International Journal of Control, Automation, and Systems 20(12) (2022) 3996-4004
http://dx.doi.org/10.1007/s12555-021-0929-8

ISSN:1598-6446 eISSN:2005-4092
http://www.springer.com/12555

RANET: A Grasp Generative Residual Attention Network for Robotic Grasping Detection

Qian-Qian Hong , Liang Yang* , and Bi Zeng

Abstract: This paper presents a novel grasp generative residual attention network (RANET) for generating antipodal robotic grasp from multi-modal images with the pixel-wise method. To strengthen the generalization ability of unknown objects, this paper proposed a new structure that differs from the previous grasp generative network in that it additionally integrates a coordinate attention mechanism and a symmetrical skip connection, respectively. Using the coordinate attention module to emphasize meaningful information of the feature map and the symmetrical skip connection to remain more fine-grained details of feature. Moreover, a multi atrous convolution module is included in the structure to capture more high-level information, while a hypercolumn feature fusion method is incorporated for getting the best from the complementation of different layers' features. Through evaluation on public datasets, the result demonstrates that we achieve 98.9% accuracy on the Cornell dataset which is the state-of-the-art performance with real-time speed($\sim$ 17 ms), meanwhile, we represent a 93.9% accuracy performance on the Jacquard dataset.

Keywords: Convolutional neural networks, deep learning, grasping detection, vision.

1. INTRODUCTION

The research of robotic grasping has attracted a wide range of interests in the past several years, owing to the various applications of this field, such as medicine [1,2], domestic chores [3], and industrial manufacture [4,5]. Although much progresses have been made in robotic grasping field, it is still challenging for a robot to decide a appropriate grasp configuration (grasp pose and grasp position), which is used to accurately capture the unknown object in the practical application environment.

In the previous work, the hand-crafted-based approaches [6-8] were employed to extract features, which are tedious and time-consuming. Along with the recent years' remarkable progress of Deep Learning (DL), deep learning methods performing a greatly improved result on grasp detection problem [9-22]. These deep learning grasping detection approaches can be broadly grouped under two categories, which are Classification-based method and Regression-based method. The Classification-based methods [10-16] employ convolutional neural network (CNN) to rank numbers of grasp position candidates, which requires additional computational time and computing resources. The Regression-based methods [19-22] directly yield the coordinates and orientation of grasp from CNN and performs well in grasping detection. Nonetheless, Regression-based methods uses the average of possible grasp configurations as output for an object may lead to invalid grasp.

To address the above issue, we present a novel grasp generative residual attention network (RANET) for robotic grasping detection. Motivated by generative-based methods [23,24], our method directly generates pixel-wise grasp configuration from the outputs of network, which are the quality, width, and angle of grasp, as presented in Fig. 1. The framework of our scheme consists of encoder path, bottle-neck layer, and decoder path. To enhance the feature representational ability of our method, the residual unit is implemented to extract the features of target in the encoder path. Subsequently, by integrating

Manuscript received November 1, 2021; revised January 12, 2022; accepted January 25, 2022. Recommended by Editor Choon Ki Ahn. This work is supported by the Science and Technology Foundation of Guangdong Province under Grant 2019B090910001 and 2021A0101180005, in part by the National Natural Science Foundation of China under Grant 61941301, Grant 61803090, Grant 11771102, and Grant 61573108, in part by China Postdoctoral Science Foundation under Grant 2018M633353, in part by the Special Program for Key Field of Guangdong Colleges under Grant 2019KZDZX1037, in part by the Natural Science Foundation of Guangdong Province under Grant 2019A1515012109, 2021A030310668 and 2022A1515010178, in part by the Scientific and Technical Supporting Programs of Sichuan Province under Grant 2017GZ0391, 2019YFG0352 and 2017GZ0392. The authors express our deep appreciations to the Editor-in-Chief, Associate Editor, and anonymous reviewers for their constructive and professional comments to improve this paper.

Qian-Qian Hong and LiangYang are with the School of Computer Engineering, University of Electronic Science and Technology of China Zhongshan Institute, Zhongshan, Guangdong 528402, China (e-mails: adaqao@foxmail.com, alex_yangliang@foxmail.com). Qian-Qian Hong and Bi Zeng are with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou, Guangdong 510006, China (e-mail: zb9215@gdut.edu.cn).
* Corresponding author.

Springer

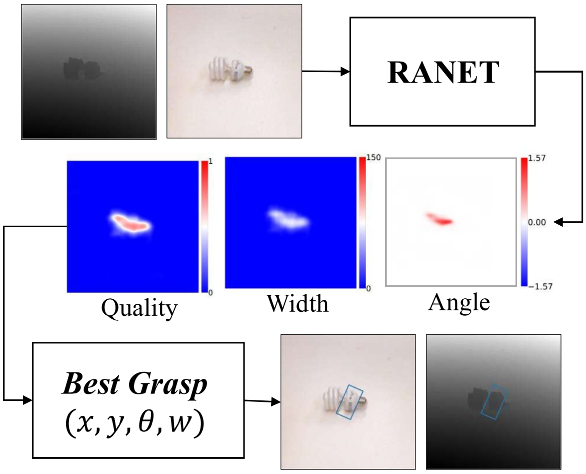


Fig. 1. The pipeline of real-time grasp detection. Our method directly generates pixel-wise grasp configuration (position, angle, width) from provided multi-modal images.

multi atrous convolution module with bottle-neck layer, our model can extracts the features from multiple scales which provide more global and local context information. Further, to recover the lost spatial feature information of down-sampling phase, each decoder module integrates the feature of the corresponding encoder module through symmetrical skip connection. Meanwhile, in combination with coordinate attention mechanism, the decoder module is newly constructed to assign a higher priority to the more important feature. In addition, to get the best from mutual complementation between the semantic information that extracted from high-level layers and the position information from low-level layers, we novelly incorporate the hypercolumn-based feature fusion method into decoder path.

To validate our method, we train and evaluate RANET on the Cornell Grasping Dataset and the Jacquard Dataset [25]. Through evaluation on these public datasets, the results show that our method achieves state-of-the-art 98.9% accuracy on the Cornell dataset and 93.9% accuracy performance on the Jacquard dataset with real-time speed ($\sim$ 17 ms).

Below is the contribution of our work.

1) Compared with the previous grasping detection methods [23,24], the newly proposed one by us additionally incorporates coordinate attention mechanism and symmetrical skip connection, which improves the performance of grasping detection on unknown objects. Specifically, the coordinate attention mechanism is leveraged to extract more meaningful contextual information of features, meanwhile, the fine-grained details of target object which is lost during the down-sampling phase are recovered from the symmetrical skip connection.
2) By resorting to the hypercolumn feature fusion method, RANET gets the best from the complementation between the semantics-aware and the position-sensitive information of features. To further enhance the discriminability of features, a multi atrous convolution module is presented to extract features from multiple receptive fields.
3) By taking advantages of the newly designed architecture, the proposed grasp generative residual attention network achieves the state-of-the-art performance on the Cornell dataset and present 93.9% accuracy on the Jacquard dataset.

2. RELATED WORK

Robotic grasping problem has been an object of research for decades. In the previous work, the methods of grasping mainly based on hand-crafted features are time-consuming and display limited generalization ability to unknown targets [26-29]. With the application of the deep learning methods, a greatly improved performance is displayed on grasp detection problem.

2.1. Classification-based method

The classification-based method employs a classifier to chose the grasp with the highest score. Lenz *et al.* [10] firstly proposed a network that trained by a sparse auto-encoder with multi-modal images to detection grasp by using sliding window. Similar to [10], Wang *et al.* [14] proposed a convolutional neural network based multi-modal classification network to identify candidate object region. In [11], a multi-stage learning method is presented to train a CNN-based network which predicts grasping location and angle by sample image patch. The presented approach of Pinto displays better generalization ability of unknown objects than previous works. With recourse of a classification-based spatial transformer network (STN) that trained by multi-stage method, Park and Chun [12] achieves 86.9% accuracy on the public robotic grasp detection dataset. Furthermore, Chu *et al.* [13] presented a CNN-based classifier with null hypothesis competition to predict graspable location on RGB-D images and performing well on Cornell dataset. These classification-based methods present a comparatively good accuracy on grasping detection, but not efficient both in terms of computation time and memory during grasp detection.

2.2. Regression-based method

The regression-based method utilizes convolutional neural network to yield the position and angle of grasp directly. In [19], a large neural network with single-stage regression method is proposed to predict grasp coordinate, meanwhile, they get 88% accuracy and 76 ms prediction time on the Cornell dataset. Following the research of [19], Watson *et al.* [30] performed the grasping experiment on real objects. In [20], a hybrid deep architecture is presented to predict the grasp configurations by combining the visual and tactile information. At the same time, Kumra and Kanan [21] designed a network that uses ResNet as feature extractor to predict the grasp from multi-modal data. Furthermore, to improve the feature expression, a multi-modal fusion architecture that integrates atrous convolution and novel loss function is constructed by Zhang *et al.* [22]. These regression-based methods performed well in grasping detection, nonetheless, yield the grasp configuration from the average of possible grasps would lead to unreasonable grasp.

2.3. Generative-based method

The generative-based method directly generates the position and angle of grasp from pixels. In [23], a generative-based convolutional neural network (GGCNN) is proposed to predict the grasp pose and grasp position from the pixels of depth image. Following the research of [23], Chalvatzaki *et al.* [31] presented an orientation-attentive method to augment feature representation and predict pixel-wise grasp configurations from depth image. Meanwhile, Xu *et al.* [32] proposed a oriented diameter circle representation method to predict the grasp from point cloud. Subsequently, Kumra *et al.* [33] embedded the residual module on the bottle-neck layer of generative grasping network to improve the ability of feature extraction. At the same time, a segment-based fully convolution network was presented in [34], which proposes pose estimation from RGB-D image, while achieving 91.02% accuracy on the Cornell Dataset. Moreover, Dolezel *et al.* [35] presented an attention squeeze parallel network with image transformation approach, which can be applied on the single board computer. Based on the similar concept of generative-based method, our work further develop on this field.

3. PROBLEM STATEMENT

In this section, our work formulate the robotic grasping problem as predict an antipodal grasp from given n-channel multi-modal images of unknown object which is lay on a planar surface. Instead of using oriented rectangle to evaluate grasp [36], we follow up the previous work which is improved by Morrison *et al.* [23] that using the contact points to evaluate the result of model. Similar to the work of Morrison, the grasp in world coordinates is defined as

$$g_r = (p_r, \theta_r, w_r, q_r), \tag{1}$$

where $p_r = (x_r, y_r, z_r)$ is the center point of robot gripper's tip, θ_r is a angle within $(-\frac{\pi}{2}, \frac{\pi}{2})$ that represents the rotation of robot's gripper around z-axis. The required width of gripper is denoted by w_r. The last element q_r represents a scalar grasp quality score $\in [0,\ 1]$ that corresponds to the probability of a successful grasp on every pixels of input image.

Further, given the n-channels image that we use as input, we denote the grasp as

$$g_i = (p_i, \theta_i, w_i, q_i), \tag{2}$$

where $p_i = (x_i, y_i)$ is the center coordinate of grasp in image, the rotation angle of camera frame around z-axis is denoted by θ_i, meanwhile, w_i refers to the required width of gripper which is limited by the max open range of the gripper.

Moreover, the definition of grasp g can be expended to the grasp set G when we have multi grasps in the input image, which is described by

$$G = (\Theta, W, Q) \in R^{3 \cdot H \cdot W}, \tag{3}$$

where Θ, W, Q represent each result $R_i^{H \cdot W}$ from the results set of our model $R^{3 \cdot H \cdot W}$, which contains grasp angle, grasp width, and grasp quality at every pixel respectively. Further, the best grasp $\tilde{g}_i$ of the grasp set G is represented by $\tilde{g}_i = \max_Q(G)$.

4. FRAMEWORK

In this section, we first present the general structure of our model. Subsequently, we will further discussion the multi atrous convolution module, hypercolumn-based feature fusion method, and coordinate attention module on the next subsections.

4.1. Overview

To make a balance between accuracy and detection speed, a new grasp generative residual attention network is proposed as illustrated in Fig. 2, in which static muti-model image is used as input.

The main structure of the grasp generative residual attention network is motivated by the encoder-decoder architecture which can be divided into three sections, which is the encoder path, bottle-neck layer, decoder path. To take advantage of the powerful representational ability of the residual method, we employ residual unit as our feature encoder module. Moreover, we observe that the main-path kernel size of the first encoder module with $\{7, 5\}$ and set the kernel size of the bypass convolution layer to 3 that gets the better performance during detection. The structure of feature encoder module is shown in Fig. 3.

After extracting features from the encoder path, we feed the feature map into bottle-neck layer which is augmented by a multi atrous convolution module that is applied to provides more local and global context information about

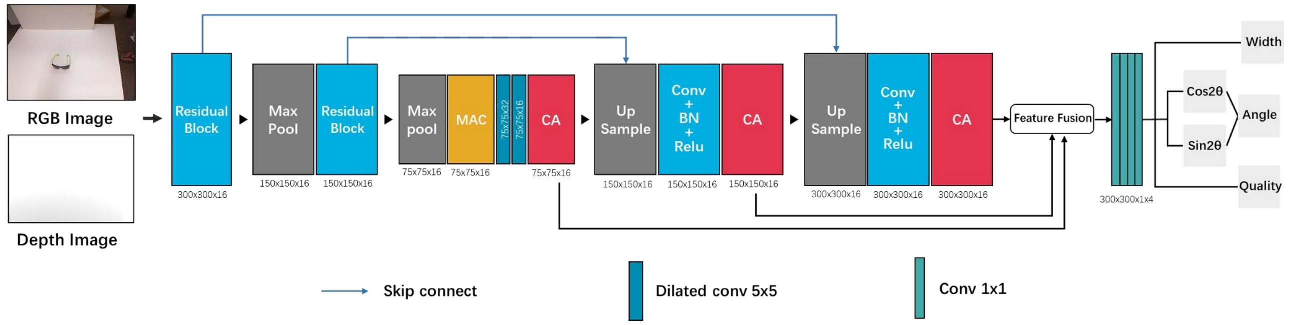


Fig. 2. An overview of the grasp generative residual attention network's architecture.

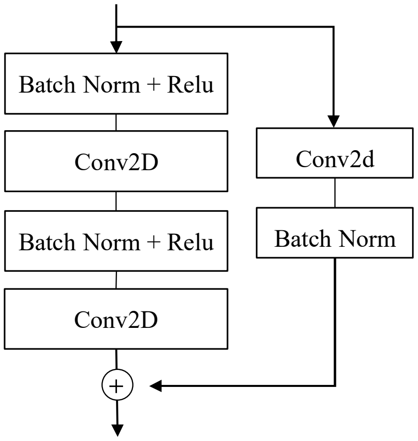


Fig. 3. The structure of the residual based encoder module.

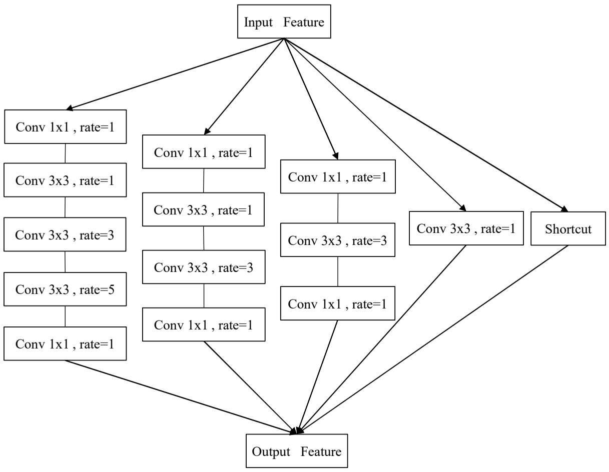


Fig. 4. The structure of multi atrous convlution module.

the target by extracting features at multiple scales. Different from the previous layers which mainly focus on enhancement the generalization ability of our model, the decoder path pays more attention to exploit the advantage of extracted features. The coordinate attention module is used to filter and emphasize meaningful features information of decoder module. Further, we integrate the hypercolumn feature fusion method into decoder-path, other than use the feature map of the last decoder module as feature representation, we concatenate the features from different layers as the input to predict pixel-wise grasp configurations.

4.2. Multi atrous convolution module

To tackle the variation of target size on the detection process, we present a multi atrous convolution module (MAC) which is motivated by the architecture of Inception-res and receptive field block, we use multi-branch to get the features of different filed-of-view that enhances the generalization ability of our model.

After getting the features from encoder module, the extracted features would go through four bypass-branch with different amounts of atrous convolution layers. To capture multi-scale feature information, the kernel size of atrous convolution is set to $\{3,1\}$ with different dilation rates $\{1,3,5\}$. Follow the atrous convolution layer, the features from bypass-branch are activated by Relu and subsequently merged by an add operation with the original feature as the output feature as displayed in Fig. 4.

4.3. Hypercolumn-based feature fusion method

The pool operation on the feature extracting process leads to the loss of feature information is one of the challenging problems of generative-based model. The loss of information makes the feature map of the last layer too coarse to make a precise localization.

To address this problem, we integrate the hypercolumn-based feature fusion method with our model which combines the complementation of features from different layers. As known to us, the previous layers on the network are more precisely in localization while the latter layers preserve more semantics information. Hence, the fusion between the features of different layer make a good trader-off between the semantics-aware information and the position-sensitive information. The hypercolumn-based feature fusion method can be formulated as

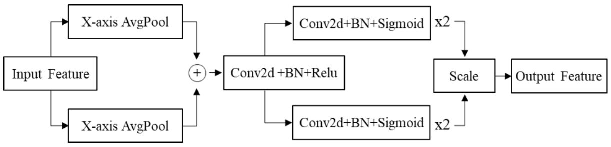


Fig. 5. The structure of coordinate attention module.

$$f_{out} = \sum_{i} concat(f_{out}, U(\alpha_i, f_i)), \tag{4}$$

where the feature map of the ith layer is denoted as f_i, α is a ratio represents the size scale between the last layer and the ith layer. The bilinear upsampling function U takes α_i and f_i as input and the output feature of upsample function will be concatenated with the feature of the last coordinate attention module that treated as the final feature map to predict grasp.

4.4. Coordinate attention module

The effectiveness of attention mechanism has been proven on vision tasks, similar to the attention with human perception system, the attention mechanism helps neural network to focus on the important part of feature map.

Hence, we introduce the coordinate attention module [37] that helps our model focus more on the important message rather than unnecessary one with slight additional computational cost, as shown in Fig. 5.

In order to keep more precise position, coordinate attention module decomposes channel attention transform process into x and y spatial direction which is accomplished by average pool in horizontal axis and vertical axis. Along with the following convolution layer, these feature maps that with 2-direction information embedded will be encoded into a pair of attention map and subsequently multiply with the original feature to enhance the representation of the target of interest. The output feature X of the coordinate attention module can be formulated as

$$X_{out} = A_h \cdot X_{in} + A_v \cdot X_{in}, \tag{5}$$

where A_h, A_v are denoted the attention map of horizontal and vertical direction separately.

5. EXPERIMENTS

To evaluate the proposed method, we carry out experiments on the Cornell dataset and the Jacquard dataset. In the next subsections, we first present the detail of datasets. Then, the define of grasp representation on our work will be introduced. Further, we display the implementation details of our model.

5.1. Dataset

Cornell grasp dataset: This dataset contains 885 images that captures from 240 different objects. Each picture has 640 x 480 pixels and three corresponding files which are depth image, negative grasp representation file, and positive grasp representation file. The negative grasp representation file is not using during the training and evaluating phase. Some examples from the Cornell dataset is presented in Fig. 6.

Jacquard grasp dataset: This dataset has 54485 images that captures from 11000 different objects. Each picture has 1024 × 1024 pixels and two corresponding files which are depth image and grasp representation file. The sample images is presented in Fig. 7.

5.2. Grasp representation

As discussed in the previous part, the grasp representation of our method on the image is defined as $g_i = (p_i, \theta_i, w_i, q_i)$, where θ_r, w_r correspond to the grasp angle map and the grasp width map generated by model, meanwhile, q_r is represented by the grasp quality map and the p_i denotes the highest grasp quality score point of q_r. The example of grasp maps are illustrated in Fig. 8.

5.3. Training implementation details

In the phase of preprocessing, we resize the input image from the Jacquard dataset and the Cornell dataset to 400 x 400 pixel and 300 x 300, respectively. Meanwhile, we split 90% of the Cornell dataset as training set and the remaining 10% of Cornell dataset as testing set. Addition-

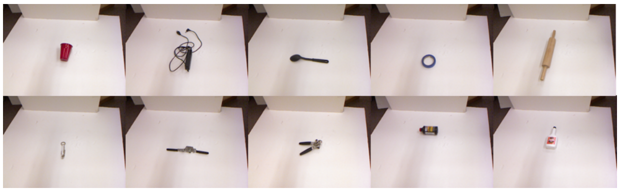


Fig. 6. The sample images from Cornell dataset.

Fig. 7. The sample images from Jacquard dataset.

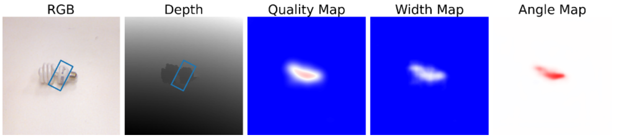


Fig. 8. The feature maps of grasp representation, which are grasp quality, grasp angle, grasp width, and the results on RGB-D images.

ally, the image from the Cornell dataset is augmented by image crop, random zoom, and rotation operation, cause of the limited quantity of the Cornell dataset.

The weights of encoder and decoder modules were all initialized using xavier uniform and the adaptive moment estimation algorithms (Adam) is employed as optimizer with an fixed initial learning rate of 0.0001. At the same time, we set the batch size of data to 64 and train our model for 50 epochs. Further, the Huber loss is adopted on our method as loss function, which is presented below

$$loss(y_{pred}, \hat{y}_{gt}) = \frac{1}{n}\sum_{i} z_i, \tag{6}$$

where the predict values and the ground truth values of grasp set G, which are represented by (Θ, W, Q), are denoted by y_{pred} and $\hat{y}_{gt}$. While z_i is given by

$$z_i = \begin{cases} 0.5(y_{pred} - \hat{y}_{gt})^2 & \text{if } |y_{pred} - \hat{y}_{gt}| < 1, \\ |y_{pred} - \hat{y}_{gt}| - 0.5 & \text{otherwise.} \end{cases} \tag{7}$$

The total loss is defined as a value that add the loss of q, θ, w together, which can be formulated as follows:

$$Loss_{total} = loss_{\Theta} + loss_{W} + loss_{Q}. \tag{8}$$

5.4. Evaluation

To better evaluate the performance of our model, following the previous wrok [36], the rectangle metric is implemented to assess the validity of the result. Based on the definition of rectangle metric, we consider a grasp representation that generated by our model as a successful grasp only if the offset of angle between the predict grasp rectangle and the ground truth rectangle is lower than 30° while the score of intersection over union between the predicted grasp representation and the ground truth representation higher than 25%.

Further, we employ the Gaussian kernel filter with the output images of our model before evaluation, which is helpful to clean up the outlier of output map while making the result more robust. Specifically, different σ setting of gaussian function is adopted on different feature map. {8, 2, 1} corresponded to the σ of quality map, angle map, and width map, respectively.

6. EXPERIMENTAL RESULT

6.1. Experimental setup

As illustrated in Fig. 9, we present the experimental results of our method. Following the setup of previous works, we apply the image-wise (IW) data split method and the object-wise (OW) data split method to deal with data, the detail is presented below.

Image wise split: This method evaluates the generalization ability of network on orientation change and size variation of objects, we randomly shuffle the data from dataset to train and evaluate our model.

Object wise split: This method focus on evaluate the generalization ability of model to new objects and divided dataset based on the object sets.

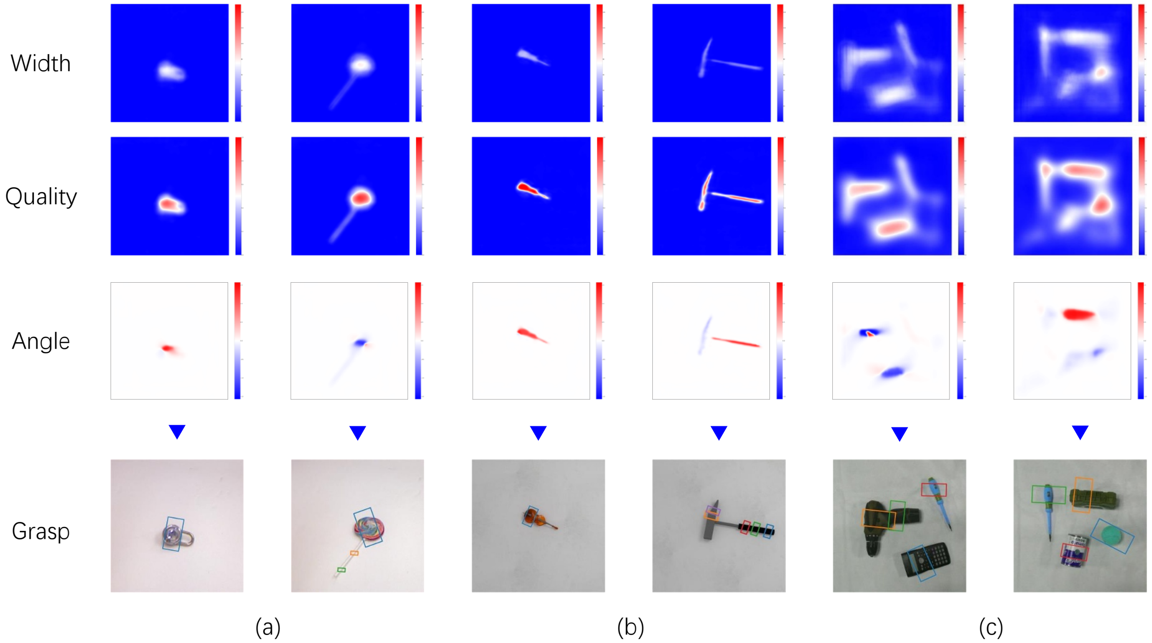


Fig. 9. Experimental results. Width, quality, and angle are the output of our model to infer the grasp configuration. (a) The results on the unknown objects of the Cornell dataset. (b) The results on the unknown objects of the Jacquard dataset. (c) The results on unknown objects in clutter.

In addition, we perform ablation experiments on the Cornell dataset to assess the impact of each module on our model. The training and evaluating process of our model are based on the Linux 16.04 system with single graphics card (Tesla V100, 32GB) and all experiments is implemented by Pytorch.

6.2. Ablation experiments on the Cornell dataset

To improve the generalization ability of our method on unknown objects, we integrate the multi atrous convolution module (MAC) , the coordinate attention module (CA) and the hypercolumn-based feature fusion method (HC) into our model. To validate the performance of presented modules, we conduct experiments with different setting of our model, the results are summarized in Table 1.

Compared with the GGCNN, our generative residual network with attention module and symmetrical skip connection makes a great improvement which takes 19.3% accuracy promotion on the Cornell dataset. Further, we adopt the hypercolumn-based fusion method and multi atrous convolution module on our model separately that both achieve 2.3% accuracy improvement compare to the original RANET. These results show that our model benefits from the abundant information of feature fusion method while the multi-scale receptive fields and atrous convolution greatly enhance the feature description ability of RANET. Further, we combine all these methods with our model, we obtain the 4.6% promotion of accuracy while reaching the state-of-the-art performance. To evaluate the performance of coordinate attention module, we combine the multi atrous convolution module and the hypercolumn-based feature fusion method with RANET separately that without integrate the attention module. Our experimental result shows that by combining the attention module to enhance the representation of the target of interest, other methods take 2.3% and 1.1% accuracy promotion, respectively.

6.3. Results on the Cornell dataset

For the Cornell dataset, We evaluate our model with the object-wise (OW) data split method and the image-wise (IW) data split method. Table 2 compares the accuracy and speed results of several previous works and our model on Cornell dataset. Our method employs multiple modalities data as input.

Table 1. Ablation experiments on the Cornell dataset.

Model	HC	MAC	CA	Accuracy (%)
GGCNN				75
RANET			√	94.3
RANET		√	√	96.6
RANET	√		√	96.6
RANET	√	√	√	98.9
RANET		√		94.3
RANET	√			95.5

As shown in Table 2, our model achieves state-of-the-art accuracy performance of 98.9% from multi-modal input images with real-time grasp detection speed ($\sim$ 17 ms), which represents that our model has the better generalization ability to unknown object than previous work while displaying the ability to be applied in real-time environment. Additionally, the same object could have different efficient grasp configuration to define a grasp pose in the practical application environment, our approach also displays the ability to directly generate multiple grasp configuration candidates without adding others complex procedures.

6.4. Results on the Jacquard dataset

We further compare our work with the existing algorithm on the Jacquard dataset. The data augment methods are not applied on the Jacquard dataset, cause of the data size of Jacquard dataset is far bigger than Cornell dataset. Our method splits 95% of the Jacquard dataset as training set and the remaining as testing set. The results are listed in Table 3. Compare with the GGCNN, our approach achieves 9.9% accuracy promotion on the Jacquard dataset.

6.5. Objects in clutter

As shown in Fig. 9, multi-object scene is common in the practical application environment. Hence, the model that trained on the Cornell dataset is employed to predict the grasp of objects in clutter. Our model demonstrates the generalization ability on objects in clutter, despite each training image of the training set only contains single object. To obtain the more robust and less redundancy grasps in the multi-objects scene, we set the min-distance between each potential grasp to 10 and grasp quality threshold to 0.3.

Table 2. Results on the Cornell dataset.

Author	Algorithm	Speed (ms)	Accuracy (%)	
			IW	OW
Jiang *et al.* [36]	Fast Search	5000	60.5	58.3
Morrison *et al.* [23]	GGCNN	19	73.0	69.0
Lenz *et al.* [10]	SAE, struce. reg.	1350	73.9	75.6
Redmon *et al.* [19]	AlexNet	76	88.0	87.1
Kumra *et al.* [21]	ResNet-50	103	89.2	88.9
Zhang *et al.* [38]	ROIGD, ResNet-101	25	92.3	91.7
Wang *et al.* [34]	Efficient FC	21	94.2	91.0
Chu *et al.* [13]	VGG16	18	95.5	96.1
Kumra *et al.* [33]	GR-ConvNet	20	97.7	96.6
Ours	RANET - D	17	94.4	96.6
	RANET - RGB	17	97.7	96.6
	RANET - RGBD	18	**98.9**	**97.7**

Table 3. Results on the Jacquard dataset.

Author	Algorithm	Accuracy (%)
Morrison *et al*. [24]	GGCNN	84
Depierre *et al*. [25]	Jacquard	72.42
Wang *et al*. [34]	Efficient FC	92.83
Zhang *et al*. [38]	ROIGD, ResNet-101	93.5
Ours	RANET	93.9

7. CONCLUSION

This paper addresses the problem of robotics grasping detection for unknown objects. A grasp generative residual attention network with coordinate attention mechanism and symmetrical skip connection is newly proposed to directly generate pixel-wise grasp configuration. To further strengthen the generalization ability of our model for unknown objects, a multi atrous convolution module is presented while a hypercolumn feature fusion method is novelly embedded in the structure to get the best from the complementation between the feature of different layers. Furthermore, by taking advantages of the newly designed architecture, our method achieves the state-of-the-art performance on Cornell dataset with real-time speed.

In the next stage, there are still lots of challenging work, for instance, the field of multi-modal feature fusion methods. It is worthy to further exploit the complementation between multiple modalities data. Meanwhile, it would be more valuable to further extend our work to different types of grasping, like suction grasping, multi-fingers grasping.

REFERENCES

[1] R. H. Taylor, A. Menciassi, G. Fichtinger, P. Fiorini, and P. Dario, "Medical robotics and computer-integrated surgery," *Springer Handbook of Robotics*, pp. 1657-1684, 2016.

[2] D. Choi, S. H. Kim, W. Lee, S. Kang, and K. Kim, "Development and preclinical trials of a surgical robot system for endoscopic endonasal transsphenoidal surgery," *International Journal of Control, Automation, and Systems*, vol. 19, no. 3, pp. 1352-1362, 2021.

[3] Y. Li, Y. Yue, D. Xu, E. Grinspun, and P. K. Allen, "Folding deformable objects using predictive simulation and trajectory optimization," *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 6000-6006, IEEE, 2015.

[4] M. Saadat and P. Nan, "Industrial applications of automatic manipulation of flexible materials," *Industrial Robot: An International Journal*, vol. 29, no. 5, pp. 434-442, 2002.

[5] J.-K. Oh, S. Lee, and C.-H. Lee, "Stereo vision based automation for a bin-picking solution," *International Journal of Control, Automation, and Systems*, vol. 10, no. 2, pp. 362-373, 2012.

[6] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Proc. of European Conference on Computer Vision*, pp. 404-417, Springer, 2006.

[7] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *Proc of International Conference on Computer Vision*, pp. 2564-2571, IEEE, 2011.

[8] S. Salti, F. Tombari, and L. Di Stefano, "SHOT: Unique signatures of histograms for surface and texture description," *Computer Vision and Image Understanding*, vol. 125, pp. 251-264, 2014.

[9] L. Chen, P. Huang, Y. Li, and Z. Meng, "Detecting graspable rectangles of objects in robotic grasping," *International Journal of Control, Automation, and Systems*, vol. 18, no. 5, pp. 1343-1352, 2020.

[10] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705-724, 2015.

[11] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50K tries and 700 robot hours," *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3406-3413, IEEE, 2016.

[12] D. Park and S. Y. Chun, "Classification based grasp detection using spatial transformer network," *arXiv preprint arXiv:1803.01356*, 2018.

[13] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355-3362, 2018.

[14] Z. Wang, Z. Li, B. Wang, and H. Liu, "Robot grasp detection using multimodal deep convolutional neural networks," *Advances in Mechanical Engineering*, vol. 8, no. 9, p. 1687814016668077, 2016.

[15] J. J. van Vuuren, L. Tang, I. Al-Bahadly, and K. M. Arif, "A 3-stage machine learning-based novel object grasping methodology," *IEEE Access*, vol. 8, pp. 74216-74236, 2020.

[16] X. Wang, X. Jiang, J. Zhao, S. Wang, and Y.-H. Liu, "Grasping objects mixed with towels," *IEEE Access*, vol. 8, pp. 129338-129346, 2020.

[17] D.-W. Kim, H. Jo, and J.-B. Song, "Irregular depth tiles: Automatically generated data used for network-based robotic grasping in 2D dense clutter," *International Journal of Control, Automation, and Systems*, vol. 19, no. 10, pp. 3428-3434, 2021.

[18] K.-H. Ahn and J.-B. Song, "Image preprocessing-based generalization and transfer of learning for grasping in cluttered environments," *International Journal of Control, Automation, and Systems*, vol. 18, pp. 2306-2314, 2020.

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

[20] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, "A hybrid deep architecture for robotic grasp detection," *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1609-1614, IEEE, 2017.

[21] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 769-776, IEEE, 2017.

[22] Q. Zhang, D. Qu, F. Xu, and F. Zou, "Robust robot grasp detection in multimodal fusion," *Proc. of MATEC Web of Conferences*, vol. 139, p. 00060, EDP Sciences, 2017.

[23] D. Morrison, J. Leitner, and P. Corke, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *Proceedings of Robotics: Science and Systems*, pp. 1-10, 2018.

[24] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *The International Journal of Robotics Research*, vol. 39, no. 2-3, pp. 183-201, 2020.

[25] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3511-3516, IEEE, 2018.

[26] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—A survey," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289-309, 2013.

[27] C.-P. Tung and A. C. Kak, "Fast construction of force-closure grasps," *IEEE Transactions on Robotics and Automation*, vol. 12, no. 4, pp. 615-626, 1996.

[28] C. Rosales, R. Suárez, M. Gabiccini, and A. Bicchi, "On the synthesis of feasible and prehensile robotic grasps," *Proc. of IEEE International Conference on Robotics and Automation*, pp. 550-556, IEEE, 2012.

[29] D. Prattichizzo, M. Malvezzi, M. Gabiccini, and A. Bicchi, "On the manipulability ellipsoids of underactuated robotic hands with compliance," *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 337-346, 2012.

[30] J. Watson, J. Hughes, and F. Iida, "Real-world, real-time robotic grasping with convolutional neural networks," *Proc. of Annual Conference towards Autonomous Robotic Systems*, pp. 617-626, Springer, 2017.

[31] G. Chalvatzaki, N. Gkanatsios, P. Maragos, and J. Peters, "Orientation attentive robotic grasp synthesis with augmented grasp map representation," *arXiv preprint*, arXiv:2006.05123, 2020.

[32] Y. Xu, L. Wang, A. Yang, and L. Chen, "GraspCNN: Real-time grasp detection using a new oriented diameter circle representation," *IEEE Access*, vol. 7, pp. 159322-159331, 2019.

[33] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," *arXiv preprint arXiv:1909.04810*, 2019.

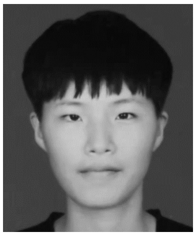
[34] S. Wang, X. Jiang, J. Zhao, X. Wang, W. Zhou, and Y. Liu, "Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images," *Proc. of IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 474-480, IEEE, 2019.

[35] P. Dolezel, D. Stursa, D. Kopecky, and J. Jecha, "Memory efficient grasping point detection of nontrivial objects," *IEEE Access*, vol. 9, pp. 82130-82145, 2021.

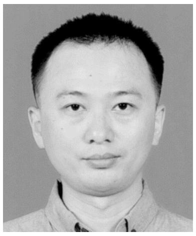
[36] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," *Proc. of IEEE International Conference on Robotics and Automation*, pp. 3304-3311, IEEE, 2011.

[37] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," *arXiv preprint arXiv:2103.02907*, 2021.

[38] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "ROI-based robotic grasp detection for object overlapping scenes," *Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4768-4775, IEEE, 2019.



Qian-Qian Hong received her B.S. degree in science and technology of computer from University of Electronic Science and Technology Zhongshan College, China in 2018. Currently, she is pursuing an M.S. degree in computer engineering with the Guangdong University of Technology. Her current research interests include computer vision and robotic grasping.



Liang Yang received his B.S. degree in electronics engineering from Nanchang University, Nanchang, China, in 2002, his M.S. and Ph.D. degrees from the School of Automation, Guangdong University of Technology, in 2005 and 2016, respectively. From 2005 to 2009, he has worked in Huawei Co. as a senior engineer. The products which he had ever involved in implementing serves millions of people. He is currently a Professor at the School of Computer Engineering, University of Electronic Science and Technology of China, Zhongshan Institute. Meanwhile, he is a postdoctoral with University of Electronic Science and Technology. His research interests include robot systems and technology, and robotics and computational intelligence.

Bi Zeng received her M.S. and Ph.D. degrees from the Guangdong University of Technology, where she is currently a Professor with the School of Computers. Her current research interests include computational intelligence, data mining, intelligent robot, and wireless sensor networks. She is a Senior Member of CCF, and Multi-Valued Logic and Fuzzy Logic Committee, China.

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