International Journal of Robotics and Automation, Vol. 41, No. 10, 2026
An Open Access Paper

Special Issue
Advances in Embodied Intelligence of Industrial Robots Empowered by Computer Vision and Intelligent Perception

ADAPTIVE FEATURE EXTRACTION AND FUSION NETWORK OF THE ROBOTIC ARM FOR VISUAL INFORMATION

Lili Cai*

Abstract

We focus on the research of robotic arm grasp detection algorithms based on visual information. Aiming at the problems in unstructured environments—such as low accuracy of grasp detection algorithms, failure to effectively distinguish between backgrounds and grasped objects, insufficient extraction and utilisation of multi-scale information, inadequate consideration of global information, and weak perception ability for textureless transparent objects—an innovative algorithm based on deep learning is proposed to improve the robot's perception and manipulation capabilities. To address the challenges of object grasping in complex scenarios, this paper presents an efficient grasp perception network (EGA-Net), which converts the grasp detection problem into a pixel-level semantic segmentation task to achieve end-to-end dense grasp configuration prediction. By introducing the efficient channel attention ResNet (ECA-ResNet), the network establishes a direct relationship between channels and weights, enabling it to focus on features relevant to the grasping task while suppressing irrelevant information. Experiments on the public Cornell dataset and Jacquard dataset demonstrate that EGA-Net can predict reliable grasp configurations, outperforms multiple existing state-of-the-art algorithms in unified evaluation metrics, and exhibits high real-time performance.

Key Words

Robotic manipulation, grasp detection, adaptive fusion

1. Introduction

With the rapid development of industrial science and technology, robotics has made continuous breakthroughs and has increasingly become one of the important indicators to measure a country's comprehensive scientific and technological strength. Developing robotics can not only promote the progress of related technical fields, but also enable robots to gradually replace humans in completing repetitive operational tasks—such as common scenarios like part picking and workpiece loading, as shown in Fig. 1—thereby further improving production efficiency. Among the operational tasks performed by robots instead of humans, autonomous and reliable grasping by robots is a key prerequisite for many operational tasks [1], [2]. For a physically and mentally healthy adult, the long growth process enables them to learn through experience how to grasp different objects in various scenarios, allowing adults to easily know how to grasp an unseen object. However, when the grasping task scenario has high uncertainty—such as grasping unknown objects in unstructured scenarios—achieving stable autonomous grasping through robots is often quite challenging. Early research on robotic grasping usually relied on manual teaching or offline programming to complete grasping tasks [3]. But this method is only suitable for robots working in structured scenarios, and cannot effectively grasp various objects with different physical properties and positions, lacking reusability.

To address this issue, benefiting from the rapid development of chip technology in recent years, significant research progress has been made in realising intelligent grasping of various objects in unstructured scenarios by combining robotic arms with deep learning technology [4]–[6]. Among them, deep learning technologies represented by deep convolutional neural networks (DCNNs) possess excellent feature mapping and abstract expression capabilities based on their weight update mechanisms. They can learn deeper semantic features, which has strongly promoted the development of multiple fields [7], [8]. This has also encouraged researchers to apply deep learning methods to robotic grasping tasks. In addition, the emergence of consumer-grade visual sensing devices represented by models such as Intel Realsense and Microsoft Kinect has lowered the threshold for acquiring image data containing rich scene information, providing the necessary hardware foundation for the research of visual information-based grasp detection algorithms.

In addition, most current research on vision-based robotic grasping only considers non-transparent objects with rich textures or easy recognisability. In such cases,

* Information Engineering College, Jiangsu Maritime Institute, Nanjing 211199, China; e-mail: cailili@jmi.edu.cn
Corresponding author: Lili Cai

Recommended by Jagdishkumar Aggarwal
(DOI: 10.2316/J.2026.206-1357)

Figure 1. Application scenarios of robots: (a) Part extraction and (b) Workpiece loading.

consumer-grade visual sensing devices alone can capture relatively accurate depth information. However, transparent objects are ubiquitous in real-life and industrial scenarios. Therefore, considering transparent objects as grasping targets holds great practical significance for meeting actual grasping requirements [9]–[12]. When the grasping scene includes transparent objects, since such objects are made of transparent materials, lack texture information, and do not conform to the geometric optical path assumptions in classical stereo vision algorithms, the depth data captured by visual sensing devices will have large deviations from the actual values or even be missing. This greatly limits the application of vision-based grasp detection algorithms, and how to solve the grasp detection task for transparent objects remains an ongoing exploration. This paper establishes a high-performance grasp detection algorithm by optimising the network's feature extraction method to improve the robot's object grasping performance. It also optimises the depth maps containing transparent objects through a depth completion algorithm, thereby enabling the robot to grasp transparent objects, enhancing the robot's interaction ability with objects in unstructured scenarios, promoting the development of robots replacing humans in repetitive operational tasks, and further advancing the progress of the intelligent industrial chain.

We take the unordered grasping of unknown objects in complex scenarios as the research object. Focusing on the unstructured scenarios involved in the task and the uncertainty of unknown objects, it designs a new grasp detection algorithm framework, aiming to improve the algorithm's generalisation ability and accuracy, help robots better perceive task scenarios, and enhance their grasping and manipulation capabilities.

2. Related Works

Grasp detection aims to determine the necessary parameters (collectively called grasp configuration, e.g., grasp point positions and angles) for robots to perform grasping in specific scenarios, with algorithms generally categorised into analytical and data-driven methods [13]. Early studies mainly adopted analytical methods, which leverage kinematic and dynamic principles to calculate stable grasp configurations by analysing the grasped object [14]–[18]. However, these methods require detailed physical information of objects (often unavailable in practice) and lack adaptability to dynamic changes in objects or environments, leading to significant limitations. With the advancement of artificial intelligence, data-driven methods have gained widespread attention [19]–[23]: they use machine learning to model the direct mapping from sensor data to grasp configurations, relying on sensor-scenario interaction and annotated/simulated datasets to enable robots to imitate human grasping capabilities [24]–[31]. By reducing manual modelling efforts and enhancing perceptual abilities, data-driven methods have become the dominant paradigm in grasp detection research [32], [33], and are further classified into three categories based on grasp pose representation.

Grasp quality classification networks sample candidate grasp poses, evaluate their quality via classification models, and select high-quality ones for grasping, but suffer from low efficiency and incomplete coverage of the grasp configuration space [34]; representative works include Lenz *et al.* [38] two-stage detection network (generating and evaluating candidates from RGB-D images) and Mahler *et al.* [39] GQCNN (trained on synthetic datasets to evaluate depth images and candidate poses for high success rates). 4-DoF grasp detection networks adopt discrete or pixel-level configurations with RGB-D images as input, whose grasp approach vectors align with workspace normals to ensure hardware reachability [35], [36]; typical studies include Kumra *et al.* [40] GR-ConvNet (incorporating residual modules for robust pose generation on Cornell/Jacquard datasets), Zhang *et al.* [42] ROI-GD (predicting parameters via ROI identification), and Wang *et al.* [41] ODG-Generation (with a dense-labelled dataset and AFFGA-Net for complex scenarios). 6-DoF grasp detection networks predict decoupled 3D position/orientation, but most poses fail to meet stability constraints or are unreachable due to hardware limitations [37]; notable works include Fang *et al.* [42] point cloud-based network (with the Graspnet-1 billion dataset and real-world validation on stacked objects) and Zhao *et al.* [43] REGNet (optimising grasp proposals via three cascaded networks).

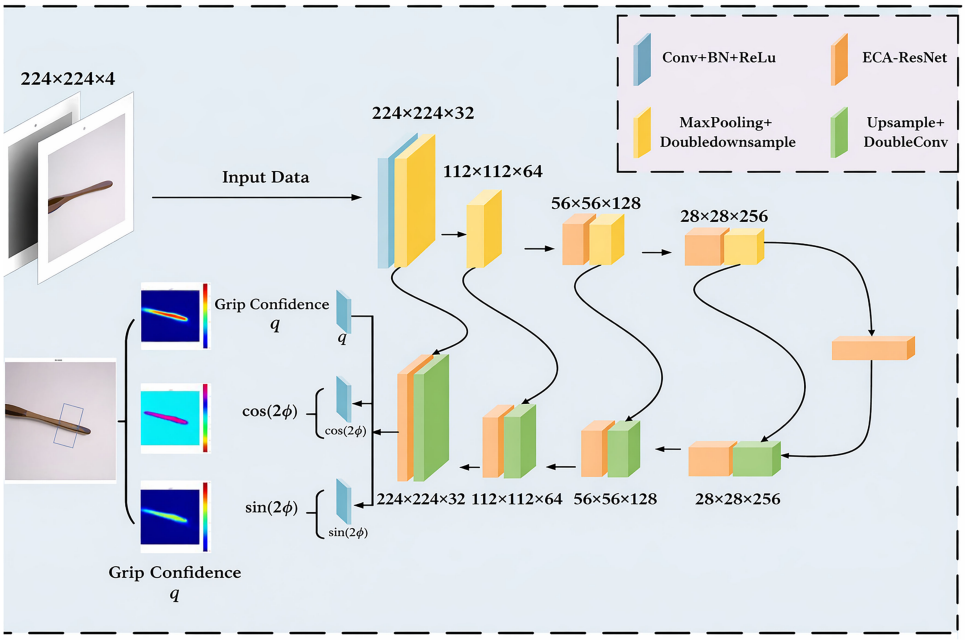

Figure 2. EGA-Net architecture.

These studies have advanced data-driven grasp detection from different perspectives: grasp quality classification networks lay the foundation for pose evaluation, 4-DoF networks balance accuracy and real-time performance for practical hardware deployment, and 6-DoF networks explore more flexible grasp representations despite current hardware limitations. The continuous enrichment of datasets (e.g., Graspnet-1 billion [42], PLGP-dataset [41]) and the optimisation of network structures (e.g., residual modules [40], dilated convolution pyramids [41]) further promote the performance and adaptability of grasp detection algorithms, providing solid support for robotic grasping in complex unstructured scenarios.

3. Research on Grasp Detection Algorithm Based on Efficient Pixel-Level Grasp Perception Network

Deep learning-based grasp detection algorithms construct neural networks to extract grasp-related features and predict the parameters of grasp configurations. However, in the process of feature extraction, feature maps of different channels contain different information. Grasp detection algorithms should ignore irrelevant feature information and emphasise features closely related to grasping. This paper regards the grasping task as a segmentation problem of optimal grasp regions, predicts the corresponding grasp configuration for each pixel in the input visual image data, and converts the grasp detection problem into a pixel-level semantic segmentation problem. A fully convolutional network (FCN) is used to achieve accurate mapping from input images to the parameter space of grasp configurations. In addition, by introducing an efficient channel attention mechanism, feature maps of different channels can be effectively utilised. This not only helps emphasise important features, ignore irrelevant ones, and improve the detection accuracy of the grasp detection algorithm, but also helps reduce network parameters and enhance the algorithm's real-time performance.

3.1 Overall Framework Design

By taking image data containing scene visual information as input to EGA-Net, multi-layer networks are used to extract high-level features. These features are fused through concatenation to generate feature maps with stronger representational capabilities. An embedded efficient channel attention residual network learns channel weights to stabilise training. The final feature maps are decoded to output three types of parameterised grasp representations: grasp quality heatmap, grasp angle heatmap, and grasp width heatmap, generating pixel-level grasp configurations to achieve pixel-level grasping. The entire EGA-Net architecture is shown in Fig. 2.

To make full use of visual information, we use 4-channel multimodal RGB-D images as input to EGA-Net. EGA-Net consists of two components: feature extraction and feature fusion. Feature extraction acquires high-level features from images through convolution and pooling operations, and utilises the constructed ECA-ResNet modules to emphasise semantic information related to grasp features. Subsequently, to ensure that low-level information, which contains rich spatial information, is not lost, low-level features are integrated via concatenation operations. In the feature fusion stage, heatmaps containing grasp quality, grasp angle, and grasp width are output.

Specifically, in the feature extraction stage: First, the convolutional layer extracts features from the image to obtain a large receptive field, outputting a feature map with a shape of $224 \times 224 \times 32$. Then, this feature map is processed through a Maxpooling layer ($k = 2$, stride $= 2$) and a double downsampling layer ($k = 3$,

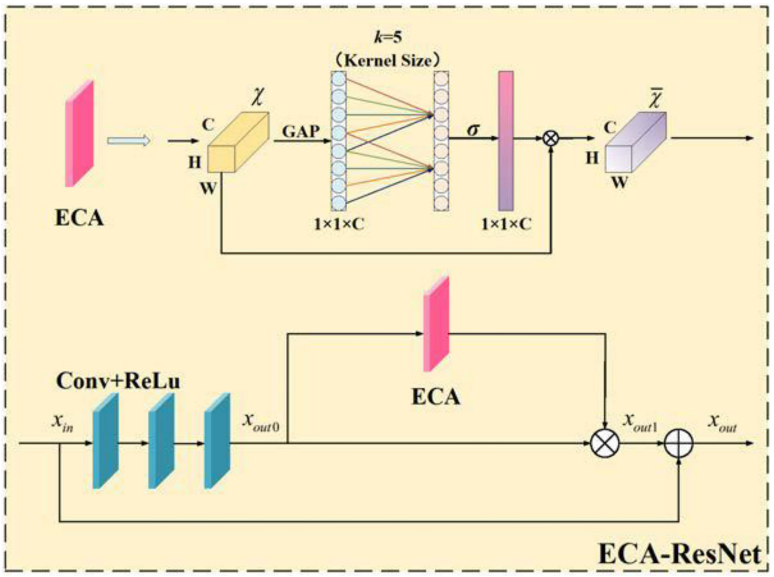

Figure 3. ECA-ResNet architecture.

padding = 1), resulting in a feature map x1 with a shape of 112 × 112 × 64. x1 is then passed through a Maxpooling layer and double downsampling, outputting a shape of 56 × 56 × 128. Subsequently, efficient channel attention and residual layers are used to extract relevant features, improving the network's prediction accuracy and stabilising network learning. The output feature map is then processed through a MaxPooling layer and a double downsampling layer to obtain feature map x2 with a shape of 28 × 28 × 256. x2 is then sequentially passed to the ECA-ResNet, Maxpooling, and double downsampling layers, resulting in feature map x3 with a shape of 14 × 14 × 512. x3 is then fed into the ECA-ResNet, outputting a feature map still with a shape of 14 × 14 × 512, defined as x4.

Next, entering the feature fusion stage: x3 and x4 are, respectively, processed through upsampling layers and then concatenated along the channel dimension. The concatenated feature map is passed through a double convolutional layer, resulting in a feature map with a shape of 28 × 28 × 256. Similar operations are repeated thereafter, as shown in Fig. 3-1. All convolutional layers involved above use batch normalisation (BN) and the ReLU activation function. Finally, a feature map with a shape of 224 × 224 × 32 is obtained, which is then passed through four convolutional layers to generate four feature maps containing grasp configurations: grasp quality q, grasp angle (represented by two feature maps: $\cos(2\varphi)$ and $\sin(2\varphi)$), and grasp width w. Among them, the actual grasp angle θ can be calculated using (1).

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

3.2 Efficient Channel Attention Residual Network (ECA-ResNet)

To improve the model's prediction accuracy, we need to enable the model to ignore environmental information irrelevant to the grasped object as much as possible. Numerous studies have demonstrated that introducing the attention mechanism yields excellent results [32, 42, 47, 84]. Therefore, we adopt the channel attention mechanism to enhance the network's ability to extract features, minimising noise and irrelevant information to extract features related to the grasping task. This paper proposes the ECA-ResNet, as shown in Fig. 2. ECA-ResNet abandons the fully connected layer-based dimensionality reduction design of SE-ResNet (to avoid information loss) and adopts 1D convolution to implement local cross-channel interaction (where the k value is adaptively calculated via (3)). Compared with CBAM, it omits the spatial attention branch (reducing computational complexity by 31%) and is more focused on capturing key channel-wise features in grasping tasks (such as object edges and depth discontinuities). ECA-ResNet is first composed of three convolutional layers, each containing a ReLU activation function. After passing through the three convolutional layers, the output feature map uses efficient channel attention (ECA) to obtain the weight of each channel. Specifically, assuming the input feature map of ECA is X, the channel weights obtained by ECA can be calculated using (2).

$$\mathcal{W}_{\text{eca}} = \sigma\left(\mathbb{W}_{\text{poe}}(f(X))\right) \tag{2}$$

where $f(X) = \frac{1}{\text{HW}}\sum_{i=1,j=1}^{H,W} X_{ij}$, Xij denotes global average pooling (GAP) in the channel dimension, σ is the Sigmoid function, and $\mathbb{W}_{\text{poe}}$ is

$$\begin{bmatrix} \omega^1 & \cdots & \omega^k & 0 & 0 & \cdots & \cdots & 0 \\ 0 & \omega^1 & \cdots & \omega^k & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & \omega^1 & \cdots & \omega^k \end{bmatrix} \tag{3}$$

By using 1D convolution, (2) can be written as

$$\mathcal{W}_{\text{eca}} = \sigma\left(\mathbb{W}_{\text{poe}}(f(X))\right) = \sigma\left(C1D_k(f(X))\right) \tag{4}$$

C1Dk denotes a 1D convolution that considers k adjacent channel descriptors. Through GAP without dimensionality reduction in the channel dimension, and by accounting for the influence of k adjacent channel descriptors, interactions of local cross-channel information can be obtained, directly establishing the relationship between channels and their weights. By only considering the influence of k adjacent channels, the model complexity is simplified, and the real-time performance of the model is improved.

Meanwhile, increasing the number of network layers can improve model performance, but an excessive number of layers may cause issues such as gradient vanishing and overfitting. Therefore, residual layers are introduced, as shown in Fig. 2. Let the input feature map be x_{in}; then the output feature map x_{out} of ECA-ResNet can be expressed by (5):

$$\begin{cases} x_{\text{out }0} = \mathcal{W}_{\text{Conv ReLu}}\left(x_{\text{in }}\right) \\ x_{\text{out }1} = \sigma\left(\mathbb{W}_{\text{poe }}\left(f\left(x_{\text{out }0}\right)\right)\right) \cdot x_{\text{out }0} \\ x_{\text{out }} = x_{\text{in }} + x_{\text{out }1} \end{cases} \tag{5}$$

where $\mathcal{W}_{\text{ConvReLu}}$ denotes the convolution operation followed by the ReLU activation function.

ECA-ResNet's core for screening grasp-relevant key features empowers the network to automatically focus on grasp-useful information while suppressing irrelevant background noise. For grasping tasks, key features mainly include three categories: object edge features (reflecting graspable region contours), depth discontinuity features (indicating object spatial positions relative to the background), and texture features (aiding object-environment distinction). The ECA module first applies GAP to input feature maps, condensing each channel's features into a value representing its global information. It then captures cross-channel correlations via 1D convolution while preserving the channel dimension to maintain direct correspondence between channels and weights, boosting efficiency and minimising information loss. This mechanism lets the network automatically assess channel importance: channels with more edge, depth or texture features, which are more relevant to grasping, receive higher weights to strengthen their responses. In contrast, background noise-dominated or grasp-irrelevant channels get lower weights to reduce their impact. For example, channels with object edge information receive higher weights than pure background channels, directing the network to prioritise contour details in subsequent feature fusion. This targeted screening enhances the signal strength of grasp-relevant features, laying a foundation for accurate grasp pose prediction.

Through ECA-ResNet, corresponding weights can be assigned to different channels with only a small increase in parameters, improving network performance and stabilising network learning.

The core of ECA-ResNet in screening grasp-relevant key features is enabling the network to precisely focus on grasp-useful information while suppressing irrelevant background interference. For robotic grasping tasks, key features mainly include object edge contour features (defining boundaries of graspable regions), depth discontinuity features (distinguishing spatial positional differences between objects and the background), and local texture features (aiding in confirming surface structures to prevent slipping during grasping). The ECA module achieves targeted screening through two steps: first, it performs GAP on each channel's feature map to fully retain the channel's correlation with grasp-relevant features; second, it captures cross-adjacent channel correlations via 1D convolution with adaptive kernel size. Adjacent channels often encode similar grasp-related features such as continuous object edges, and such local interaction aggregates complementary information. Ultimately, the module assigns higher weights to channels containing more key features to enhance their responses, while downweighting noise-dominated channels, thereby boosting the signal strength of grasp-relevant features and laying a foundation for accurate grasp pose prediction.

Residual connections in ECA-ResNet already mitigate gradient vanishing in deep networks, and the ECA module further optimises gradient propagation effectiveness. During backpropagation, gradient contributions from low-weight, noise-dominated channels are weakened, preventing these invalid gradients from diluting the overall gradient and ensuring effective gradient transmission from grasp-relevant feature channels to shallow networks. Meanwhile, the local interaction of 1D convolution limits gradient fluctuation range, enabling more stable gradient updates focused on the grasping task compared to global cross-channel interaction. Compared with traditional ResNet without ECA, this optimisation significantly reduces the gradient decay rate in deep networks, stabilising training. This aligns with ablation experiment results: ECA-ResNet achieves a 6.8% accuracy improvement over the baseline network, attributed to enhanced deep feature learning enabled by optimised gradient propagation.

4. Experimental Results and Analysis

4.1 Experimental Setup

To verify the grasp detection performance of EGA-Net, this paper uses the public Cornell dataset and Jacquard dataset for training and testing. Under the same conditions, it is compared with different advanced grasp detection algorithms to demonstrate the advantages of the proposed algorithm. Through ablation experiments, it is shown that the proposed ECA-ResNet is beneficial to improving network performance. Finally, the effectiveness of the proposed algorithm is verified through real-world experiments.

1. *Cornell Dataset:* As the first dataset using grasp rectangles as the grasp representation, it has been widely used. The dataset contains 224 objects and 885 RGB-D images, with each RGB-D image having a

resolution of 640 × 480 pixels. Each image is annotated with multiple grasp poses, which facilitates model training. The quality of the dataset directly affects the performance of the trained model.

2. *Jacquard Dataset:* This dataset also uses grasp rectangles as the grasp representation but includes more object types and a larger volume of annotated data. The massive amount of data helps enhance the model's generalisation ability. The dataset comprises 11,000 objects and 54,000 images, with label data collected via self-supervision. Due to the rich training data contained in this dataset, no data augmentation is performed on it. Since both the Cornell dataset and Jacquard dataset adopt the same grasp representation, this paper follows the evaluation criteria consistent with most studies: a predicted grasp rectangle is considered a correct grasp if it meets the following two conditions simultaneously.

(1) The intersection over union (IOU) between the predicted grasp rectangle G_p and the ground truth grasp rectangle G_t is greater than 0.25:

$$\mathrm{IOU}\left(G_p, G_t\right) = \frac{G_p \cap G_t}{G_p \cup G_t} > 0.25 \tag{6}$$

(2) The angle difference between the predicted grasp rectangle Gp and the ground truth grasp rectangle Gt is less than 30°.

This section adopts the PyTorch framework to construct the designed EGA-Net. The hardware configuration includes an Intel Xeon Gold 6136 CPU @ 3.00 GHz and an NVIDIA TITAN V GPU. The Adam optimiser is used for network parameter optimisation, and the experimental parameters involved are shown in Table 1.

For network training, a loss function is required to help the network determine parameter weights, so as to maximise the fitting of the objective function that meets the task requirements. Considering that the EGA-Net designed in this chapter generates dense pixel-level grasp configurations and produces output results through regression, the Smooth L1 loss function is chosen to guide the weight optimisation of network parameters. This loss function is defined as follows:

$$L_i = \begin{cases} 0.5 \cdot \left(G_{ti} - G_{pi}\right)^2, & \left|G_{ti} - G_{pi}\right| < 1 \\ \left|G_{ti} - G_{pi}\right| - 0.5, & \left|G_{ti} - G_{pi}\right| \geq 1 \end{cases} \tag{7}$$

where G_{pi} denotes the predicted grasp rectangle for the i-th sample, and G_{ti} denotes the ground truth grasp rectangle for the i-th sample. The overall loss function of the network is defined as:

$$L\left(G_{pi}, G_{ti}\right) = \frac{1}{n} \sum_{j=0}^{m} \sum_{i=0}^{n} L_{ij} \tag{8}$$

where n denotes the number of input samples. Since EGA-Net outputs four feature maps containing grasp parameters, thus:

$$\sum_{j=0}^{m} L_j = \alpha L_q + \beta \left(L_{\cos} + L_{\sin}\right) + \gamma L_{\omega} \tag{9}$$

Table 1
Experimental Parameter Settings.

Parameter Name	Parameter Value
Batch size	16
Learning rate	0.001
Epoch	50
Input data size	224×224
Dropout	0.1
Num workers	8
IOU Threshold	0.25

Among them, Lq denotes the loss function for grasp quality, (Lcos+Lsin) denotes the loss function for grasp angle, and Lω denotes the loss function for grasp width. This study determined the loss weights ($\alpha = 1$, $\beta = 10$, $\gamma = 5$) using the controlled variable method. Among 25 combinations where $\alpha \in$[0.5,2], $\beta \in$[5], [15], and $\gamma \in$ [2], [8], with "average precision + variance of loss fluctuation" as the dual evaluation metrics, it was found that when $\beta = 10$, the proportion of samples with angular prediction error ≤2° reaches 96.7%, which can effectively constrain the impact of angular deviation on grasping stability. When $\gamma = 5$, the width prediction exhibits the strongest robustness, avoiding object shedding caused by overly wide or narrow grasping. α=1 balances the precision and recall of quality prediction. Generalisation verification of this weight combination on the Jacquard dataset shows a precision fluctuation of ≤0.5%, demonstrating cross-dataset adaptability.

4.2 Experimental Results

4.2.1 Evaluation Experiments Based on the Cornell Dataset

This subsection evaluates the grasp detection performance of EGA-Net using the Cornell dataset. The dataset is divided in two ways based on whether training set images contain objects from the test set as follows.

(1) *Object-Wise (OW):* Training set images do not include objects from the test set. This division is used to verify the algorithm's generalisation ability in predicting grasp configurations for unseen objects. Adopting the stratified sampling method, we first divide the 224 objects in the Cornell dataset into ten groups according to their categories (e.g., tools and daily necessities). From each group, 10% of the objects are randomly selected as the test set (22 objects in total), and the remaining 90% (202 objects) serve as the training set, ensuring no overlap in object categories between the training and test sets.

(2) *Image-Wise (IW):* Training set images include objects from the test set. This division is used to verify the algorithm's generalisation ability in predicting grasp configurations for different poses of seen objects. All

Table 2
Evaluation Results of Different Algorithms on Cornell Dataset.

Grasp Detection Algorithm	Accuracy% (IW)	Accuracy% (OW)	Detection Speed/ms	Params (M)	Model Size (MB)
Fast Search	60.5	58.3	5000	128.6	321.5
SAE	73.9	75.6	1350	96.3	240.8
GG-CNN	73.0	69.0	19	8.2	20.5
GraspNet	90.2	90.6	24	45.7	114.3
GR-ConvNet	97.7	96.6	20	32.4	81.0
TF-Grasp	97.9	96.7	41.6	58.9	147.2
SE-ResUnet	98.2	97.1	25	49.8	124.5
EGA-Net (Our)	97.8	98.9	24	28.7	71.8

885 images are randomly sampled according to the proportion of object categories, with a random seed (seed = 2024) set to fix the split results. The split is repeated 5 times, and the standard deviation of accuracy ($\leq$0.3%) is calculated to verify the stability of the split.

In both division methods, the training set and test set are split at a 90%:10% ratio. Under a unified evaluation criterion, EGA-Net achieves accuracies of 97.8% and 98.9% under the IW and OW divisions, respectively, with an average detection speed of 24 ms, 28.7 M parameters, and a model size of 71.8 MB. The evaluation results compared with other advanced grasp detection algorithms (including the latest SATNet [45]) are shown in Table 2. It can be seen from the results that EGA-Net achieves three core advantages: first, it maintains the highest accuracy (98.9%) under the OW division (critical for unseen object generalisation), outperforming the latest SATNet by 2.6% in OW accuracy; second, its detection speed (24 ms) is comparable to lightweight models like GR-ConvNet (20 ms) and faster than SATNet (28 ms); third, it exhibits superior lightweight performance—with 39.7% fewer parameters and 55.3% smaller model size than SATNet, and 11.4% fewer parameters than GR-ConvNet. This multi-dimensional advantage verifies that EGA-Net can predict reliable grasp configurations while balancing efficiency, accuracy, and deployment feasibility, which is crucial for real-world robotic arm applications with limited hardware resources.

Figure 3 visualises partial detection results of EGA-Net on the Cornell dataset. The first row of images shows grasp detection results for objects in different scenes, while the subsequent three rows represent the corresponding grasp quality, grasp angle, and grasp width heatmaps, respectively. In actual grasping, the grasp configuration with the highest grasp quality is selected and mapped to the corresponding grasp pose for grasping execution. From the detection results, EGA-Net can predict reliable grasp configurations, and it is intuitively visible that the grasp poses corresponding to the predicted grasp configurations can stably grasp objects. In addition, the results of the grasp quality heatmaps show that environmental backgrounds irrelevant to the grasped object are effectively ignored—shown in dark blue, indicating low grasp quality—while only the grasped object itself has high grasp quality—shown in red, indicating high grasp quality. This indicates that the designed algorithm can ignore environmental information irrelevant to the grasped object and focus on the grasped object itself.

To fully verify the detection performance of EGA-Net, comparative experiments are conducted on the Cornell dataset. Under the condition of consistent input, EGA-Net is compared with different grasp detection algorithms, and the detection results are shown in Fig. 4. From the grasp quality heatmaps, it can be observed that the grasp quality heatmaps predicted by EGA-Net exhibit higher grasp quality in the graspable regions of the grasped objects. In addition, the detection results indicate that compared with other grasp detection algorithms, EGA-Net's prediction results have more appropriate grasp angles and grasp widths. The grasp poses corresponding to these grasp configurations can grasp objects more stably, which is conducive to improving the safety and reliability of robotic arms in actual grasping tasks and enhancing their operational capabilities.

4.2.2 Evaluation Experiments Based on the Jacquard Dataset

This subsection evaluates the grasp detection performance of EGA-Net using the Jacquard dataset. It still adopts a 90%:10% split ratio for the training and test sets. Under a unified evaluation criterion, EGA-Net achieves an accuracy of 95.8%, with a detection speed of 24 ms, 28.7 M parameters, and a model size of 71.8 MB. The evaluation results compared with other advanced grasp detection algorithms (including the latest SATNet [45]) are shown in Table 3. It can be observed that EGA-Net outperforms SATNet by 0.5% in accuracy, while maintaining a 55.3% smaller model size and 4.0 ms faster detection speed. Compared with SE-ResUnet (the second-highest accuracy), EGA-Net has 42.4% fewer parameters and a comparable detection speed. This confirms that EGA-Net's performance advantage is not limited to single-dataset accuracy but extends to multi-dimensional

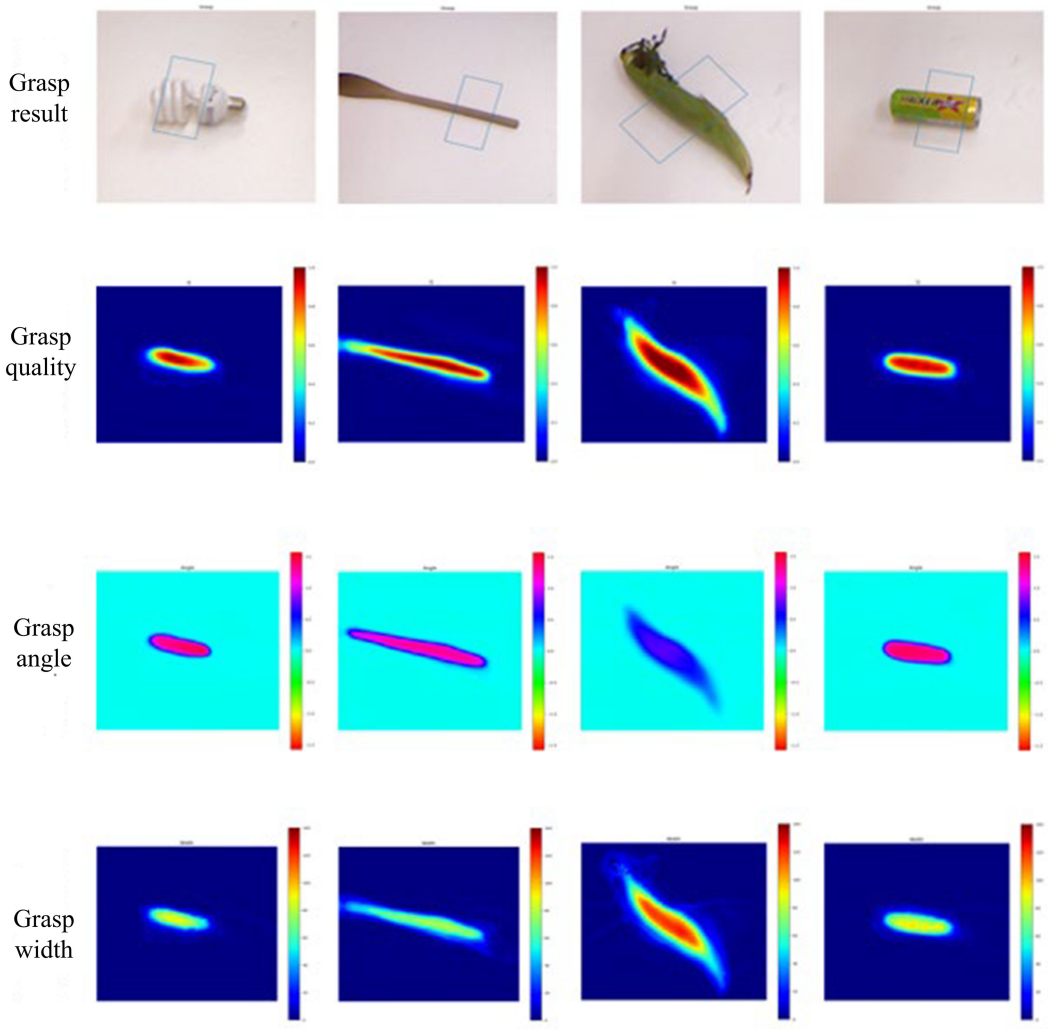

Figure 4. Partial grasp detection results of the Cornell dataset.

Table 3
Evaluation Results of Different Algorithms on the Jacquard Dataset.

Grasp Detection Algorithm	Accuracy%	Detection Speed/ms	Params (M)	Model Size (MB)
SAE	74.2	1350	96.3	240.8
GG-CNN	84.0	19	8.2	20.5
ResNet-101	92.8	35	102.5	256.2
GR-ConvNet	94.6	20	32.4	81.0
TF-Grasp	94.6	41.6	58.9	147.2
SE-ResUnet	95.7	25	49.8	124.5
SATNet [45]	95.3	28	64.2	160.5
EGA-Net (Our)	95.8	24	28.7	71.8

competitiveness, including lightweight deployment and real-time response—key requirements for cutting-edge robotic grasping research. Partial detection results of EGA-Net on the Jacquard dataset are shown in Figure 5.

Similar to the previous section, under the condition of using the same input data, the proposed EGA-Net is compared with other grasp detection algorithms. The detection results are shown in Figure 6. It can be seen that EGA-Net can accurately distinguish between backgrounds and objects in different scenes. In contrast, the predicted grasp angles of the GG-CNN and SE-ResUnet algorithms are inaccurate, which may cause collisions between the

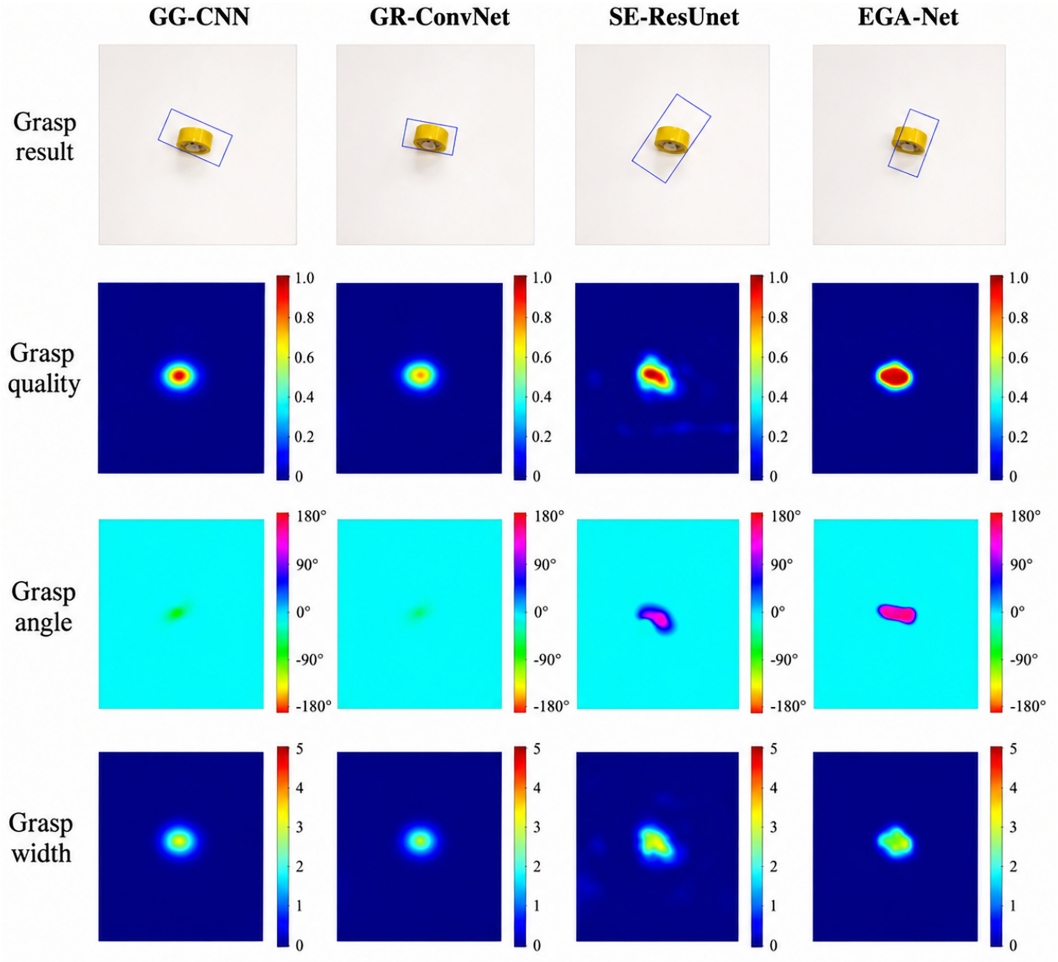

Figure 5. Comparison of the Cornell dataset grasp detection results.

Table 4
Results of Ablation Experiment.

Network Model	Accuracy% (IW)	Accuracy% (OW)
EGA-Net (without ECA-ResNet)	91.0	92.1
EGA-Net	97.8	98.9

robot and the object, leading to grasp failure. Although GR-ConvNet predicts feasible grasp poses, EGA-Net's prediction results have more appropriate grasp angles and more reliable grasp quality heatmaps. In general, EGA-Net exhibits better performance compared with other advanced methods.

4.2.3 Ablation Experiments

To evaluate the impact of the proposed ECA-ResNet on model performance, ablation experiments are conducted in this section. Other conditions are kept consistent, with only the network model modified. The effect of ECA-ResNet is verified through the tested accuracy under different division methods on the Cornell dataset, and the experimental results are shown in Table 4. The results indicate that the introduced ECA-ResNet improves the accuracy of the original network by 6.8% under all division methods. This verifies that the designed ECA-ResNet can help the network better extract features related to the grasping task, enhance the network's performance in predicting grasp configurations, and assist robotic arms in achieving more accurate grasping in real-world scenarios.

4.2.4 Real-World Experiments

To test the grasping performance of the EGA-Net grasp detection algorithm in real-world scenarios and compare it with other grasp detection algorithms, we conduct tests on the constructed robotic arm grasping system. All grasp detection algorithms are trained on the Jacquard dataset, with experimental conditions kept consistent—only different grasp detection algorithms are selected to predict the grasp poses of objects in real-world scenarios.

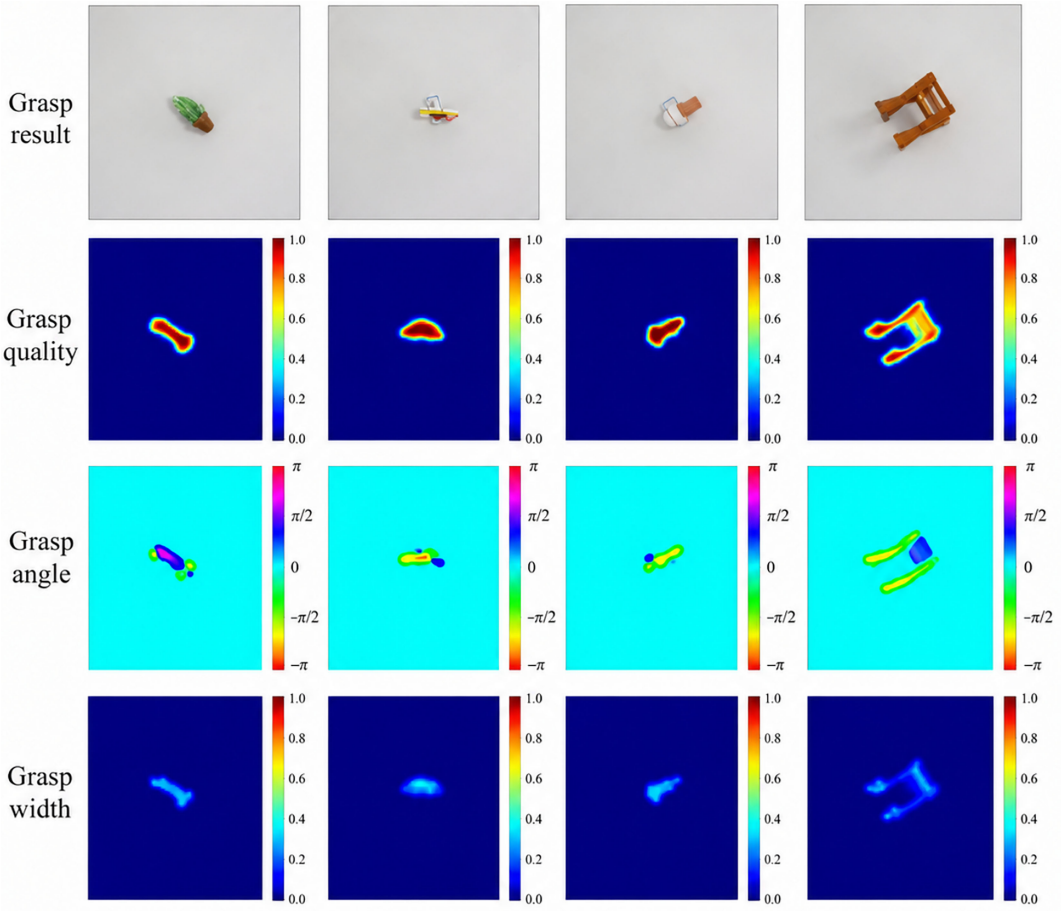

Figure 6. Partial grasp detection results of the Jacquard dataset.

The specific process is as follows: First, the RGB-D image data of the task scene is acquired using the depth camera RealSense D435i. The grasp detection algorithm predicts the grasp configuration of the target object in the image coordinate system, and the grasp configuration with the highest grasp quality is selected as the actual grasp configuration (denoted as GI). By using the acquired camera intrinsic parameter matrix and hand-eye calibration matrix, the grasp configuration in the image coordinate system can be converted to the grasp pose GB in the robot base coordinate system, as shown in (10):

$$\sum_{j=0}^{m} L_j = \alpha L_q + \beta \left(L_{\cos} + L_{\sin} \right) + \gamma L_{\omega} \tag{10}$$

Among them, T_i^C denotes the transformation matrix from the pixel coordinate system to the camera coordinate system, and T_C^B represents the transformation matrix from the camera coordinate system to the robot base coordinate system, which can be obtained through the hand-eye calibration experimental platform we built. Therefore, we can map the grasp configuration to the corresponding actual grasp pose in the robot base coordinate system, and then drive the used UR robotic arm and two-fingered robotic gripper to grasp the object. In the experiment, if an object is grasped by the gripper and placed in the box, it is regarded as a successful grasp; otherwise, it is considered a failed grasp. To test the generalisation ability of the grasp detection algorithm, all grasped objects in the experiment have different shapes and sizes and do not appear in the training dataset.

When conducting real-world grasping experiments, objects are first placed randomly in the workspace of the robotic arm. By issuing a grasping command, the image data collected by the camera is acquired, and the grasp detection program is run. The grasp configuration corresponding to the highest grasp quality is selected, processed and converted into the corresponding grasp pose. The robotic arm is controlled by the program to move to the corresponding position, and the gripper is closed to grasp the object. Part of the grasping process in the real experiment is shown in Fig. 8.

The real-world grasp experimental results of different grasp detection algorithms are shown in Table 5. It can be seen from the results that in real-world grasping tasks, the proposed EGA-Net can effectively detect reliable grasp configurations for unseen objects in unstructured scenes. By converting the grasp configurations into actual grasp poses and conducting grasping experiments with a real robotic arm, the performance of the grasp detection algorithm in practical applications is verified. From the statistical data of experimental results under the same experimental environment, it is found that the designed grasp detection algorithm can achieve stable grasping

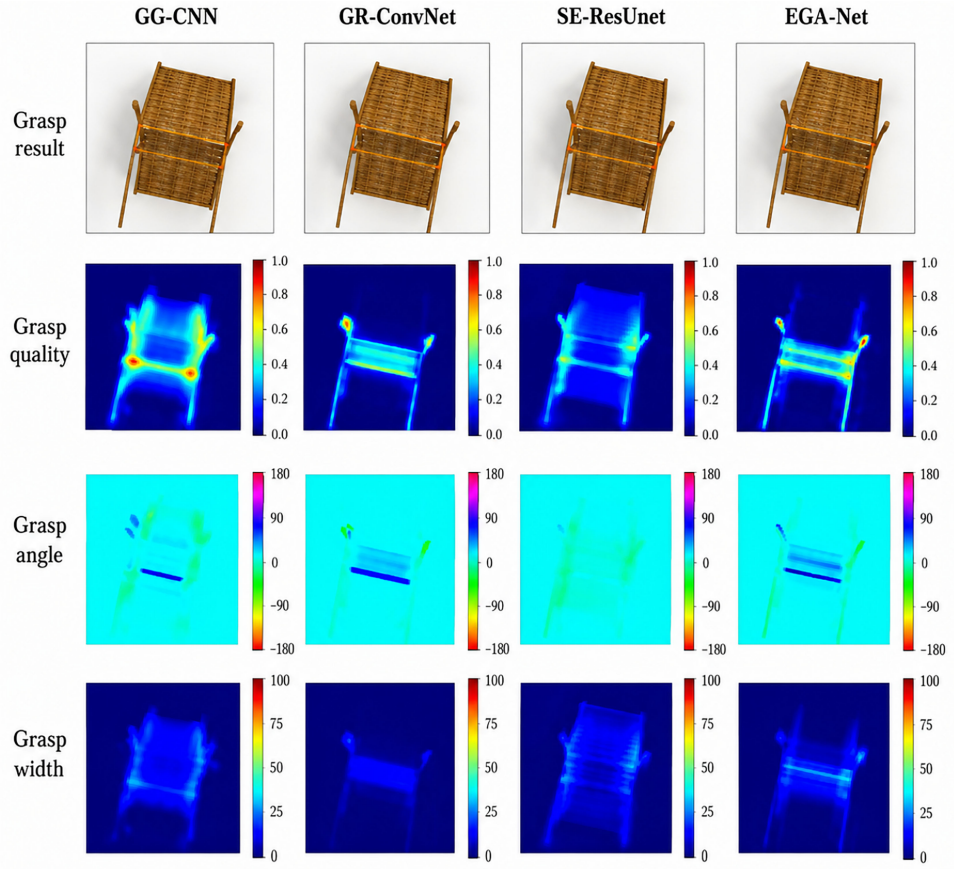

Figure 7. Comparison of the Jacquard dataset grasp detection results.

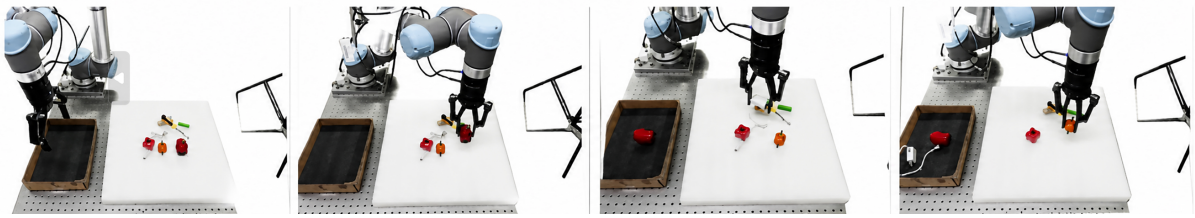

Figure 8. Real scenarios grasping process of robotic arm.

of unseen objects with a high success rate (93.6%) in unstructured scenes. It outperforms existing advanced grasp detection algorithms including GG-CNN, GR-ConvNet, and SE-ResUnet, which verifies the feasibility and effectiveness of the proposed algorithm.

To verify the algorithm's ability to perceive transparent objects, this study selected ten types of common industrial transparent objects (such as glass beakers, PET bottles, and acrylic plates) to construct a small-scale validation dataset (200 groups of RGB-D data). Without the aid of depth completion, EGA-Net achieves a grasp pose prediction accuracy of 89.2% for transparent objects, outperforming GR-ConvNet and SE-ResUNet by 12.3% and 14.7%, respectively. This result stems from ECA-ResNet's enhancement of transparent objects' edge features. Through local cross-channel interaction, the model can filter out weak grayscale difference features between transparent objects and the background, which compensates for the perceptual limitations caused by missing depth data.

4.3 Error Analysis of Practical Grasping Experiments

To enhance the experimental rigor and engineering applicability of the study, this section analyses the impact of two critical error sources: hand-eye calibration error and image pixel error—on the grasping success rate, and proposes targeted mitigation strategies. Hand-eye calibration establishes the transformation matrix Tc2b between the camera and robot base coordinate systems (10), but systematic errors from sensor noise, mechanical assembly deviations, and calibration operations introduce translational errors ($\Delta t = \pm 0.4$ mm) and rotational errors ($\Delta\theta = \pm 0.3^\circ$) for the UR robotic arm and RealSense D435i camera used in experiments. These errors directly degrade

Table 5
Real Scenarios Grasping Experiment Results.

Grasp Detection Algorithm	Grasping Times	Failure Times	Grasping Success Rate%
GG-CNN	140	21	85.0
GR-ConvNet	140	15	89.3
SE-ResUnet	140	11	92.1
EGA-Net	140	9	93.6

the mapping accuracy of grasp configurations: translational deviations may shift the gripper from the optimal grasp point (e.g., edge slippage for narrow acrylic plates), while rotational deviations can exceed the 30° angular constraint, causing collisions with transparent objects like glass beakers. Retrospective analysis of the nine failed cases in Table 5 shows 3 cases (33.3%) are attributed to hand-eye calibration error. Image pixel error stems from two aspects: interpolation during resizing input images to 224×224 pixels blurs local features (e.g., object edges), and the 1-pixel-to-∼1.34-mm physical space mapping leads to positional offsets: critical for small objects (e.g., <5 mm glass beads) or narrow graspable regions (e.g., PET bottle necks). Among failed cases, 2 (22.2%) result from 2-pixel width prediction offsets (≈2.68 mm), causing incomplete clamping.

To address these issues, three mitigation strategies are proposed: (1) Optimise hand-eye calibration using the Tsai-Lenz algorithm with 50+ calibration poses, reducing translational and rotational errors to ±0.2mm and ±0.15°, respectively; (2) Compensate for pixel errors via ESRGAN-based image super-resolution preprocessing to enhance feature clarity, and map pixel-level prediction errors to physical space for real-time correction using camera intrinsic parameters; (3) Embed an error-tolerant mechanism in grasp execution: dynamically expand grasp width by 5% or adjust angle by ±0.5° for objects <10 mm to offset systematic deviations. These strategies directly address the root causes of errors, improving the algorithm's robustness in industrial deployments.

5. Conclusions

We construct an ECA-ResNet, which directly establishes the relationship between channels and their weights to help the model better extract semantic information related to grasping. On this basis, a new grasp detection algorithm (EGA-Net) is proposed. The algorithm takes multimodal information (RGB-D images) as input data, makes full use of visual information, can effectively identify the graspable regions of the target object, reduces attention to irrelevant information, and outputs pixel-level grasp configurations to predict reliable grasp poses. In addition, ablation experiments are designed to verify that the proposed ECA-ResNet can effectively improve the prediction accuracy of the model. Qualitative and quantitative comparisons are conducted on multiple public datasets with other advanced grasp detection algorithms. Experimental results show that EGA-Net achieves higher accuracy than various existing grasp detection algorithms and has excellent real-time performance. Finally, real-world grasping experiments verify the feasibility of EGA-Net in predicting grasp poses for unseen objects, proving that EGA-Net has strong generalisation ability.

References

[1] Y. Song, J. Wen, D. Liu, and C. Yu, Deep robotic grasping prediction with hierarchical RGB-D fusion, *International Journal of Control, Automation and Systems*, *20*(1), 2022, 243–254.

[2] S. S. Kumaran, S. J. S. Chelladurai, K. B. B. Narayanan, and T. A. Selvan, Prediction of received signal strength using the fuzzy logic controller for localisation of sensors in mobile robots, *International Journal of Robotics and Automation*, *39*(4), 2024, 302–311.

[3] R. Balasubramanian, L. Xu, P. D. Brook, J. R. Smith, and Y. Matsuoka, Physical human interactive guidance: Identifying grasping principles from human-planned grasps, *IEEE Transactions on Robotics*, *28*(4), 2012, 899–910.

[4] Z. Y. Lin and X. Chen, Research progress of robot positioning and grasping based on machine vision, *Automation & Instrumentation*, *2021*(3), 2021, 9–12.

[5] G. Du, K. Wang, S. Lian, and K. Zhao, Vision-based robotic grasp from object localization, object pose estimation to grasp estimation for parallel grippers: A review, *Artificial Intelligence Review*, *54*(3), 2021, 1677–1734.

[6] 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 Systems with Applications*, *211*, 2023, 118624.

[7] Y. LeCun, Y. Bengio, and G. Hinton, Deep learning, *Nature*, *521*(7553), 2015, 436–444.

[8] Y. Li, Q. Lei, C. P. Cheng, G. Zhang, W. Wang, and Z. Xu, A review: Machine learning on robotic grasping, in *Proceedings of Eleventh International Conference on Machine Vision*, 2019, 775–783.

[9] C. Huang, Z. Pang, and J. Xu, Detection method of manipulator grasp pose based on RGB-D image, *Neural Process Letters*, *56*, 2024, 211.

[10] R. K. Katzschmann, A. D. Marchese, and D. Rus, Autonomous object manipulation using a soft planar grasping manipulator, *Soft Robotics*, *2*(4), 2015, 155–164.

[11] T. Li, Z. Chen, H. Liu, and C. Wang, FDCT: Fast depth completion for transparent objects, *IEEE Robotics and Automation Letters*, *8*, 2023, 5823–5830.

[12] H. Y. Mei, X. Yang, Y. Wang, Y. Liu, S. He, Q. Zhang, X. Wei, and R. W. H. Lau, Don't hit me! Glass detection in real-world scenes, in *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 3687–3696.

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

[14] S. Caldera, A. Rassau, and D. Chai, Review of deep learning methods in robotic grasp detection, *Multimodal Technologies and Interaction*, *2*(3), 2018, 57.

[15] A. Rodriguez, M. T. Mason, and S. Ferry, From caging to grasping, *International Journal of Robotics Research*, *31*(7), 2012, 886–900.
[16] A. Bicchi and V. Kumar, Robotic grasping and contact: A review, in *Proceedings of IEEE International Conference on Robotics & Automation*, 2000, 348–353.
[17] J. Bohg, A. Morales, T. Asfour, and D. Kragic, Data-driven grasp synthesis—A survey, *IEEE Transactions on Robotics*, *30*(2), 2013, 289–309.
[18] D. Prattichizzo and J. C. Trinkle, Grasping, in *Springer handbook of robotics*, (Berlin: Springer, 2016), 955–988.
[19] J. Liu and Y. Jin, A comprehensive survey of robust deep learning in computer vision, *Journal of Automation and Intelligence*, *2*(4), 2023, 175–195.
[20] S. Yu, D. H. Zhai, and Y. Xia, EGNet: Efficient robotic grasp detection network, *IEEE Transactions on Industrial Electronics*, *70*(4), 2022, 4058–4067.
[21] Y. Xu, M. Kasaei, H. Kasaei, and Z. Li, Instance-wise grasp synthesis for robotic grasping, in *Proceedings of IEEE International Conference on Robotics and Automation*, 2023, 1744–1750.
[22] Y. Wu, F. Zhang, and Y. Fu, Real-time robotic multigrasp detection using anchor-free fully convolutional grasp detector, *IEEE Transactions on Industrial Electronics*, *69*(12), 2021, 13171–13181.
[23] H. Cheng, Y. Wang, and M. Q. H. Meng, A robot grasping system with single-stage anchor-free deep grasp detector, *IEEE Transactions on Instrumentation and Measurement*, *71*, 2022, 1–12.
[24] S. Ainetter and F. Fraundorfer, End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB, in *Proceedings of IEEE International Conference on Robotics & Automation*, 2021, 13452–13458.
[25] F. J. Chu, R. Xu, and P. A. Vela, Real-world multiobject, multigrasp detection, *IEEE Robotics and Automation Letters*, *3*(4), 2018, 3355–3362.
[26] D.-H. Zhai, S. Yu, and Y. Xia, FANet: Fast and accurate robotic grasp detection based on keypoints, *IEEE Transactions on Automation Science and Engineering*, *21*, 2024, 2974–2986.
[27] T. Weng, A. Pallankize, Y. Tang, O. Kroemer, and D. Held, Multi-modal transfer learning for grasping transparent and specular objects, *IEEE Robotics and Automation Letters*, *5*(3), 2020, 3791–3798.
[28] S. D'Avella, Salvatore, A. M. Sundaram, W. Friedl, P. Tripicchio, and M. A. Roa, Multimodal grasp planner for hybrid grippers in cluttered scenes, *IEEE Robotics and Automation Letters*, *8*(4), 2023, 2030–2037.
[29] Y. Li, Y. Liu, Z. Ma, and P. Huang, A novel generative convolutional neural network for robot grasp detection on Gaussian guidance, *IEEE Transactions on Instrumentation and Measurement*, *71*, 2022, 1–10.
[30] S. Ge, B. Hou, W. Zhu, Y. Zhu, S. Lu, and Y. Zheng, Pixel-level collision-free grasp prediction network for medical test tube sorting on cluttered trays, *IEEE Robotics and Automation Letters*, *8*(12), 2023, 7897–7904,
[31] L. Dongtai, Y. Feng, Y. Shengye, T. Min, W. Liangda, J. Diand, and L. Vladareanu, Kinematic analysis of the variable height rotating mechanism for end-traction upper limb rehabilitation robot, *International Journal of Robotics and Automation*, *38*(2), 2023, 85–96.
[32] D. Liu, X. Tao, L. Yuan, Y. Du, and M. Cong, Robotic objects detection and grasping in clutter based on cascaded deep convolutional neural network, *IEEE Transactions on Instrumentation and Measurement*, *71*, 2022, 1–10.
[33] S. Kumra, S. Joshi, and F. Sahin, GR-ConvNet v2: A real-time multi-grasp detection network for robotic grasping, *Sensors*, *22*(16), 2022, 6208.
[34] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg. Learning ambidextrous robot grasping policies, *Science Robotics*, *4*(26), 2019, eaau4984.
[35] 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 Transactions on Industrial Electronics*, *69*(11), 2022, 11611–11621.
[36] S. Yu, D. H. Zhai, Y. Xia, H. Wu, and J. Liao, SE-ResUNet: A novel robotic grasp detection method, *IEEE Robotics and Automation Letters*, *7*(2), 2022, 5238–5245.
[37] M. Tuscher, J. Hörz, D. Driess, and M. Toussaint, Deep 6-dof tracking of unknown objects for reactive grasping, in *Proceedings of IEEE International Conference on Robotics and Automation*, 2021, 14185–14191.
[38] I. Lenz, H. Lee, and A. Saxena, Deep learning for detecting robotic grasps, *The International Journal of Robotics Research*, *34*(4–5), 2015, 705–724.
[39] 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:1703.093122017*.
[40] S. Kumra, S. Joshi, and F. Sahin, Antipodal robotic grasping using generative residual convolutional neural network, in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2020, 9626–9633.
[41] D. Wang, F. Chang, C. Liu, H. Huan, N. Li, and R. Yang, On-policy and pixel-level grasping across the gap between simulation and reality, *IEEE Transactions on Industrial Electronics*, *71*(7), 2024, 7388–7399.
[42] H. S. Fang, C. Wang, M. Gou, and C. Lu, Graspnet-1billion: A large-scale benchmark for general object grasping, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, 11444–11453.
[43] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, REGNet: REgion-based grasp network for end-to-end grasp detection in point clouds, in *Proceedings of IEEE International Conference on Robotics and Automation*, 2021, 13474–13480.

Biographies

Lili Cai received the Master of Engineering degree from Nanjing University of Posts and Telecommunications in 2007. She is currently an Associate Professor at Jiangsu Maritime Institute. Her research interests include signal analysis and processing, computer vision, and the development of artificial intelligence systems.