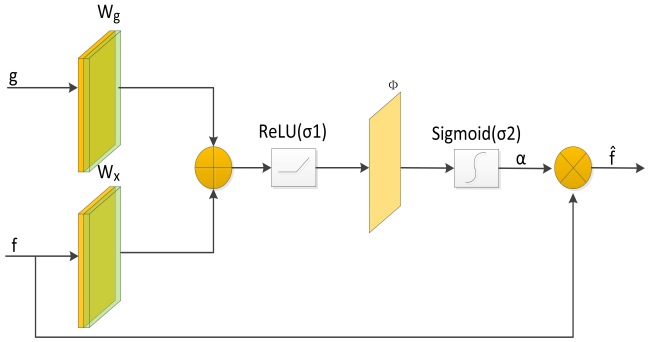

Fig. 5. Schematic of the attention mechanism model. Where g is the gating signal, f is the input feature and $\widehat{f}$ is the output feature.

Our AGN model uses common fetures from the convolutional layers and deconvolutional layers for both recongition and detection, which processses a RGB image in a single pass to predict the objects and generates a good grasp. In addtion, Our model has achieved good performance in Cornell Grasping Dataset [8].

IV. Experiments And Evaluation

To evaluate our network model, we choose Cornell Grasping Dataset [8] similar to other literatures[1,2,3,4,9,10]. The dataset consists 885 RGB-D images of 244 real objects, with 5110 human-labelled positive and 2909 negetive grasps. Each distinct image is labelled with mutiple grasps, which suits our pixelwise grasp representation.

A. Data Processing

In order to estimate of the full grasp map, we use a distinct image to represent one grasp. We assume that 5110 human-labelled positive grasps are right, so we use the random croping, zooming and rotating methods to process the Cornell Grasping Dataset to generate associated grasp map images with positive grasps. Each RGB image corresponds to three grasp feature maps: Quality, Width and Angle.

- Quality: Grasp quality q set to 0 to 1. The ground-truth positive grasp repesents the value of 1,and other pixels are 0.We need to calculate the grasp quality of each pixel in a RGB image. The higher the quality, the higher rate of sucessful grasp.
- Width: We compute the gripper width $\widetilde{w}$ of each pixel in a RGB image, and set the value of width from 0 to 150.
- Angle: The angle of each grasping rectangle is in the range $[-\frac{\pi}{2}, \frac{\pi}{2}]$, so we encode the angle of each pixel in a image from -1 to 1.The range of rotation angle θ around Z-axis in actual grasp is $[-\frac{\pi}{2}, \frac{\pi}{2}]$. The Model predicts the grasp angle $\widetilde{\theta}$, then the θ can be calculated.

B. Training

Our models do not need any pre-training model, and our model has few parameters to learn. For training, our models run on two NVIDIA GTX2080TI with 11GB memory. The dataset splits two part, 90% as training set and 10% as cross validation. The batch size is set as 8, and the epoch is set as 30. We use Adam algorithm with the learning rate 0.001 to optimize our model. The training process can be visualized as follows Figure 6. Our model predicts a grasp for every pixels.

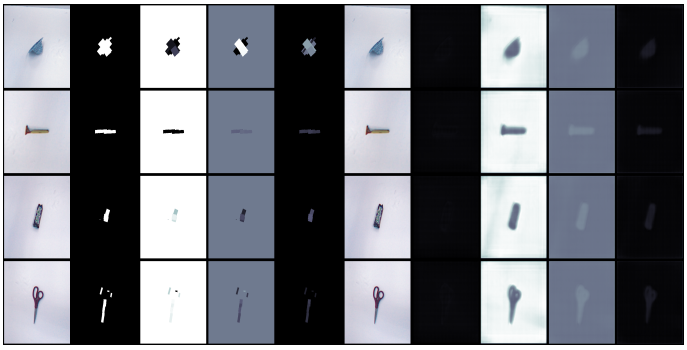

Fig. 6. The process of training

We can see the trainning result on Cornell Grasping Dataset in Figure 6. In order to further explain the positive effect of the model integrated the attention mechanism, we will compare our model with the model without the attention mechanism in the same picture (Figure 7). In terms of the training loss, the model with the attention mechanism (AGN) decreases faster, and the model without the attention mechanism easily overfits. From the perspective of IoU, IoUs of AGN are almost above 0.80, while IoUs of the model without attention mechanism are below 0.80. We could conclude that our integration of attention mechanism greatly enhances the performance of the model.

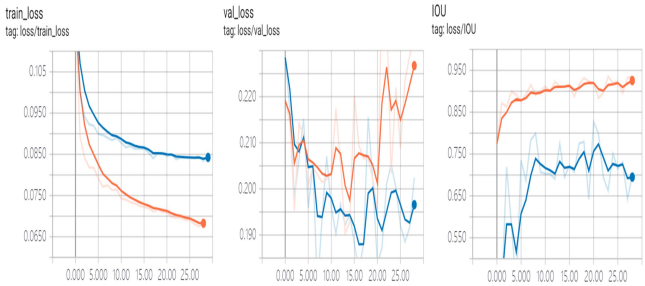

Fig. 7. The result of training. The blue line represents a model with no attention mechanism, and the yellow line represents the model with attention mechanism

C. Evaluation Metrics

When evaluating grasp detection results on Cornell Grasping Dataset, previous work uses two different metrics. One is the point metric, which calculates the distance from the centre of predicted grasp to the centre of each one of the ground truth grasps. If one of these distances is less than some threshold, the grasp is considered as a successful grasp.

The point metric have some problems. Most directly, the metric does not take angle into consideration. So we use the other metric is the rectangle metric. A predicted grasp is regarded as a successful grasp if both:

- The grasp angle is within 30° of the ground truth grasp.
- The Jaccard index of predicted grasp and the ground truth is greater than 25%.

Where the Jaccard index is defined as:

$$J(T, \hat{T}) = \frac{T \cap \hat{T}}{T \cup \hat{T}} \tag{4}$$

The Jaccard index is similar to the metrics (IoU) used in object detection. The T is the area of the predicted grasp rectangle, and $\hat{T}$ is the ground truth grasp rectangle. Jaccard index is the insersection of these two rectangles divided by the union of these two rectangles.

V. Results

Our model is a light weight network model, which does not require pre-training model. Compared with other CNNs such as Resnet, Vgg and so on, the parameters it needs to learn are seldom at 7.1k. Furthermore, our AGN model performs well on Cornell Grasping Dataset. Combining attention mechanism with fully convolutional network has achieved good results, as shown in Figure 8. The grasp rectangles of learning errors in general fully convolutional network has been corrected in our model. It is easy to find that the three feature maps of an image are clear and accurate after integrating the attention mechanism. A robotic grasp model proposed requires possessing practicality. Robotic grasp detection needs to satisfy three characteristics: stability, accuracy and rapidity in complex industrial environment.

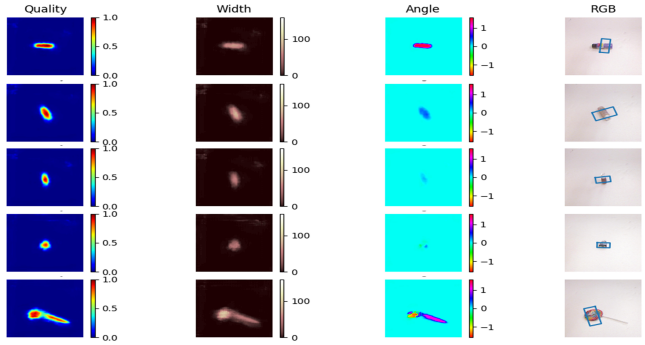

Fig. 8. Grasp map and result learned by our AGN models

A. Stability

Stability is an important condition in robot grasping. It determines whether the robot can have robust performance in complex environment. We test our model performance by changing the jaccard threshold and compare it with other models. The result can been see in Tabel1. Our model has the highest accuracy without varying with the change of Jaccard indexes.The performance of outcomes decreases with stricter Jaccard indexes but maintains competitiveness even at the Jaccard threshold 0.45. Compared to other models in the Table 1, our model is not only more accurate but also has a strong interference ability.

TABLE I
PREDICTION ACCURACY (%) AT DIFFERENT JACCARD THRESHOLDS

split	0.25	0.30	0.35	0.4	0.45
Chu et al. [3]	96.0	94.9	92.1	84.7	*
Morrison et al. [2]	85.4	83.1	77.5	75.3	70.8
Our AGN	**97.8**	**97.8**	94.4	92.1	84.3

B. Accuracy and Rapidity

Accuracy is undoubtedly an important factor in measuring the quality of a model. The Real-time performance is important, which determines whether our model can be online. Due to the real-time requirements in the industrial environment, we do not pursue the ultimate in accuracy. We make the accuracy as high as possible when the model responds quickly. Tabel 2 contains the accuracy and speed of prediction. Compared to other models, our models are competitive. We have the higher accuracy and the faster response speed.

VI. Conclusion

In this paper, we propose an Attention Grasping Network (AGN) model to generate a grasp. We predict a grasp for each pixel based on our model. These grasps can form a grasp map, and we find the highest quality points in the grasp map as the optimal grasp. Compared with the methods of sliding windows and bounding boxes, our method can effectively avoid repeated calculation and get more accurate rotation angle. In addtion, we integrate an attention mechanism to make our model focused on the salient features of the input image, which increases the sensitivity of the model and accuracy of prediction. Furthermore, our model architecure is able to directly generate grasp poses from a RGB image on a pixelwise basis instead of sampling and classifying grasp candidates like other CNN models.

TABLE II
ACCURACY OF DIFFERENT METHONS(JACCARD THRESHOLD IS 25%)

approach	Accuracy(%)	speed(fps)
Jiang et al. [4]	60.5	0.02
Lenz et al. [6]	75.6	0.07
Redmon et al. [1]	88.0	3.31
Guo et al. [9]	93.2	-
Kumra et al. [10]	88.9	16.03
Fu-jen et al. [3]	96.0	17.24
Douglas et al. [2]	85.4	**52.63**
Our AGN	**97.8**	45.46

We evaluate our AGN model on the Cornell Grasping dataset. The training result indicates that the training loss of our model decreases faster than the model without attention mechanisms, and our model is not easy to overfit.Moreover, our model has a higher IoU than that of the model without attention mechanisms. The testing result shows that our model can maintains competitiveness with the stricter Jaccard indexes. Our model has 84.3% accuracy even at the Jaccard threshold 0.45. And our model achieve 97.8% grasp success rate respectively on the rectangle metric. At the same time, our model has a relatively fast response speed.

References

[1] Redmon J, Angelova A. Real-time grasp detection using convolutional neural networks[C]// IEEE International Conference on Robotics & Automation. 2015.
[2] Morrison D , Corke P , Leitner, Jrgen. Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach[J]. 2018.
[3] Fu-Jen Chu, Xu R, Patricio A. Vela. Real-World Multiobject, Multigrasp Detection[J]. IEEE Robotics and Automation Letters, 2018, 3(4):3355-3362.
[4] Yun, Jiang, S. Moseson, and A. Saxena. "Efficient grasping from RGBD images: Learning using a new rectangle representation." IEEE International Conference on Robotics & Automation 2011.
[5] Oktay O , Schlemper J , Folgoc L L , et al. Attention U-Net: Learning Where to Look for the Pancreas[J]. 2018.
[6] Lenz I , Lee H , Saxena A . Deep Learning for Detecting Robotic Grasps[J]. 2013.
[7] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.
[8] Lenz I , Lee H , Saxena A . Deep Learning for Detecting Robotic Grasps[J]. 2013.
[9] Di, Guo, et al. "A hybrid deep architecture for robotic grasp detection." IEEE International Conference on Robotics & Automation 2017.
[10] Bicchi A, Kumar V. Robotic grasping and contact: a review[C]// IEEE International Conference on Robotics & Automation. 2002.
[11] Zhang H, Zhou X, Lan X, et al. A Real-time Robotic Grasp Approach with Oriented Anchor Box[J]. 2018.
[12] Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2014, 39(4):640-651.
[13] Bohg J , Morales A , Asfour T , et al. Data-Driven Grasp SynthesisA Survey[J]. IEEE Transactions on Robotics, 2014, 30(2):289-309.

[14] K. Shimoga, Robot Grasp Synthesis Algorithms: A Survey,Int. Jour.of Robotic Research , vol. 15, no. 3, pp. 230266, 1996.
[15] Pinto L, Gupta A. Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours[C]// IEEE International Conference on Robotics & Automation. 2016.
[16] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation[C]// International Conference on Medical Image Computing & Computer-assisted Intervention. 2015.
[17] Kumra S , Kanan C . Robotic Grasp Detection using Deep Convolutional Neural Networks[J]. 2016.
[18] Symes E, Tucker M, Ellis R, et al. Grasp preparation improves change detection for congruent objects[J]. J Exp Psychol Hum Percept Perform, 2008, 34(4):854-871.
[19] Treisman A M, Gelade G. A feature-integration theory of attention[J]. Cognitive Psychology, 1980, 12(1):97-136.
[20] Posner M I, Snyder C R, Davidson B J. Attention and detection of signals[J]. Journal of Experimental Psychology, 1980, 109(2):160-174.
[21] Johns E, Leutenegger S, Davison A J. Deep Learning a Grasp Function for Grasping under Gripper Pose Uncertainty[J]. 2016.
[22] Lecun Y, Bengio Y, Hinton G. Deep learning.[J]. Nature, 2015, 521(7553):436.
[23] Deng L, Yu D. Deep Learning: Methods and Applications[J]. Foundations & Trends in Signal Processing, 2014, 7(3):197-387.
[24] Schmidhuber J. Deep Learning in neural networks: An overview.[J]. Neural Netw, 2015, 61:85-117.
[25] Dai J, Li Y, He K, et al. R-FCN: object detection via region-based fully convolutional networks[J]. 2016.