Robotica (2024), **42**, pp. 3190–3210
doi:10.1017/S0263574724001358

CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

A pixel-level grasp detection method based on Efficient Grasp Aware Network

Haonan Xi, Shaodong Li and Xi Liu

School of Electrical Engineering, Guangxi University, Nanning, China
Corresponding author: Shaodong Li; Email: lishaodongyx@126.com

Received: 14 October 2023; **Revised:** 11 May 2024; **Accepted:** 2 August 2024; **First published online:** 18 September 2024

Keywords: grasping; attention mechanism; grasp detection; robotic control; control of robotic systems

Abstract

This work proposes a novel grasp detection method, the Efficient Grasp Aware Network (EGA-Net), for robotic visual grasp detection. Our method obtains semantic information for grasping through feature extraction. It efficiently obtains feature channel weights related to grasping tasks through the constructed ECA-ResNet module, which can smooth the network's learning. Meanwhile, we use concatenation to obtain low-level features with rich spatial information. Our method inputs an RGB-D image and outputs the grasp poses and their quality score. The EGA-Net is trained and tested on the Cornell and Jacquard datasets, and we achieve 98.9% and 95.8% accuracy, respectively. The proposed method only takes 24 ms for real-time performance to process an RGB-D image. Moreover, our method achieved better results in the comparison experiment. In the real-world grasp experiments, we use a 6-degree of freedom (DOF) UR-5 robotic arm to demonstrate its robust grasping of unseen objects in various scenes. We also demonstrate that our model can successfully grasp different types of objects without any processing in advance. The experiment results validate our model's exceptional robustness and generalization.

1. Introduction

With the development of robotics and artificial intelligence, robot intelligent operation has excellent research potential. Among many critical operational capabilities, the ability to grasp is fundamental and essential, which will bring enormous power to the industry, home, service, warehouse, medical field, and so on refs. [1–5]. Moreover, effectively grasping transparent and specular objects remains challenging in robot grasp operations [6–8].

Nowadays, methods for grasping tasks can be summarized as analytic and data-driven methods. Though methods based on analytic methods can reliably grasp objects, these approaches usually require miscellaneous information about the object, such as the geometry of the object, a physical model, kinematics, dynamics, and mechanical analysis, and which are often not easily accessible [9–11].

In contrast, many modern data-driven methods for grasping objects rely on only visual information. It is very efficient for application because visual information is often easier and quicker to obtain. Meanwhile, with the improvement of computability, more and more researchers have developed deep learning methods for robotic manipulations [12–18]. Considering robots often inevitably need to grasp objects in unstructured scenes, many studies around deep learning methods based on visual information have become active [19, 20].

Most modern grasp detection [21–23] is based on convolutional neural networks, which have become the mainstream for grasp tasks based on visual information. Specifying a suitable grasp configuration is an essential issue in the image from the visual sensor. Reference [24] proposed a method that represented the grasp configuration with a rectangle. Besides, some researchers proposed that a series of grasp candidates be generated first, and then the optimal grasp be selected from them [25]. However, these

© The Author(s), 2024. Published by Cambridge University Press.

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

methods are relatively time-consuming and limited to real-time performance in generating candidates. Meanwhile, there is still room for improvement in the success rate of grasping.

Grasp tasks require both efficient real-time performance and accuracy. Reference [12] proposed a novel grasp detection method that is based on Unet [26]. The method can predict the pixel-level grasping pose and grasping quality based on the input image without generating candidates. Therefore, it is sufficiently fast for real-time work, and considering it incorporates the squeeze-and-excitation (SE) attention mechanism [27], the approach achieves impressive accuracy on public datasets. However, the SE attention mechanism still has certain shortcomings, such as when obtaining channel features, using dimensionality reduction to a low-dimensional space first and then mapping it back to the original dimension, which makes it impossible to establish a direct relationship between the channel and its weight, which can affect both accuracy and real-time performance.

In order to solve these problems and improve the performance of grasp detection, this paper presents a visual grasp detection framework. We adopt the residual block with the Efficient Channel Attention (ECA) mechanism incorporated. It overcomes the reaction of dimensionality reduction in the traditional SE attention mechanism and utilizes informative features to improve the network model's predicted accuracy by establishing a direct relationship between the channel and its weight. It can efficiently and accurately identify the graspable area of the object in the input image and output pixel-wise quality scores to obtain appropriate grasp configuration. We also evaluate our method on the Cornell dataset and the Jacquard dataset. The result shows our method achieves 98.9% and 95.8% accuracy, respectively, which is competitive with the state-of-the-art. Through using a 6-degree of freedom (DOF) UR-5 robotic arm for real-world grasp experiments, our method achieves a 93.6% grasp success rate in unstructured real-world scenes.

The contributions can be summarized as follows:

1) We have constructed the Efficient Channel Attention Residual Network (ECA-ResNet) module based on ref. [28] that directly establishes the relationship between channels and weights, which is beneficial for improving channel attention learning and reducing model complexity. The module enables the Efficient Grasp Aware Network (EGA-Net) to distinguish between objects and backgrounds and avoids the gradient vanishing problem caused by increasing the number of layers through residual layers, thereby improving learning efficiency.
2) The EGA-Net is proposed to generate pixel-level grasp poses and their quality score. The EGA-Net consists of two parts: feature extraction and feature fusion. Feature extraction is used to emphasize and obtain semantic information relevant to grasping features, while feature fusion is used to fuse low-level information containing rich spatial information. Meanwhile, it is only trained on the dataset that contains a single type of object; it still has a high success rate in grasping different types of objects.
3) Our method achieves competitive results versus state-of-the-art grasp methods on public datasets and real-world robotic grasp experiments.

2. Related work

2.1. Methods for grasp task

In the earlier period, most researchers used analytical methods to grasp tasks based on hand-engineering the features in specific tasks [29–31]. Meanwhile, though these methods possess high performance for grasping known objects, they cannot be extended to grasp unknown objects. When facing different scenes or types of objects, this method is often no longer applicable. With the rapid development of artificial intelligence, a new grasp detection method called the data-driven method has shown great potential in grasp tasks [13, 14]. These methods can solve the shortcomings of the analytic methods. It typically uses machine learning to develop models that map information read by sensors directly to ground truth from humans, physical, and simulation trials. These methods often require large-scale

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

datasets to support training. The most important conception of the method is to enable robots to imitate human grasp strategies [32]. Meanwhile, it typically requires close interaction between sensors and the environment.

2.2. Deep learning methods for grasp detection

In recent years, deep learning methods have achieved enormous advances in computer vision, natural language processing, and reinforce learning [18, 33, 34]. Inspired by these fields, many robotic researchers have begun to attempt to apply deep learning techniques to the field of grasp detection. Reference [35] proposed a visual enhancement guided grasp detection model (VERGNet), which includes a low-light feature enhancement strategy to improve the effectiveness of robot grasping under low-light conditions. Reference [36] further improved the performance of the Generative Residual Convolutional Neural Network (GR-ConvNet) by adding a dropout layer and replacing the rectified linear unit (ReLU) activation function with Mish as the activation function. Reference [37] proposes Selective Kernel convolution Grasp detection Network (SKGNet) based on Selective Kernel Convolution. It achieved excellent performance on single-type object datasets. However, a large number of model parameters compromised the real-time performance of the network. The current notable grasp detection methods can be divided into regression-based, classification-based, and detection-based methods.

The method by which the model directly predicts the grasp configuration by inputting relevant observation data of objects that can be grasped implicitly into the model is called the regression-based method. Reference [25] directly regresses the raw RGB-D image to obtain the grasp configuration instead of classifying many small patches in a sliding window type method and only needs to consider a single image and make a global prediction.

Rather than regression-based methods, ref. [38] plans the grasp representation orientation mapping as a classification task, which defines the learning problem to be classified with null hypothesis competition.

For detection-based methods, this type of method draws inspiration from the idea of semantic segmentation, transforming the prediction of grasp configuration into a prediction of the segmentation problem of optimal grasp area, such as refs. [13, 14].

In general, these detection-based methods can directly predict reliable grasp configurations and are highly effective. The above methods have achieved remarkable results in the field of grasp detection. However, they still need to consider the use of attention mechanisms to distinguish objects and backgrounds better, and there is still room for improvement in the accuracy of grasp detection. In addition, many detection methods do not consider grasping different types of objects, such as opaque, transparent, and specular objects. Thus, we propose a novel approach, EGA-Net, to address these issues by ECA-ResNet module based on ref. [28], which enables the network to reweigh channel weights to extract grasp-related information better. So, our method can distinguish between objects and backgrounds. Meanwhile, by using RGB-D multi-modal data as input to the EGA-Net, our method can effectively grasp different types of objects.

3. Problem statement

We describe the grasping task as a robot using the generated grasp configuration to perform grasping. Specifically, by inputting the RGB-D image of the grasping scene into our model, the model will output four heat maps representing the grasp configuration. We selected the grasp configuration with the highest grasp quality as the action that the robotic arm will perform, and we used a depth camera to obtain the RGB-D image of the grasping scene.

Most studies on grasp detection mainly use two grasp representations: grasp contact points and grasp rectangle. The grasp contact points representation method has the characteristics of being simple, direct, and easy to identify. It can perform dense grasp detection on the image in tasks. However, it cannot

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

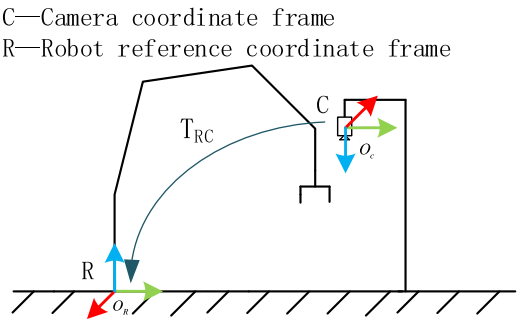


Figure 1. *The transformation between the camera coordinate frame and the robot reference coordinate frame.*

indicate the width of the opening or the orientation of the end-effector in detail [39]. Considering the characteristics of our method, we choose the grasp rectangle as the grasp detection representation and use an improved version of the grasp rectangle [14] to represent an appropriate grasp configuration and overcome the limitation of the grasp point representation method. We denote the grasp configuration in the image coordinate frame as:

$$G_I(q, p, \theta, \omega) \tag{1}$$

where q represents the grasp quality, $p(u, v)$ corresponds to the gripper's center position in image coordinate frame, the gripper's rotation angle around the z-axis is represented by θ, and ω represents the required gripper width.

Grasp quality takes a value of [0, 1]; the closer the score is to 1, the more likely the robotic arm is to grasp the object under the given grasp configuration successfully. To enable model learning smoothly, we calculate the gripper's rotation angle θ using $\sin(2\phi)$ and $\cos(2\phi)$ during post-processing by:

$$\theta = \frac{1}{2}\arctan\frac{\sin(2\phi)}{\cos(2\phi)}, \theta \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \tag{2}$$

Because the gripper is symmetrical around $\pm\frac{\pi}{2}$. Thus, θ is given in the range $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$. For ω, its pixel values are set in the range of $[0, \omega_{max}]$, and ω_{max} can be adjusted according to the maximum width of the gripper. The above grasp configuration is described in the image coordinate frame. We can convert the grasp configuration to the robot reference coordinate frame by:

$$G_R = T_{RC}(T_{CI}(G_I)) \tag{3}$$

where T_{CI} is the transformation of the image coordinate frame to the camera coordinate frame, and T_{RC} converts the camera coordinate frame to the robot reference coordinate frame by the hand-eye calibration result (see Figure 1).

4. Method

The grasp detection task we need to complete requires the use of a model to process the input image and classify the objects to be grasped in the image and the environment, similar to a segmentation task. In addition, we also need to extract semantic information about the object's properties, such as position and posture. Considering practical applications, we also need to simplify model complexity as much as possible to ensure its real-time performance.

Based on the above requirements, we consider selecting a segmentation network with outstanding segmentation and relatively lightweight performance as the backbone. We propose a novel model, a framework used initially for biomedical image segmentation [26]. To improve the accuracy of prediction, we introduce the ECA-ResNet module. Our robot visual grasp detection pipeline can be divided into

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

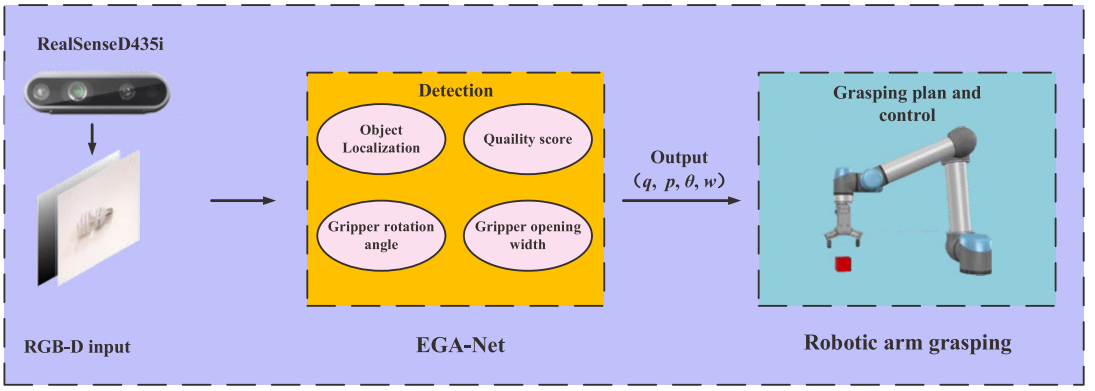


Figure 2. *The robot visual grasp detection pipeline.*

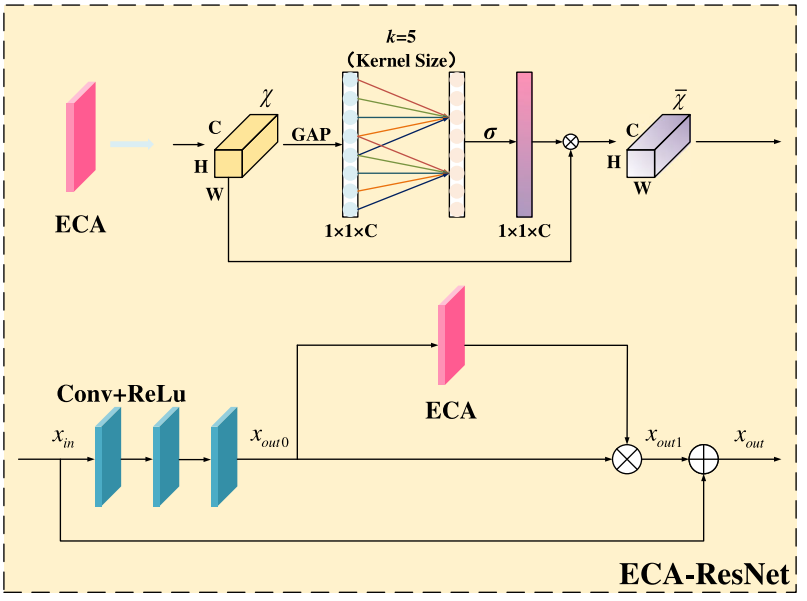


Figure 3. *ECA-ResNet module.*

three parts: first, we use a depth camera to obtain the required visual information, and then our model serves as the perception function. Finally, we send the appropriate grasp configuration to the robotic arm to complete the grasp action (see Figure 2).

4.1. Efficient Channel Attention Residual Network (ECA-ResNet)

In order to improve the prediction accuracy of the model, we need to make the network ignore environmental information unrelated to the object as much as possible. Introducing attention mechanisms has been proven effective by most works [12, 15, 27, 28, 40]. Therefore, we adopt a channel attention mechanism to improve the network's ability to obtain global features, reduce attention to noise and trivial information, and extract feature channels related to grasping tasks.

We proposed the ECA-ResNet module into our model; the module is given in Figure 3. It consisted of three convolution layers at first. Then, the global average pooling (GAP) and ECA module, where the reduction ratio r is set to 16, was used. Specifically, assuming the input is X, the channel weight

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

calculated by the ECA can be represented by Equation 4:

$$\mathcal{W}_{eca} = \sigma(\mathbb{W}_{poe}(f(X))) \tag{4}$$

where $f(X) = \frac{1}{HW}\sum_{i=1,j=1}^{H,W} X_{ij}$ is channel-wise GAP, σ is a Sigmoid function, and $\mathbb{W}_{poe}$ has

$$\begin{bmatrix} \omega^1 & \cdots & \omega^k & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \omega^1 & \cdots & \omega^k & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \omega^1 & \cdots & \omega^k \end{bmatrix} \tag{5}$$

As for Equation 5, such method can be efficiently implemented by a fast 1D convolution with kernel size of k:

$$\sigma(\mathbb{W}_{poe}(f(X))) = \sigma(C1D_k(f(X))) \tag{6}$$

After channel-wise GAP without dimensionality reduction, the ECA module obtains local cross-channel interaction by considering the impact of k neighbors [28]. Replacing the two full connections in the traditional SE block with $1D$ convolution can directly establish the relationship between channels and their weights. In the SE block, the channel weight calculated can be represented by Equation 7:

$$\mathcal{W}_{se} = \sigma(W_2 ReLu(W_1(f(X)))) \tag{7}$$

where there are a total of $2 \times (\frac{C^2}{r})$ parameters, W_1 and W_2 contain parameters $C \times (\frac{C}{r})$ and $(\frac{C}{r}) \times C$, respectively. The fact that the ECA module only contains k parameters, which is usually less than $2 \times (\frac{C^2}{r})$. So, we simplified the complexity of the model. We improve the model's prediction accuracy and real-time performance based on the above reasons. Meanwhile, we can improve the effect of our model by increasing the number of layers. However, excessive layers may prevent smooth learning; they induce problems such as vanishing gradients, overfitting, and dimensionality errors, which harm performance [41]. So, we use the residual layer to solve the problems above. As shown in the bottom frame of Figure 3, assuming the input is x_{in}, the output x_{out} of the ECA-ResNet module can be represented by Equation 8.

$$\begin{cases} x_{out0} = \mathcal{W}_{Conv}(x_{in}), \\ x_{out1} = \sigma(\mathbb{W}_{poe}(f(x_{out0})))x_{out0}, \\ x_{out} = x_{in} + x_{out1} \end{cases} \tag{8}$$

where $\mathcal{W}_{Conv_Relu}$ indicates three convolution operations, with a ReLU after each convolution operation.

4.2. Network architecture

We present the architecture of the EGA-Net in Figure 4, which is inspired by ref. [26]. To fully utilize visual information, we use the 4-channel multi-modal RGB-D image as input to the model [6]. The model consists of two parts: feature extraction and feature fusion. The feature extraction obtains high-level information in the image through convolution and pooling operations and emphasizes semantic information relevant to grasping features using the constructed ECA-ResNet module. Afterward, to ensure that low-level information containing rich spatial information is not lost, we concatenate low-level features and ultimately output a heatmap that includes grasp quality, grasp angle, and grasp width in the feature fusion section. The process can be described as follows: First, features of the 4-channel image are extracted by a convolutional layer to get a sizeable receptive field; the output size is $224 \times 224 \times 32$. Then, the output as input forwards Maxpooling ($kernelsize = 2, stride = 2$) and Doubledownsample layers ($kernelsize = 3, padding = 1$), and the size of the output is $112 \times 112 \times 64$, defined as x_1. We will pass this output through Maxpooling and Doubledownsample again; the output size is $56 \times 56 \times 128$.

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

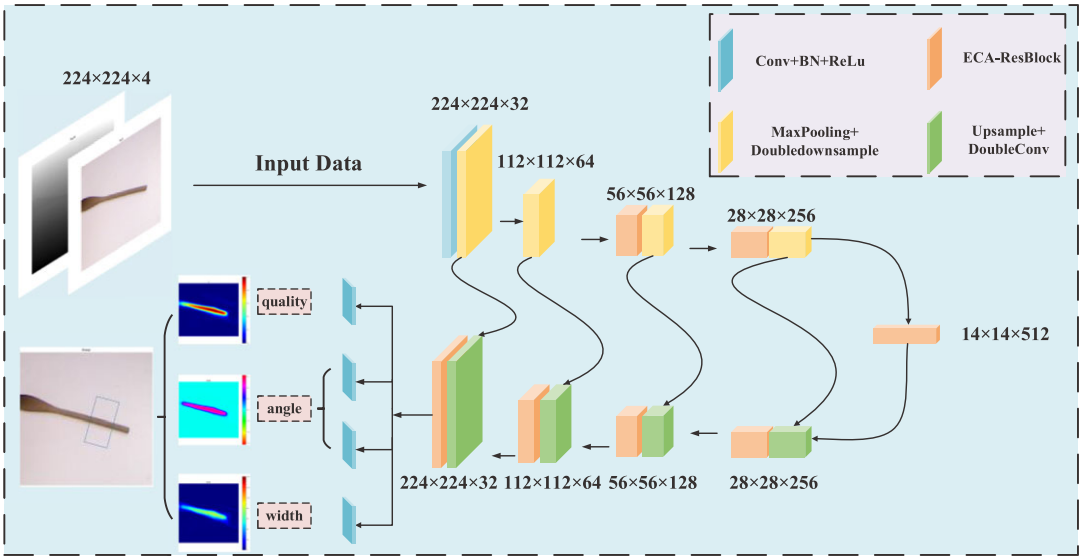


Figure 4. *Architecture of the Efficient Grasp Aware Network.*

After that, the efficient channel attention and residual layer are used to suppress the corresponding trivial background area, which smooths learning, extracts the saliency features, and improves the predicted accuracy of the network. MaxPooling and Doubledownsample layers were followed after this ECA module, and the output size is $28 \times 28 \times 256$, defined as x_2. The output as input forwards the ECA module and Maxpooling and Doubledownsample layers again; the output size is $14 \times 14 \times 512$, defined as x_3. Then, x_3 is fed into the ECA module, and the output size is also $14 \times 14 \times 512$, defined as x_4. We feed x_3 and x_4 into the Upsample layer, concatenate them, and then pass them to the double convolutional layer. The output size is $28 \times 28 \times 256$. Meanwhile, we use the batch normalization and the ReLU function in all convolution layers.

Our EGA-Net model processes the 4-channel image in a single pass to predict an appropriate grasp configuration, which includes four components: grasp quality, $\sin(2\phi)$, $\cos(2\phi)$, and the opening width of the gripper. We choose the pixel $p = (u, v)$ with the highest grasp quality in the grasp quality heat map as the grasp position and use the grasp angle and grasp width represented by the same pixel position in the grasp angle heat map and grasp width heat map as the grasp configuration.

A grasp rectangle can represent the necessary parameters mentioned above. Specifically, in the image coordinate frame, we use the position of the pixel $p = (u, v)$ with the highest grasp quality in the grasp quality map output from the EGA-Net as the vertical projection of the center of the gripper's fingertips, the angle θ of the rectangle as the rotation angle of the gripper around the z-axis, and the length ω of the rectangle as the gripping width, as shown in the Figure 5.

4.3. Evaluation metric

To fairly evaluate the performance of our model, we use the rectangle metric for our method. According to ref. [42], it can be considered successful when a grasp configuration meets the following conditions:

- The intersection over union (IOU) score between the predicted grasp G_p and the ground truth grasp G_t is higher than 25%
- The difference between the grasp orientation of the predicted grasp rectangle and the grasp orientation of the ground truth rectangle is less than 30°.

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

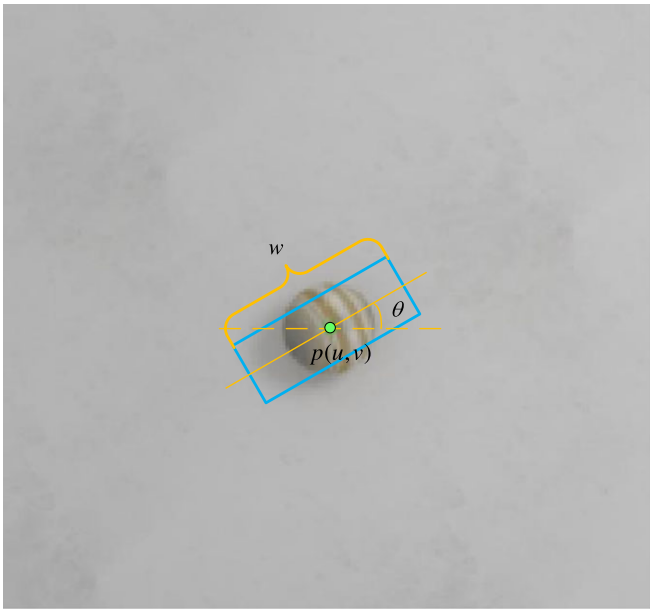


Figure 5. *The grasp rectangle represents grasp configuration.*

Where the IOU score for the predicted grasp G_p and the ground truth grasp G_t is defined as:

$$\text{IOU}(G_p, G_t) = \frac{G_p \cap G_t}{G_p \cup G_t} \tag{9}$$

where G_p is the predicted grasp rectangle and G_t is the ground truth grasp rectangle. The IOU score is the quotient of the area of intersection between two rectangles and the area of union.

4.4. Loss function

In order to train our model well, we need to choose a stable and efficient loss function. Thus, we calculate all losses in the output of the EGA-Net and choose the sum of losses to optimize our model. The entire function is defined as:

$$Loss(G_p, G_t) = \frac{1}{n} \sum_i l_i \tag{10}$$

where n represents the number of samples, $l_i = \alpha l_{pos} + \beta(l_{cos} + l_{sin}) + \gamma l_\omega$, $l_{pos}, l_{cos}, l_{sin}$, and l_ω are the smooth $L1$ losses of all outputs of the model, respectively. Where *smoothL*1 loss is defined as:

$$smoothL1 = \begin{cases} 0.5(G_t - G_p)^2, & if |G_t - G_p| < 1 \\ |G_t - G_p| - 0.5, & otherwise \end{cases} \tag{11}$$

A suitable grasp configuration is crucial for successfully grasping an object. Thus, the position and rotation angle of the gripper are very important. Besides, an appropriate gripper opening width can avoid unnecessary collisions. Thus, we experimentally set $\alpha = 1$, $\beta = 10$, and $\gamma = 5$.

4.5. Training details

All networks involved in this paper are implemented with PyTorch. All networks are trained end-to-end with a single NVIDIA TITAN V GPU, which has a total of 192 GB of memory. The batch size is set

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

to 8, and the optimizer used in this paper is Adam. The learning rate is set to 0.0001. We take a center crop of 224×224 pixels for input data.

5. Experiment setup

To verify the performance of the proposed method, we choose to test it on the Cornell dataset and the Jacquard dataset. The Cornell dataset is widely used in the field of grasp detection, but it contains comparatively less data, in fact. Some researchers have proposed methods to improve the performance of models trained on limited datasets and have demonstrated through experiments that these methods are effective [43, 44]. Thus, we also chose to test the performance of the method using the large-scale Jacquard dataset. Meanwhile, we conducted ablation studies to test the contribution of each component in the network, demonstrating the high performance of our method through comparison experiments. In addition, we implemented the method on a real robotic arm to verify its feasibility.

5.1. Cornell dataset

The Cornell dataset is the first dataset to use the grasp rectangle as a grasp representation and is widely used. The Cornell dataset contains a total of 224 objects with 885 RGB-D images. Each RGB-D image has a resolution of 640 × 480 pixels. It has approximately 5110 positive and 2909 negative grasps. Each image is labeled with multiple grasp poses, which is helpful for our model to output a pixel-level grasp configuration. The quality of the dataset will directly affect the performance of the trained model. We improve the performance of the model by implementing an augmented Cornell dataset using random zooms, crops, and rotations.

We divided the dataset into two parts, where 90% of the dataset is used as the training set, and 10% is the testing set. In order to demonstrate the performance of our model fairly, we divide the Cornell dataset into two sets: image-wise (IW) and object-wise (OW).

5.2. Jacquard dataset

Compared to the Cornell dataset, the Jacquard dataset has a wider variety of object types and a larger scale of labeled data, which is beneficial for improving the model's generalization ability. The Jacquard dataset contains 11,000 objects with 54,000 images, which collects labeled data through self-supervised simulation of robotic arms. Because the dataset is enough for the training process, data augmentation will not be performed in this dataset. Similarly, we use 90% of the dataset as the training set and the other 10% as the testing set.

5.3. Grasp experiment

We use a 6-DOF UR-5 robotic arm with a Robotiq 2F-140 gripper for robotic grasp-and-place. We adopt an Intel RealSense D435i RGB-D camera that uses stereo vision to calculate depth and an infrared projector that is fixed 0.8 m above the workspace (see Figure 6) to obtain information about the objects. The EGA-Net is performed on a PC running Windows 10 with a 2.6 GHz Intel Core i7-10750 CPU, NVIDIA GTX 1650Ti graphics card, and 16 GB of RAM allocated per job. To verify the generalization ability of our model, we selected a series of objects with different shapes and types for the grasp experiment in the physical world, all of which the model had never seen before.

We perform grasp tests of the isolated scene, the cluttered scene, and the non-clutter scene. All objects used in the experiment are shown in Figure 7. It is worth noting that even when facing challenging types (transparent and specular) of objects, our model can still achieve a great success rate in a variety of scenes.

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

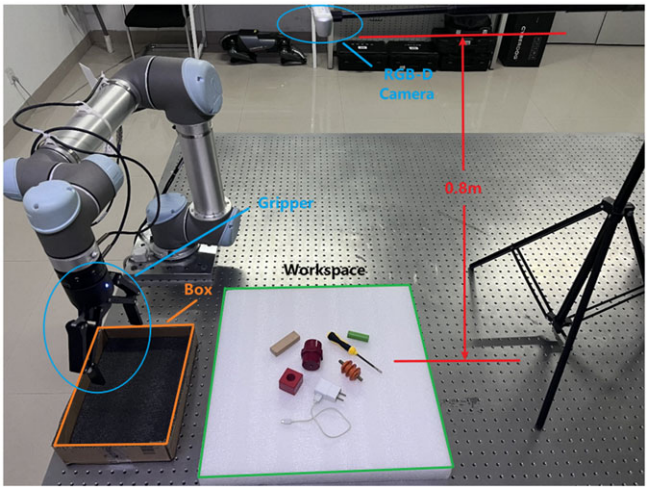


Figure 6. *Grasp experiment.*

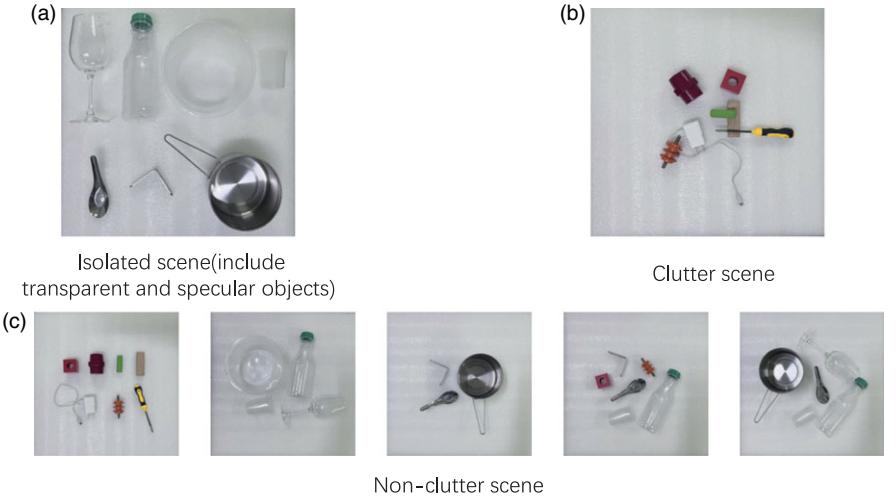


Figure 7. *Objects of a variety of scenes.*

6. Experiments

6.1. Grasp detection on cornell dataset

Because the EGA-Net combines the color space containing rich texture information with depth data containing object shape information, the method can predict the angle and width of the object to be grasped well. In addition, through the ECA-Resnet module, the EGA-Net can distinguish objects from the background and segment suitable grasp areas.

To show the effectiveness of our method, the EGA-Net is compared with recent methods in Table I. The EGA-Net outperforms these methods and achieves an accuracy of 98.9%, which is competitive with the state-of-the-art. It is worth noting that EGA-Net can accurately locate the object to be grasped from the environment through the application of attention mechanisms; meanwhile, a dropout layer is added to each of the output layers of the network to improve generalization ability. Data augmentation is used to enrich the dataset during training, making the network achieve better results under OW. Due to the lighter weight of the ECA-ResNet modules we use, the average inference speed of the model can reach

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

Table I. *Results on Cornell dataset.*

Authors	Algorithm	IW(%)	OW(%)	Speed(ms)
Jiang [42]	Fast Search	60.5	58.3	5000
Lenz [24]	SAE, struct. reg	73.9	75.6	1350
Morrison [13]	GG-CNN	73.0	69.0	19
Asif [45]	GeaspNet	90.2	90.6	24
Kumra [14]	GR-ConvNet-RGB-D	97.7	96.6	20
Wang [15]	TF-grasp	97.9	96.7	41.6
i Yu [12]	SE-ResUnet	98.2	97.1	25
Our	**EGA-Net**	97.8	**98.9**	24

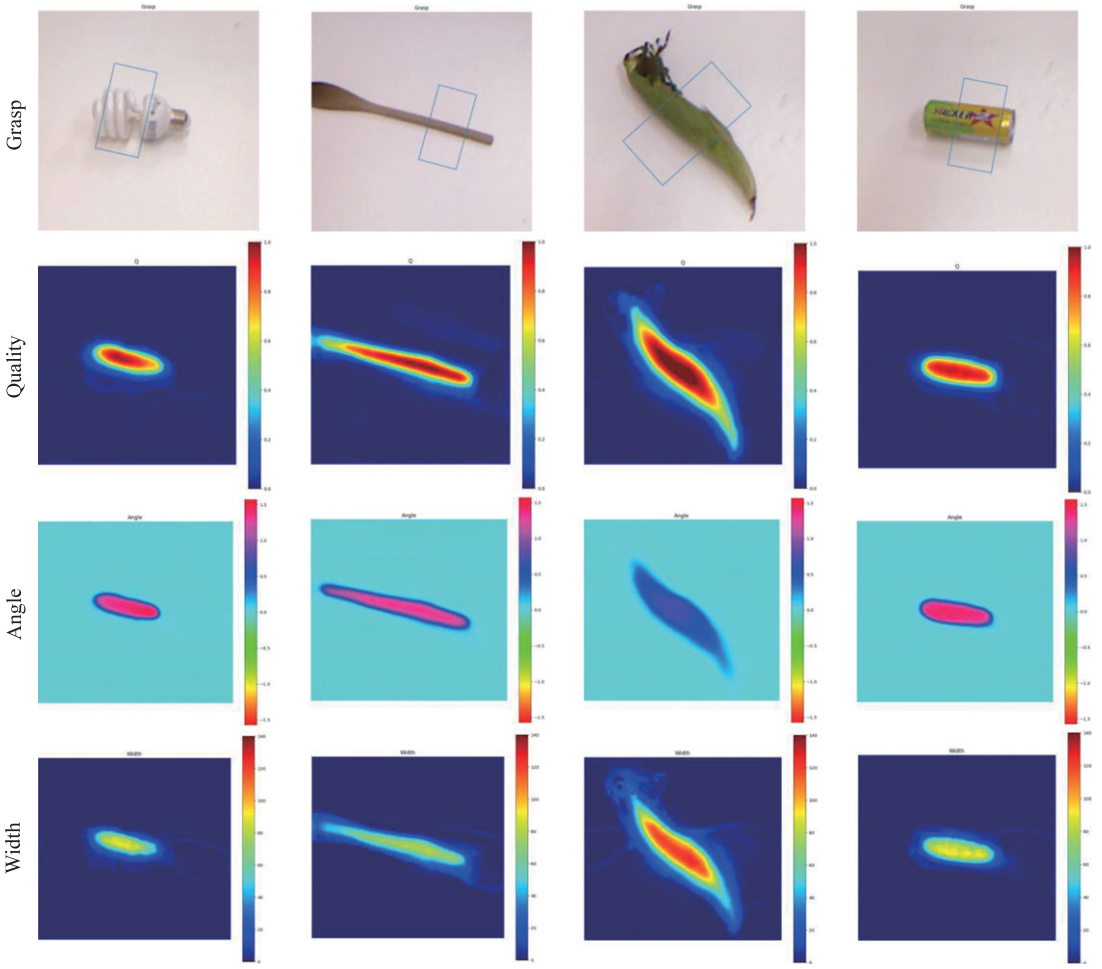


Figure 8. *The part grasp detection results on the Cornell dataset.*

24 ms per image, with high real-time performance. We also present the part prediction results of EGA-Net on the Cornell dataset in Figure 8. We chose the grasp configuration with the highest grasp quality for the robotic arm to execute, which is shown in the top row of the Figure 8. The other rows are heat maps of grasp quality, angle, and width, respectively. It shows our method achieved impressive results on the Cornell dataset.

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

Table II. *Results on Jacquard dataset.*

Authors	Algorithm	Accuracy(%)
Morrison [13]	GG-CNN	84
Kumra [14]	GR-ConvNet-RGB-D	94.6
Wang [15]	TF-Grasp	94.6
Yu [12]	SE-ResUnet	95.7
Our	**EGA-Net**	**95.8**

6.2. Grasp detection on jacquard dataset

The performance on the Jacquard dataset is shown in Table II, and our proposed method achieved high accuracy, achieving a prediction accuracy of 95.8%. We have also provided heat maps of prediction results on the part of the Jacquard datasets, as shown in Figure 9. From the Figure 9, we can qualitatively observe that our method can predict secure and robust grasp configuration, which contributes to increased prediction accuracy for grasp detection tasks.

6.3. Ablation studies

In order to assess the influence of the incorporated components on the predictive performance of the model, we conducted ablation experiments. The accuracy of the model trained on the Cornell dataset was measured using the same training methodology to evaluate the contributions of each component, as presented in Table III. Our findings from these experiments reveal that incorporating an attention mechanism is beneficial for improving network performance. Specifically, we found that the ECA mechanism outperforms other attention mechanisms in terms of effectiveness. Additionally, when combined with residual networks, it further enhances overall model performance.

6.4. Comparison studies

We conducted comparison studies and used the same input data and training settings to compare the performance of our method with other outstanding methods like GG-CNN [13], GR-ConvNet [14], and SE-ResUnet [12] on the Cornell dataset, Jacquard dataset, and real-world scenes. In the comparison studies, for fairness, we used same training settings and the same input data for each dataset and only changed the network for the experiment.

For the Cornell dataset, as shown in Figure 10, our method achieves higher grasp quality in terms of the object's graspable region. Additionally, for a given grasp point, our method determines a more suitable grasp angle and width. This advantage enhances the safety and reliability of robotic arm grasp actions in real-world scenes.

For the Jacquard dataset, as shown in Figure 11, our method can accurately distinguish between the background and object of each scene. GG-CNN's grasp angle of prediction and SE-ResUnet's grasp angle of prediction are incorrect, which can cause collisions between the gripper and the object and prevent successful grasp. Although GR-Convnet predicts a feasible grasp configuration, our method predicts a more suitable grasp angle and a more reliable grasp quality heat map. Overall, compared to other advanced methods, our method is able to generate more reliable grasp configuration under the same input data.

Furthermore, we conducted a comparative analysis of the performance of various methods in real-world scenes. The objects involved were unseen, and some of them possessed unique physical properties. Firstly, our evaluation focused on predicting the grasp configuration of individual objects, as illustrated in Figure 12. For the left-side specular object, all methods effectively predicted the grasp configuration. However, GG-CNN's grasp width of prediction was suboptimal, while GR-Convnet's angle of

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

Table III. *The experimental results of different component networks.*

	Accuracy(%)	
Network	**IW**	**OW**
Unet	91.0	92.1
SE-Unet	92.1	93.3
EGA-Net (without residual layer)	93.3	94.4
EGA-Net	97.8	98.9

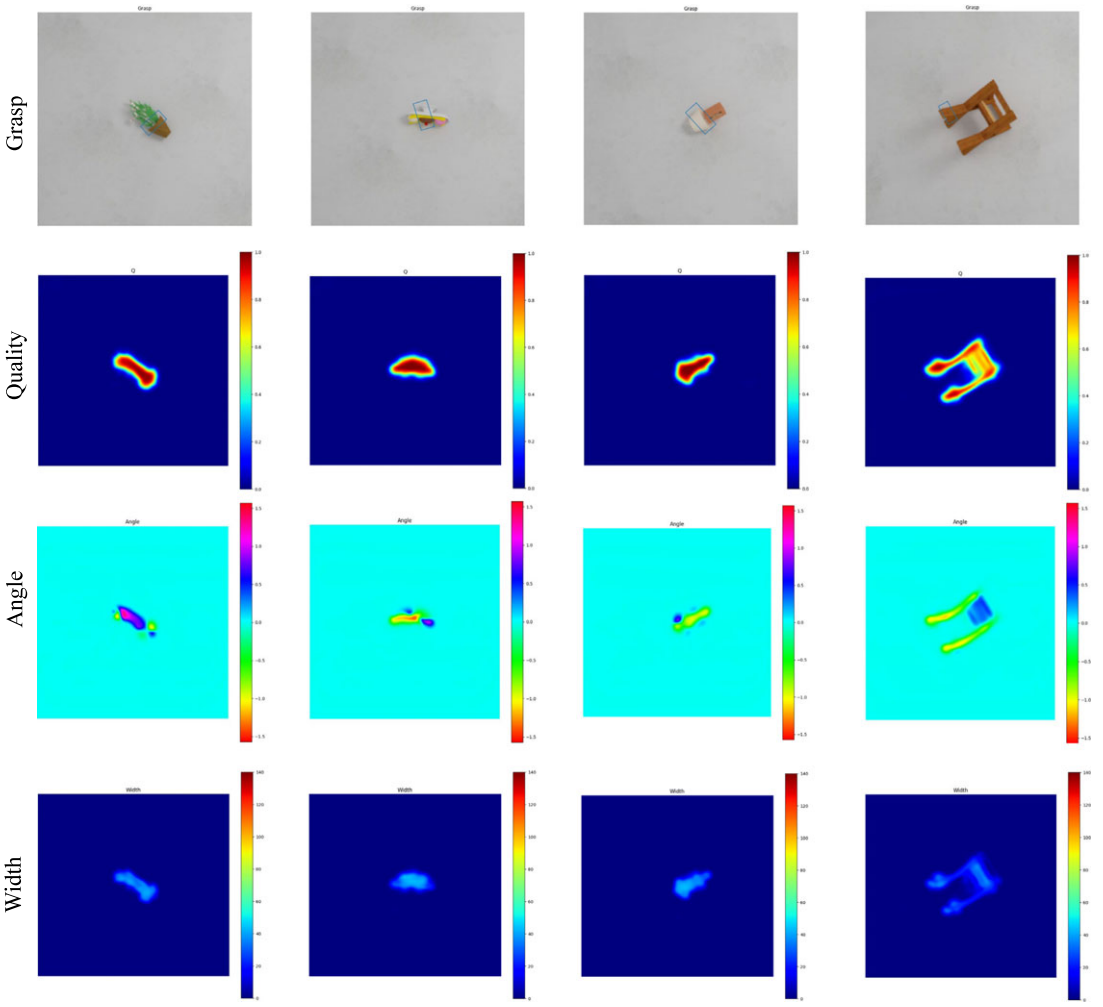


Figure 9. *The part grasp detection results on the Jacquard dataset.*

prediction risked collisions between the gripper and the object. Although SE-ResUnet generates a suitable grasp position and angle, its grasp width requires refinement. In contrast, our method produced a more appropriate grasp configuration, which not only reliably grasps objects but also has a more suitable grasp width. For transparent objects on the right side, because transparent objects have unique physical properties; even humans can sometimes find it difficult to distinguish them from their environment. GR-Convnet and SE-ResUnet view objects and the environment as one and do not detect the existence of

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

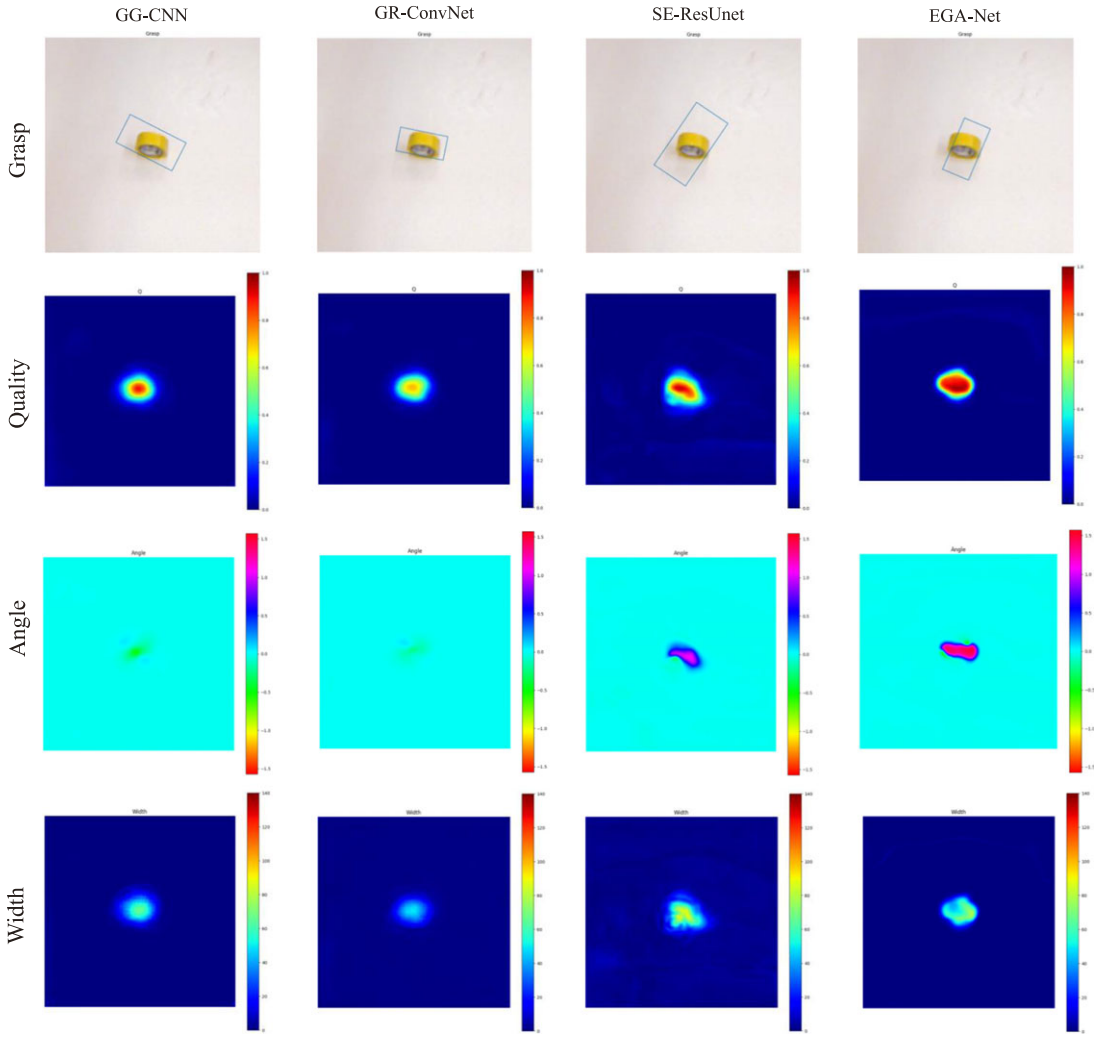


Figure 10. *Comparison of grasp detection results on the Cornell dataset.*

objects. GG-CNN has detected the object, but the angle and width could be better. Compared to opaque objects, our model has shown a decrease in the prediction of graspable areas but still predicts a suitable grasp pose. Secondly, while our models are trained only on data sets with individual objects, we verify the generalization of our method by comparing the results of the prediction of multiple objects, as shown in Figure 13. For all opaque objects on the left side of Figure 13, GG-CNN's predicted grasp poses will result in collisions, and the GR-Convnet and SE-ResUnet predicted grasp poses can grasp objects, but overall, they are not as accurate as the predictions of the EGA-Net. For objects containing various different physical properties in the scene on the right side of the Figure 13, only the EGA-Net accurately predicted the appropriate grasp position, and other methods ignored the presence of transparent objects due to the inability to distinguish between the transparent objects and the environment.

We also conducted comparison experiments on the most difficult-to-predict transparent objects and specular objects separately, as shown in Figure 14. Due to the introduction of multiple objects in the scene, there is less environmental information unrelated to grasping, which reduces the difficulty of detecting object positions and enables most algorithms to generate detection results. The results show

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

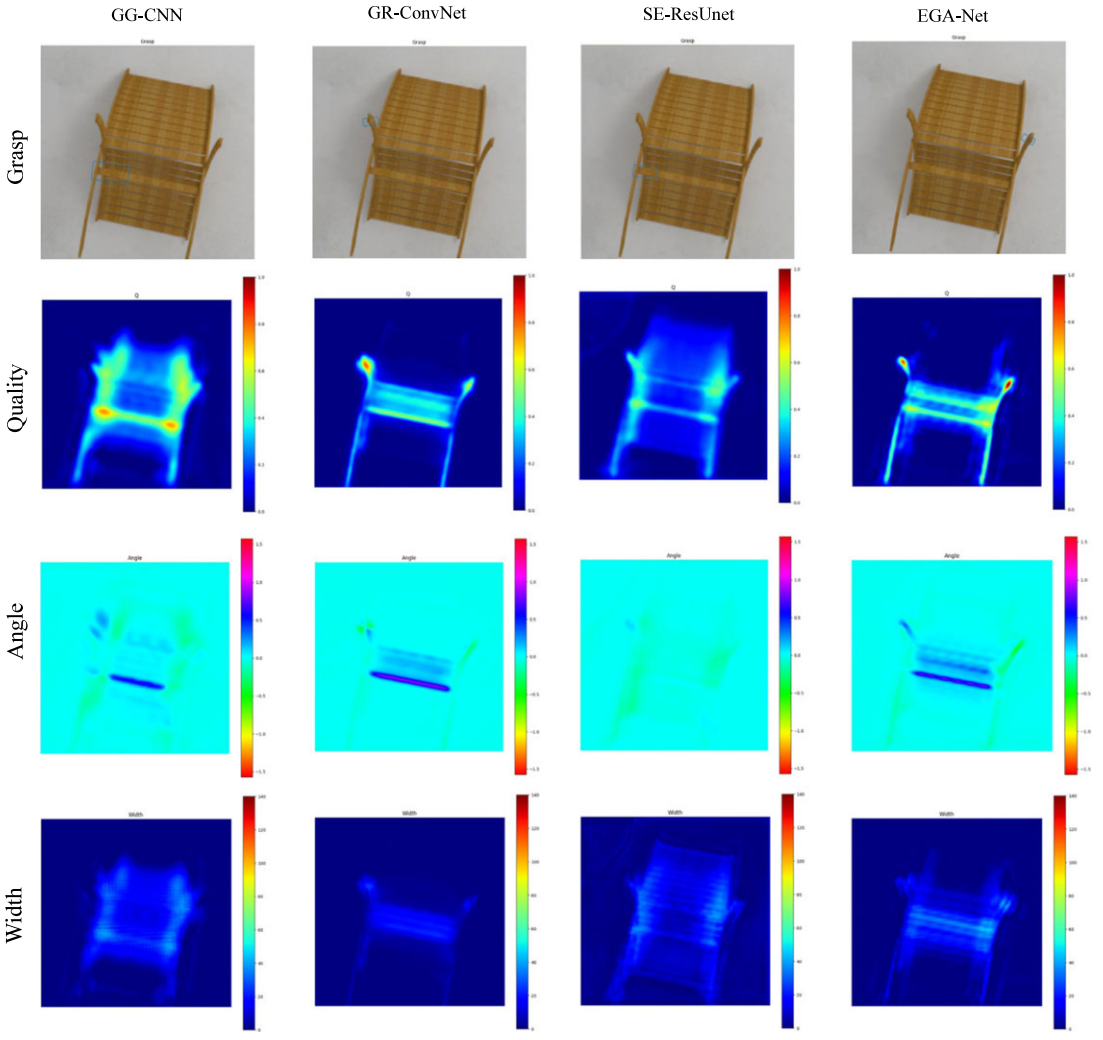


Figure 11. *Comparison of grasp detection results on the Jacquard dataset.*

that EGA-Net can predict the appropriate grasp configuration for different types of objects with higher robustness and prediction accuracy, which is superior to the current multiple methods in the real world.

6.5. The robot visual grasp detection pipeline

The 4-channel RGB-D image from the RealSense D435i camera is input to the EGA-Net, which predicts the appropriate grasp configuration for our robot and executes the grasp action. When an object is picked up by the gripper and placed in the box, we think it is a successful grasp; otherwise, it is considered a failed one.

6.5.1. Objects in isolated scene(Including trasparent and specular objects)

As far as we know, there are methods for grasping transparent objects [14, 46]. However, the position grasped by these methods is in an opaque area of the object. The performance of these methods could be better when facing completely transparent objects.

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

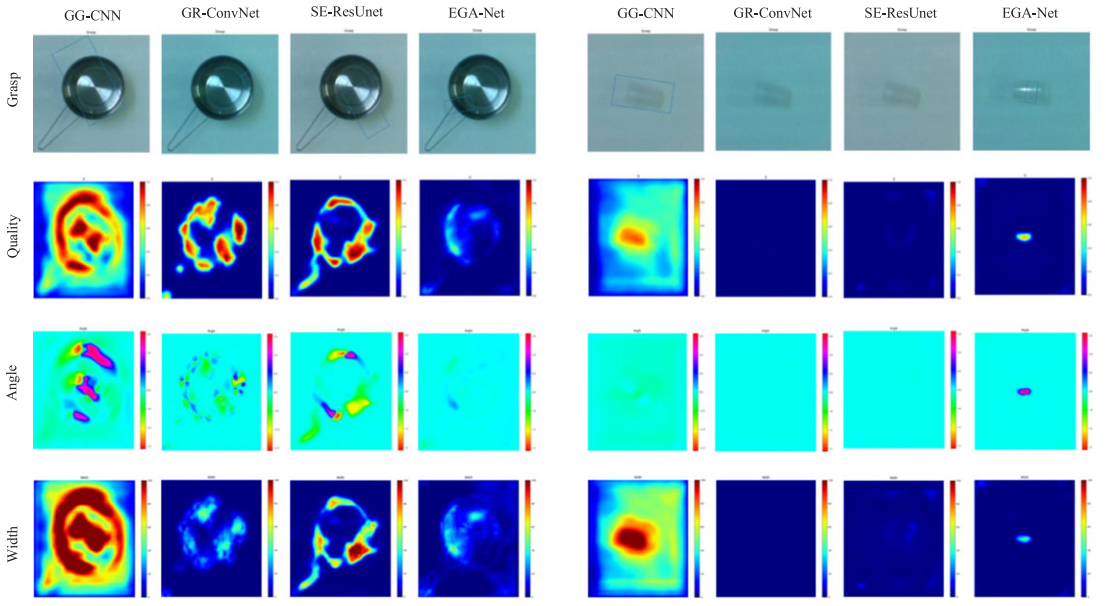


Figure 12. *Comparison studies with individual objects.*

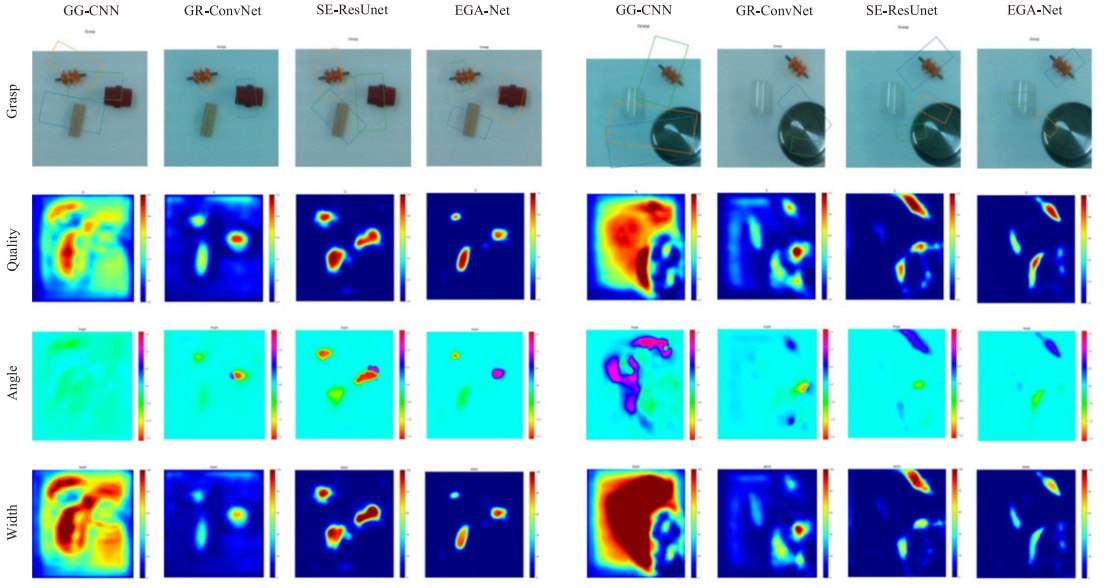


Figure 13. *Comparison studies with multiple objects.*

Although transparent and specular objects possess unique visual properties, our method adopts a multi-modal fusion approach and takes the RGB-D image as the input. Thus, it can still predict appropriate grasp configuration (see Figure 15). The robotic arm performed 93 successful grasps of the total 100 grasp attempts, achieving a success rate of 93% for transparent and specular objects. The failure to grasp is due to the network's inability to detect transparent objects and generate appropriate grasp configuration.

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

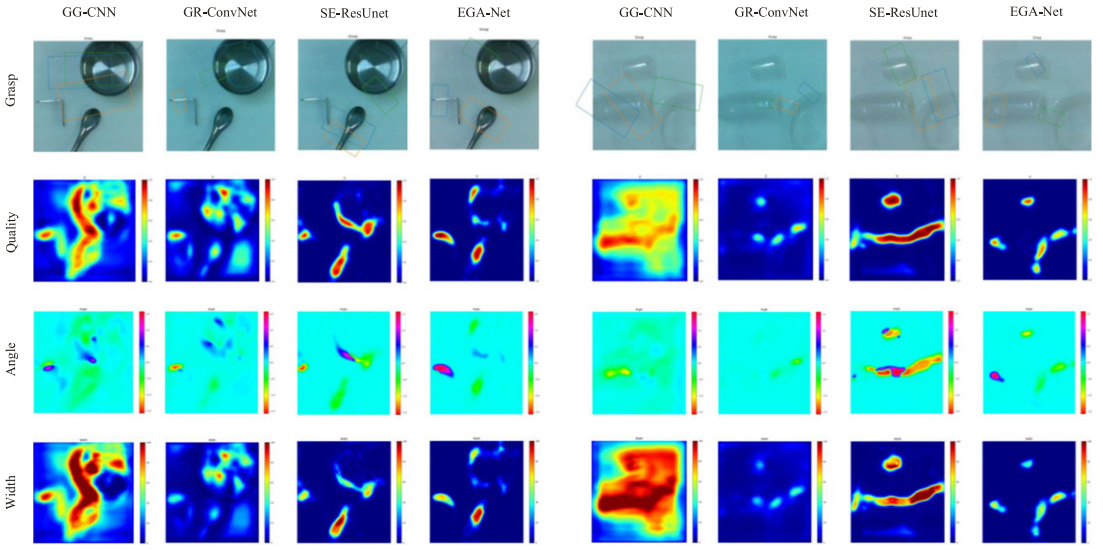


Figure 14. *Comparison studies of specular objects and transparent objects.*

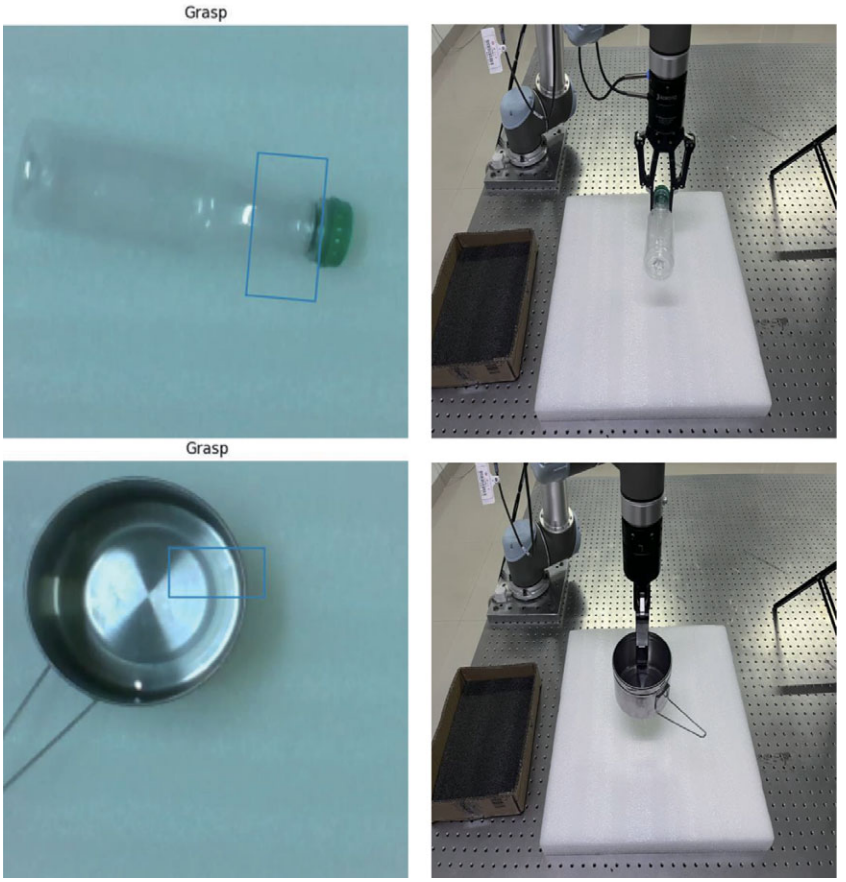


Figure 15. *Grasp for transparent and specular objects.*

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

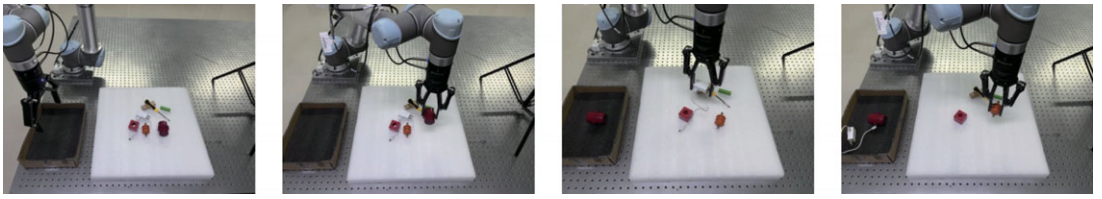


Figure 16. *Physical grasp in clutter.*

6.5.2. Objects in clutter scene

We conduct grasp experiments in the clutter scene. We stack objects from non-clutter experiments to build the clutter scene. In this scene, multiple objects are placed in the workspace, and there is occlusion between objects. This makes it more difficult for the model to predict the appropriate grasp configuration. Under the circumstances, we also achieved a success rate of 91% in a total of 100 grasp attempts. The experiment scene is shown in Figure 16. The most common gripping failure is caused by collisions between the gripper and surrounding objects.

6.5.3. Objects in non-clutter scene

To verify the generalization ability of our model, we conducted grasp experiments with different types of objects (see Figure 17). In 20 runs of the experiment (a) (only including opaque objects), we performed 140 attempts, of which 131 were successful (93.6%). Table III shows the results of data-driven methods for the real robotic grasp experiment. We performed the same experimental setup, only replacing objects to be grasped. Through comparison, our method outperforms current, same-type methods in real-world experiments. Experiments have shown that our model has a high generalization ability. Even in different environments (b), (c), (d), and (e) in Figure 17, it still achieves an average success rate of 88.6% (121/140) for grasping different types of objects. We encountered the most common failure cause: collision of the gripper with other objects. This may be due to the poor quality of the ground truth about width in the original dataset.

6.5.4. Error analysis

The output of the grasp detection method has the most critical impact on the successful grasp of objects. However, some external factors still have an impact on grasp performance.

- The input data used in this paper is the RGB-D image obtained by the Intel RealSense D435i RGB-D camera. Therefore, adopting an industrial camera with higher performance to improve the quality of the acquired image, especially the quality of the depth data will enhance the performance of grasp detection.
- Errors from hand-eye calibration results and the motion of the 6-DOF UR-5 robotic arm can also affect the grasp's performance.
- When facing transparent and specular objects, the unique physical properties of these objects would distort the light path. Therefore, illumination would also affect the performance of grasping these objects.

7. Conclusion

In this paper, we constructed the ECA-ResNet module to establish the direct relationship between channels and their weights directly. We help the model better extract semantic information about grasp

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

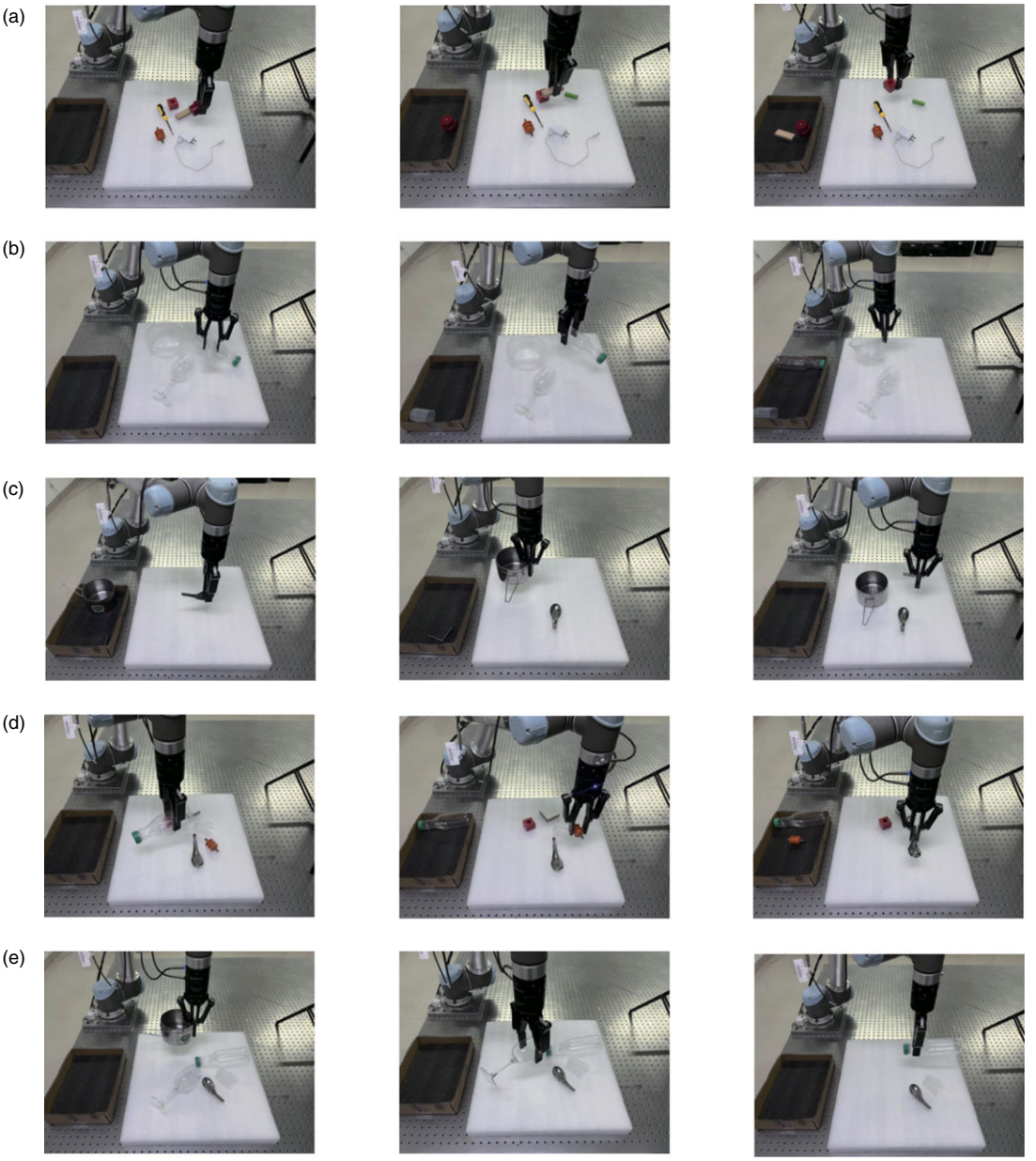


Figure 17. *Physical grasp in non-clutter. (a) only including opaque objects. (b) only including transparent objects. (c) only including specular objects. (d) Including opaque,transparent and specular objects. (e) including transparent and specular objects.*

and propose a novel model, EGA-Net, for grasp detection. Our method takes multi-modal information (RGB-D image) as input to utilize visual information fully. It can efficiently distinguish important features, reduce attention to trivial features, and output a pixel-level quality score to obtain appropriate grasp configuration. The experiments show that our proposed method outperforms previous methods and achieves 98.9% and 95.8% on the Cornell and Jacquard datasets, respectively. Moreover, our model does not require any specific processing for input; it can also grasp transparent and specular objects

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

with a high success rate, even in different environments. Nevertheless, in the grasp detection of transparent and specular objects, our method still encounters situations where transparent objects cannot be detected, possibly due to missing depth values. In the future, we plan to use the depth prediction method to solve the problem of missing depth values. We aim to improve the performance of robots in grasping transparent objects, enabling robots to replace humans in grasping dangerous transparent objects, such as post-assay test tubes, which will reduce the risk of infection for relevant practitioners. In addition, it is valuable to develop multi-modal fusion for grasping, such as visual information (RGB and depth), temperature, and audio modalities.

Author contribution. Haonan Xi: Conceptualization, design of methodology, designing computer programs, specifically performing the experiments, and video demonstration. Shaodong Li: Supervision, provision of study materials, computing resources and instrumentation, and oversight and leadership responsibility for the research activity planning and execution, specifically critical review on previous versions of the manuscript. Xi Liu:Supervision. All authors read and approved the final manuscript.

Financial support. This work was supported in part by the Natural Science Foundation of Guangxi (Grant No. 2023GXNSFBA026069), and in part by the funding of basic ability promotion project for young and middle-aged teachers in Guangxi's colleges and universities (Grant No. 2022KY0008).

Competing interests. The authors declare no competing interests exist.

Ethical approval. None.

References

[1] H. Tian, T. Wang, Y. Liu, X. Qiao and Y. Li, "Computer vision technology in agricultural automation-a review," *Inf Process Agric* **7**(1), 1–19 (2020).
[2] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou and Y. Mezouar, "Robotic manipulation and sensing of deformable objects in domestic and industrial applications: A survey," *Int J Robot Res* **37**(7), 688–716 (2018).
[3] Y. Hu, X. Wu, P. Geng and Z. Li, "Evolution strategies learning with variable impedance control for grasping under uncertainty," *IEEE Trans Ind Electron* **66**(10), 7788–7799 (2018).
[4] R. K. Hota and C. S. Kumar, "Effect of design parameters on strong and immobilizing grasps with an underactuated robotic hand," *Robotica* **40**(11), 3769–3785 (2022).
[5] W. Yan, Z. Deng, J. Chen, H. Nie and J. Zhang, "Precision grasp planning for multi-finger hand to grasp unknown objects," *Robotica* **37**(8), 1415–1437 (2019).
[6] T. Weng, A. Pallankize, Y. Tang, O. Kroemer and D. Held, "Multi-modal transfer learning for grasping transparent and specular objects," *IEEE Robot Autom Lett* **5**(3), 3791–3798 (2020).
[7] H. Fang, H.-S. Fang, S. Xu and C. Lu, "Transcg: A large-scale real-world dataset for transparent object depth completion and a grasping baseline," *IEEE Robot Autom Lett* **7**(3), 7383–7390 (2022).
[8] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng and S. Song, "Clear Grasp: 3d Shape Estimation of Transparent Objects for Manipulation," **In:** *2020 IEEE International Conference on Robotics and Automation (ICRA)*, (IEEE, 2020) pp. 3634–3642.
[9] Z. Ju, C. Yang, Z. Li, L. Cheng and H. Ma, "Teleoperation of Humanoid Baxter Robot Using Haptic Feedback," **In:** *2014 International Conference on Multisensor Fusion and Information Integration for Intelligent Systems (MFI)*, (IEEE, 2014) pp. 1–6.
[10] M. Dong and J. Zhang, "A review of robotic grasp detection technology," *Robotica* **41**(12), 3846–3885 (2023).
[11] Q. M. Marwan, S. C. Chua and L. C. Kwek, "Comprehensive review on reaching and grasping of objects in robotics," *Robotica* **39**(10), 1849–1882 (2021).
[12] S. Yu, D.-H. Zhai, Y. Xia, H. Wu and J. Liao, "Se-resunet: A novel robotic grasp detection method," *IEEE Robot Autom Lett* **7**(2), 5238–5245 (2022).
[13] D. Morrison, P. Corke and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int J Robot Res* **39**(2-3), 183–201 (2020).
[14] S. Kumra, S. Joshi and F. Sahin, "Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network," **In:** *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (IEEE, 2020) pp. 9626–9633.
[15] S. Wang, Z. Zhou and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," *IEEE Robot Autom Lett* **7**(3), 8170–8177 (2022).
[16] S. Yu, D.-H. Zhai and Y. Xia, "Egnet: Efficient robotic grasp detection network," *IEEE Trans Ind Electron* **70**(4), 4058–4067 (2022).
[17] D. Wang, C. Liu, F. Chang, N. Li and G. Li, "High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network," *IEEE Trans Ind Electron* **69**(11), 11611–11621 (2021).

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press

[18] N. Lu, Y. Cai, T. Lu, X. Cao, W. Guo and S. Wang, "Picking out the impurities: Attention-based push-grasping in dense clutter," *Robotica* **41**(2), 470–485 (2023).
[19] H. Tian, K. Song, S. Li, S. Ma, J. Xu and Y. Yan, "Data-driven robotic visual grasping detection for unknown objects: A problem-oriented review," *Expert Syst Appl* **211**, 118624 (2023).
[20] O. Kundu, S. Dutta and S. Kumar, "A novel method for finding grasping handles in a clutter using RGBD Gaussian mixture models," *Robotica* **40**(3), 447–463 (2022).
[21] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," (2017).arXiv preprint arXiv: 1703.09312.
[22] 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)*, (IEEE, 2018) pp. 7223–7230.
[23] D. Kim, A. Li and J. Lee, "Stable robotic grasping of multiple objects using deep neural networks," *Robotica* **39**(4), 735–748 (2021).
[24] I. Lenz, H. Lee and A. Saxena, "Deep learning for detecting robotic grasps," *Int J Robot Res* **34**(4-5), 705–724 (2015).
[25] J. Redmon and A. Angelova, "Real-Time Grasp Detection Using Convolutional Neural Networks," **In:** *2015 IEEE international conference on robotics and automation (ICRA)*, (IEEE, 2015) pp. 1316–1322.
[26] O. Ronneberger, P. Fischer and T. Brox, "U-net: Convolutional Networks for Biomedical Image Segmentation," **In:** *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference*, (Springer, 2015) pp. 234–241. Proceedings, Part III 18.
[27] Q. Gu, J. Su and X. Bi, "Attention Grasping Network: A Real-Time Approach to Generating Grasp Synthesis," **In:** *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, (IEEE, 2019) pp. 3036–3041.
[28] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo and Q. Hu, "ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks," **In:** *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (IEEE, 2020) pp. 11534–11542.
[29] S. Caldera, A. Rassau and D. Chai, "Review of deep learning methods in robotic grasp detection," *Multimodal Technol Interact* **2**(3), 57 (2018).
[30] A. Rodriguez, M. T. Mason and S. Ferry, "From caging to grasping," *Int J Robot Res* **31**(7), 886–900 (2012).
[31] A. Bicchi and V. Kumar, "Robotic Grasping and Contact: A Review," **In:** *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation Symposia Proceedings*, (IEEE, 2000) pp. 348–353.
[32] T. Tan, R. Alqasemi, R. Dubey and S. Sarkar, "Formulation and validation of an intuitive quality measure for antipodal grasp pose evaluation," *IEEE Robot Autom Lett* **6**(4), 6907–6914 (2021).
[33] Y. LeCun, Y. Bengio and G. Hinton, "Deep learning," *Nature* **521**(7553), 436–444 (2015).
[34] H. Van Hasselt, A. Guez and D. Silver, "Deep Reinforcement Learning with Double Q-learning," *Proc AAAI Conf Artif Intell* **30**(1) (2016).
[35] M. Niu, Z. Lu, L. Chen, J. Yang and C. Yang, "Vergnet: Visual enhancement guided robotic grasp detection under low-light condition," *IEEE Robot Autom Lett* **8**(12), 8541–8548 (2023).
[36] S. Kumra, S. Joshi and F. Sahin, "Gr-convnet v2: A real-time multi-grasp detection network for robotic grasping," *Sensors* **22**(16), 6208 (2022).
[37] S. Yu, D.-H. Zhai and Y. Xia, "Skgnet: Robotic grasp detection with selective kernel convolution," *IEEE Trans Autom Sci Eng* **20**(4), 2241–2252 (2022).
[38] F.-J. Chu, R. Xu and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robot Autom Lett* **3**(4), 3355–3362 (2018).
[39] F. Zhang, J. Leitner, M. Milford, B. Upcroft and P. Corke, "Towards vision-based deep reinforcement learning for robotic motion control," (2015).arXiv preprint arXiv: 1511.03791.
[40] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," **In:** *Proceedings of the IEEE conference on computer vision and pattern recognition*, (IEEE, 2018) pp. 7132–7141.
[41] K. He, X. Zhang, S. Ren and J. Sun, "Deep Residual Learning for Image Recognition," **In:** *Proceedings of the IEEE conference on computer vision and pattern recognition*, (IEEE, 2016) pp. 770–778.
[42] Y. Jiang, S. Moseson and A. Saxena, "Efficient Grasping from rgbd Images: Learning using a New Rectangle Representation," **In:** *2011 IEEE International conference on robotics and automation*, (IEEE, 2011) pp. 3304–3311.
[43] P. Shukla, N. Pramanik, D. Mehta and G. C. Nandi, "Generative model based robotic grasp pose prediction with limited dataset," *Appl Intell* **52**(9), 9952–9966 (2022).
[44] V. Kushwaha, P. Shukla and G. C. Nandi, "Generating quality grasp rectangle using Pix2Pix GAN for intelligent robot grasping," *Mach Vision Appl* **34**(1), 15 (2023).
[45] U. Asif, J. Tang and S. Harrer, " Graspnet: An Efficient Convolutional Neural Network for Real-Time Grasp Detection for Low-Powered Devices," **In:** *IJCAI*, (2018) pp. 4875–4882.
[46] A. Saxena, J. Driemeyer and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int J Robot Res* **27**(2), 157–173 (2008).

Cite this article: H. Xi, S. Li and X. Liu (2024). "A pixel-level grasp detection method based on Efficient Grasp Aware Network", Robotica **42**, 3190–3210. https://doi.org/10.1017/S0263574724001358

https://doi.org/10.1017/S0263574724001358 Published online by Cambridge University Press