applied sciences

MDPI

Article

FAGD-Net: Feature-Augmented Grasp Detection Network Based on Efficient Multi-Scale Attention and Fusion Mechanisms

Xungao Zhong [1,2,*], Xianghui Liu [1], Tao Gong [1], Yuan Sun [1,2], Huosheng Hu [3] and Qiang Liu [4]

[1] School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen 361024, China; 2122031347@stu.xmut.edu.cn (X.L.); gongtao@stu.xmut.edu.cn (T.G.); sunyuan@xmut.edu.cn (Y.S.)
[2] Xiamen Key Laboratory of Frontier Electric Power Equipment and Intelligent Control, Xiamen 361024, China
[3] School of Computer Science and Electronic Engineering, University of Essex, Colchester CO4 3SQ, UK; hhu@essex.ac.uk
[4] School of Engineering Mathematics and Technology, Faculty of Engineering, University of Bristol, Beacon House, Queens Rd, Bristol BS8 1QU, UK; qiang.liu@bristol.ac.uk
* Correspondence: zhongxungao@163.com or zhongxungao@xmut.edu.cn; Tel.: +86-189-5921-6800

Abstract: Grasping robots always confront challenges such as uncertainties in object size, orientation, and type, necessitating effective feature augmentation to improve grasping detection performance. However, many prior studies inadequately emphasize grasp-related features, resulting in suboptimal grasping performance. To address this limitation, this paper proposes a new grasping approach termed the Feature-Augmented Grasp Detection Network (FAGD-Net). The proposed network incorporates two modules designed to enhance spatial information features and multi-scale features. Firstly, we introduce the Residual Efficient Multi-Scale Attention (Res-EMA) module, which effectively adjusts the importance of feature channels while preserving precise spatial information within those channels. Additionally, we present a Feature Fusion Pyramidal Module (FFPM) that serves as an intermediary between the encoder and decoder, effectively addressing potential oversights or losses of grasp-related features as the encoder network deepens. As a result, FAGD-Net achieved advanced levels of grasping accuracy, with 98.9% and 96.5% on the Cornell and Jacquard datasets, respectively. The grasp detection model was deployed on a physical robot for real-world grasping experiments, where we conducted a series of trials in diverse scenarios. In these experiments, we randomly selected various unknown household items and adversarial objects. Remarkably, we achieved high success rates, with a 95.0% success rate for single-object household items, 93.3% for multi-object scenarios, and 91.0% for cluttered scenes.

Keywords: robotic grasping; deep network model; attention mechanism; feature fusion

Citation: Zhong, X.; Liu, X.; Gong, T.; Sun, Y.; Hu, H.; Liu, Q. FAGD-Net: Feature-Augmented Grasp Detection Network Based on Efficient Multi-Scale Attention and Fusion Mechanisms. *Appl. Sci.* **2024**, *14*, 5097. https://doi.org/10.3390/app14125097

Academic Editors: Miguel Angel Cazorla, Francisco Gomez-Donoso and Félix Escalona Moncholí

Received: 6 April 2024
Revised: 29 May 2024
Accepted: 29 May 2024
Published: 12 June 2024

1. Introduction

With the improvement in robot autonomous manipulation, grasping has been extensively studied in recent years [1,2]. Artificial Intelligence (AI)-based robot grasping methods possess significant application prospects, such as in industrial workpiece sorting [3]. While traditional robot grasping approaches applied in industry with 2D or 3D vision require knowledge of the candidate object model (size, orientation and type), deep learning grasping methods do not rely on such information because of their strong learning and reasoning capabilities. However, the precise grasping of a random object still remains challenging because of uncertainty in target size, orientation, and type. Thus, there are issues that need further resolution in order to model the relationship between visual image and reliable grasping pose, which is required to extract grasp-related features effectively and generalize the learned knowledge to new objects.

Data-driven approaches such as deep learning [4–8] for robot grasping are becoming mainstream and showing great potential in unstructured scenarios. In the work [9], a pioneering representation method for grasping and detecting rectangular boxes was

introduced. Subsequently, numerous deep learning methods were developed. Among of them, detection-based approaches demonstrate superior performance. Detection-based approaches generate pixel-level heatmaps of grasp detection from an N-channel input feature, where each pixel indicates a distinct level of grasp confidence. Morrison et al. [4] proposed a Generative Grasping CNN structure to model the relationship between depth images and grasping positions. This approach was also adopted by Kumra et al. [6], who improved model performance by continuously stacking residual blocks. However, the proposed model lacks sufficient emphasis on extracting grasp-related features, resulting in suboptimal grasping performance.

In recent years, several studies have incorporated attention mechanisms into grasp detection networks, and their experiments have confirmed that attention mechanisms can enhance feature extraction efficiency and improve network performance. S. Wang et al. [10] integrated the self-attention mechanism of Transformer into a grasp model to learn global features. However, this approach may lead to the loss of some local features. The works [11,12] employed Squeeze-and-Excitation attention [13] in grasp detection networks to reweight the importance of channels. However, this method overlooks spatially dependent features, thereby limiting the performance of grasp models. Zhou et al. [14] incorporated Coordinate Attention [15] into a grasp model for capturing position-related features; however, their method ignores mutual features across the entire spatial position, and the convolution is not conducive to cross-channel interactions. Furthermore, many grasp detection networks rely on encoders and decoders. While the encoding process yields advanced grasp semantic features, the gradual reduction in resolution during encoding results in the loss of spatial features. It is noteworthy that the spatial features of objects are closely related to grasp poses, and multi-scale features can help the learning system gain a comprehensive understanding of the object's appearance and structure, which significantly impacts robot grasping quality.

To overcome these problems, we developed the Feature-Augmented Grasp Detection Network (FAGD-Net) based on encoder–decoder architecture. Within this network, we introduced two modules to enhance its focus on grasp-related features. Firstly, we proposed a Residual Efficient Multi-Scale Attention (Res-EMA) module, aimed at effectively regulating the significance of feature channels while preserving precise spatial information. Additionally, we introduced a Feature Fusion Pyramidal Module (FFPM) as an intermediary between the encoder and decoder, effectively addressing potential oversights or losses of grasp-related features as the encoder network deepens. Summarizing, the main contributions of this paper are as follows:

(1) A new module termed Residual Efficient Multi-Scale Attention (Res-EMA) was proposed to adjust the importance of the feature channels and ensure the preservation of accurate grasping space information within the channels.
(2) A Feature Fusion Pyramid Module (FFPM) was constructed between the encoder and decoder. This module aims to enhance the grasp-related features that may otherwise be overlooked or lost during the encoding process.
(3) A Feature-Augmented Grasp Detection Network (FAGD-Net) was developed based on the Res-EMA and FFPM network frameworks, which is used for the real-time prediction of optimal grasp configuration. Experimental evaluations on the Cornell dataset and Jacquard dataset demonstrate exceptional performance, where the proposed method achieved high accuracy rates of 98.9% and 96.5%, respectively.
(4) We integrated the network into an actual robot grasping system and conducted real-world grasping tests to validating the performance of our approach. It achieves advanced performance in both public datasets and real-world grasping tasks.

2. Related Works

2.1. Deep Learning for Robot Grasping

In recent years, deep learning has gained prominence in the field of robot grasping because of its superior feature extraction and generalization capabilities. Deep learning

methods for robot grasping can be broadly categorized into regression-based, classification-based, and detection-based approaches.

Regression-based approaches resemble object detection bounding box regression and have demonstrated effectiveness, but they often involve a two-stage process, leading to slower detection speeds. For instance, ROI-GD [8] utilizes features within the Region of Interest (ROI) for grasp detection, initially extracting ROIs from a scene and subsequently employing a grasp detector based on the extracted ROI features. Inspired by Fast-RCNN, Chu et al. [16] employed a grasp proposal network to identify grasping regions and then used a filter to determine the optimal grasping pose. Similarly adopting a two-stage approach, SISG-Net [17] incorporates a semantic segmentation capability, albeit at the expense of increased computational costs. Suwoyo et al. [18] integrated YOLO with grasp detection, but this method is limited to grasping single objects and cannot handle grasping in cluttered environments. Additionally, Liu et al. [19] applied Mask-RCNN for grasp feature extraction and introduced Y-Net for angle estimation and Q-Net for quality evaluation.

Classification-based approaches, pioneered by Lenz et al. [9], treat grasping as a classification problem. It is noteworthy that SAE [9] is the first approach that employed deep learning methods to address robot grasping problems, avoiding the laborious process of manual feature engineering. J. Mahler et al. [5] developed a model known as the Grasp Quality Convolutional Neural Network (GQ-CNN) for the classification of stable grasps in depth images. Recently, Zhang et al. [20] developed an angle-matching strategy based on the oriented anchor box mechanism.

Detection-based approaches treat grasping as a pixel-level grasp configuration detection problem, as illustrated by Morrison et al. [4]. Morrison et al. [4] pioneered an approach that involves regression to determine grasp configurations by predicting heatmaps for grasp confidence, angles, and widths. Other studies, such as S. Kumra et al. [6], introduced a residual convolutional neural network for the task of antipodal robot grasping, achieving high accuracies of 97.7% on the Cornell dataset and 94.6% on the Jacquard dataset. H. Cao et al. [21] introduced a Gaussian-based grasp representation method based on a lightweight generative structure network. This method achieved accuracies of 97.8% on the Cornell dataset and 95.6% on the Jacquard dataset. Teng et al. [7] utilized depth-wise separable convolutions to make a grasping network more lightweight. Fu et al. [22] also utilized depth-wise separable convolution to create a lightweight network. Tian et al. [23] employed an RGBD dense fusion approach, but the grasping performance was mediocre.

2.2. Attention Mechanism and Feature Fusion Methods

To enhance grasp-related features, various approaches have been adopted, including attention mechanisms and feature fusion. By integrating these techniques, models can improve their modeling capabilities, adaptability, and generalization to input data, thus enabling more effective handling of complex input scenarios.

In recent years, some methods have begun to utilize attention mechanisms to enhance grasping performance. While Transformer's self-attention mechanism [10] was effective in learning global features, it suffered from local feature loss. Squeeze-and-Excitation attention [11,12] yielded effective channel characteristics but ignored spatial information. The latest AAGDN [14] employed Coordinate Attention [15], capturing position-related features but overlooking mutual features across spatial positions, while the 1×1 convolution was not conducive to cross-channel interactions. To address these limitations, we developed a novel Residual Efficient Multi-Scale Attention (Res-EMA) module.

The approach of attention mechanisms has also been widely applied in other fields. For example, one study [24] utilized attention mechanisms in standard convolutions to highlight important features for better detection of the dynamic behavior of melt pools in metal additive manufacturing (AM) processes. Another study [25] employed channel attention mechanisms in models to identify rock thin sections. Additionally, in [26], atten-

tion mechanisms were incorporated into the YOLOv5 model for improved detection of tomato viruses.

Feature fusion, a technique in deep learning, involves merging feature information from different levels or branches. In encoder–decoder networks, low-level features provide spatial details, and high-level features offer semantic richness. However, down-sampling during the encoder process can result in spatial resolution loss. Feature fusion methods, such as those in [27,28], address this by combining low-level and high-level features. Inspired by Deeplab [29], we developed a feature fusion module with a smaller dilation rate to mitigate information loss during encoder subsampling, thereby enhancing the network's ability to capture grasp-related features.

3. Methods

3.1. Grasp Configuration

To facilitate the representation of robot grasp detection, we abstract the grasp into the determination of a grasp point and a line segment, which offers a more concise and efficient representation compared with the use of grasp rectangle [4,6]. Specifically, in the image space, the grasp configuration is detected from a depth image $D = \mathbb{R}^{h \times w}$, which can be described as:

$$G_i = (x_i, y_i, \theta_i, W_i, Q) \tag{1}$$

where (x_i, y_i) is the grasp center, θ_i is the rotation angle in image coordinates, with a range of $[-\frac{\pi}{2}, \frac{\pi}{2}]$, and W_i is the required grasp width in the image, within a range of $[0, W_{\max}]$. Q is the grasp confidence corresponding to each pixel, with values between 0 and 1. The goal of robot grasping is to infer the grasp configuration G_i^* with the maximum confidence from the detected grasp configuration, which can be described as:

$$G_i^* = \text{argmax}_Q\{G_i\} \tag{2}$$

In the robot's 3D workspace, a grasp pose can be described as:

$$G_\text{r} = (\mathbf{P}_\text{r}, \theta_r, W_r, Q) \tag{3}$$

where $\mathbf{P}_\text{r} = (x, y, z)$ represents the center position of the gripper tip, z is derived from the depth value of predicted grasp point by coordinate transformation [30], θ_r denotes the gripper orientation around the Z-axis, W_r denotes the closing width of the robot gripper, and Q is the grasp quality score used to predict the probability of successful grasping.

To effectively perform robot grasping tasks in real-world scenarios, it is essential to translate the grasp configurations from the image space to the robot workspace. We utilize the hand–eye calibration method [30] to compute the transformation matrix from the camera coordinate system to the robot base coordinate system. This transformation is shown in Figure 1 and described in Equation (4) [6].

$$G_r = T_{cr}(T_{ic}(G_i^*)) \tag{4}$$

where T_{ic} refers to converting the grasp configuration from the image coordinate system to the camera's 3D coordinate system, while T_{cr} refers to converting the camera's 3D coordinate system to the robot's world coordinate system.

In the experiment, within the robot's workspace, the robot approached the object based on the predicted grasp position $\mathbf{P}_\text{r} = (x, y, z)$, adjusted the gripper orientation according to the predicted angle θ_r, and closed the gripper to grasp the object based on the predicted width W_r.

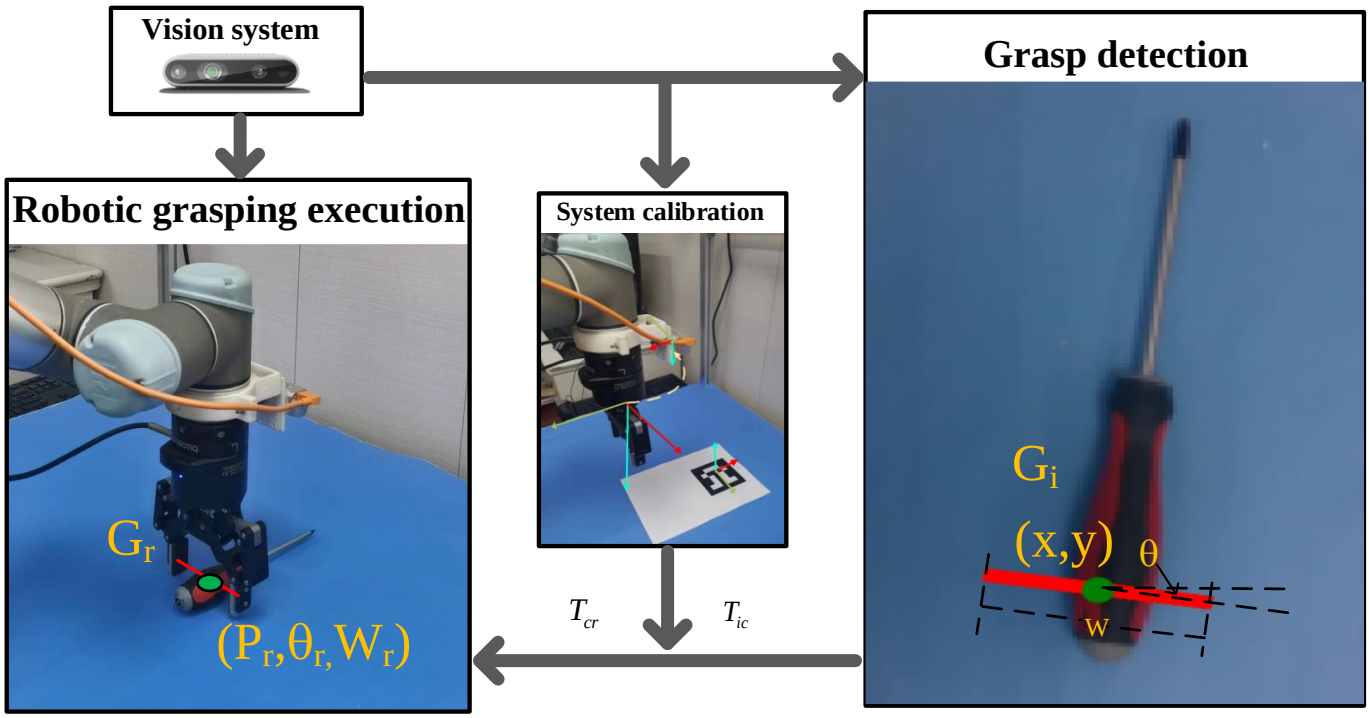


Figure 1. The transformation of grasp configurations from the image space to the robot workspace.

3.2. The Proposed FAGD-Net for Grasping Detection

The architecture of the proposed grasping network, FAGD-Net, is depicted in Figure 2, the primary goal in designing of the FAGD-Net framework is to equip the grasping system with robust feature extraction capabilities to ensure high accuracy in grasp detection while keeping its architecture lightweight. FAGD-Net employs an encoder–decoder architecture to enhance contextual understanding from input data.

Initially, a depth image undergoes feature extraction through a backbone network, which comprises only four down-sampling modules to ensure model lightweightness. Each down-sampling module consists of convolutional layers, batch normalization, and ReLU activation functions. The first down-sampling module utilizes a 9×9 convolutional kernel, while subsequent modules use a 4×4 kernel. These larger convolutional kernels offer broad receptive fields and rich feature representations. The size of the output feature maps is halved after each down-sampling module. Subsequently, features are further processed through the Residual Efficient Multi-Scale Attention (Res-EMA) module, which adjusts the importance of feature channels to highlight effective features while preserving precise spatial information within channels. In the Res-EMA module shown in Figure 2, "g" denotes the number of groups along the feature channel dimension after passing through conv block 2, "X avg pool" represents global pooling along the horizontal direction, and "Y Avg Pool" indicates global pooling along the vertical direction. Further details are provided in the subsequent sections.

Following the encoder stage, we introduce the Feature Fusion Pyramidal Module (FFPM) between the encoder and decoder to enhance information utilization without compromising the receptive field. This effectively addresses potential oversights in grasp features during encoder down-sampling.

Finally, the up-sampling module is utilized to expand and restore image dimensions. It consists of transposed convolutional layers, batch normalization, and ReLU activation functions, with the transposed convolutional kernel size set to 4×4. Ultimately, the final grasp configuration is obtained through transposed convolution, completing the network's comprehensive grasp detection process, as described by Equations (5) and (6).

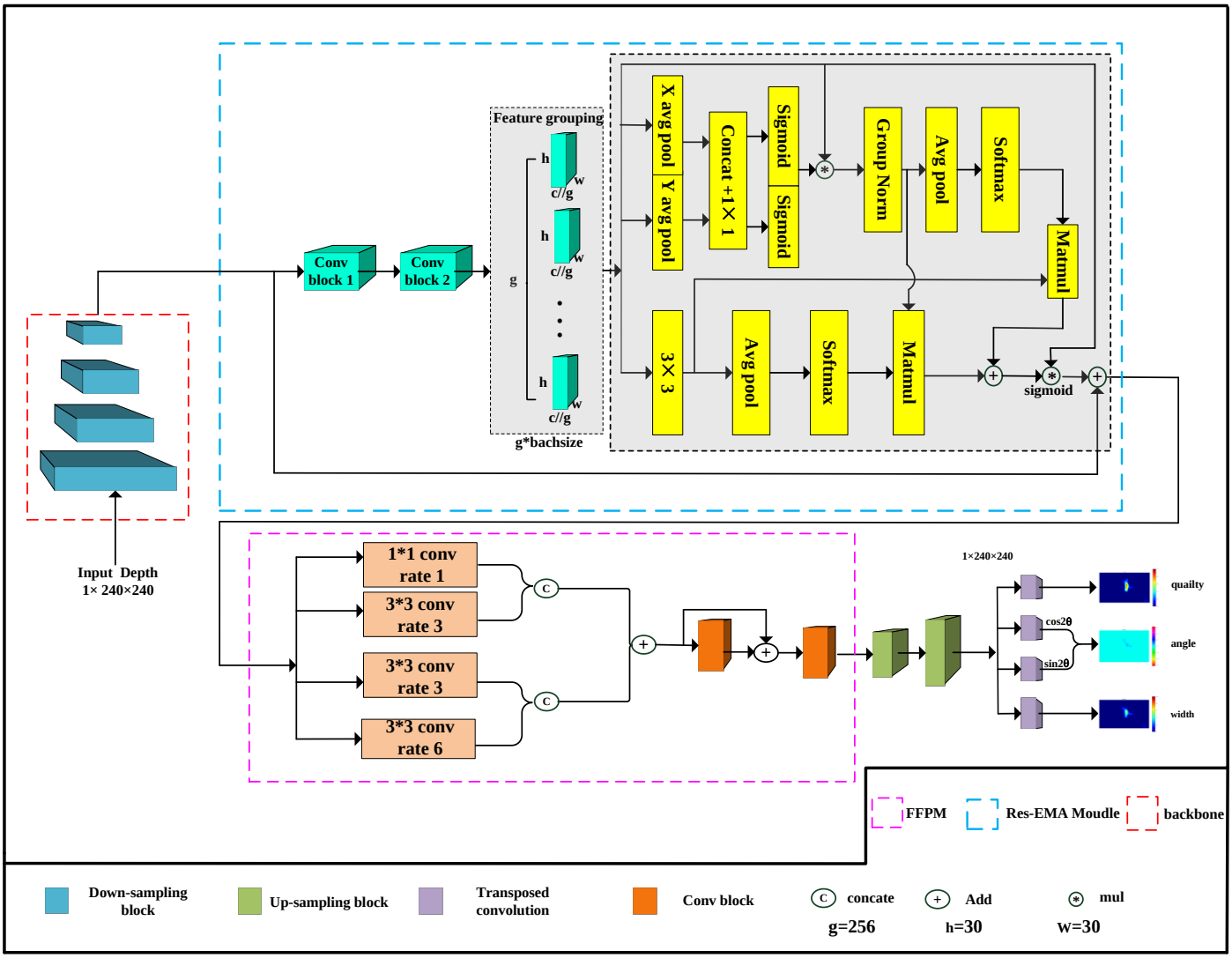


Figure 2. The structure of the Feature-Augmented Grasp Detection Network (FAGD-Net).

$$\begin{cases} F_1 = f_{backbone}\left(I_{depth}\right) \\ F_2 = f_{\text{Res}-EMA}(F_1) \\ F_3 = f_{FFPM}(F_2) \\ F_4 = f_{Up-sampling1}(F_3) \\ F_5 = f_{Up-sampling2}(F_4) \end{cases} \tag{5}$$

where F_1, F_2, F_3, F_4, and F_5 represent the feature maps after passing through the backbone ($f_{backbone}$), Res-EMA module ($f_{\text{Res}-EMA}$), Feature Fusion Pyramidal Module (f_{FFPM}), the first up-sampling block ($f_{Up-sampling1}$), and the second up-sampling block ($f_{Up-sampling2}$), respectively.

$$\begin{cases} qualityMap = f_{trans\_conv}(F_5) \\ AngleMap_{\sin} = f_{trans\_conv}(F_5) \\ AngleMap_{\cos} = f_{trans\_conv}(F_5) \\ WidthMap = f_{trans\_conv}(F_5) \end{cases} \tag{6}$$

where f_{trans_conv} represents the transposed convolution and F_5 denotes the feature map output by the up-sampling block. Ultimately, the network generates feature maps of grasp quality, grasp angle, and grasp width.

It is worth noting that we do not directly output the final grasping angle θ. Instead, we separately output $\sin(2\theta)$ and $\cos(2\theta)$ to eliminate discontinuities near $\pm\frac{\pi}{2}$ and maintain

a unique mapping between θ and the interval $\left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$. Finally, the grasping angle is generated using Equation (7).

$$\theta = \arctan\left(\frac{\sin(2\theta)}{\cos(2\theta)}\right)/2. \tag{7}$$

4. The Components of FAGD-Net

4.1. Res-EMA Module

Existing works claim that attention mechanisms play a crucial role in improving the expressive capability of grasping networks, as evidenced by studies such as [10–12,14]. However, there is still potential for optimization in these attention mechanisms. To bridge this gap, we introduce advanced Multi-Scale Attention (EMA) [31] into the realm of grasping, thereby introducing a novel Res-EMA module.

Figure 3 illustrates the structure of Res-EMA. The first part comprises the residual block structure [32], and the second part is EMA. Through skip connections, the features outputted by the backbone are aggregated with those from the attention module, aiding the network in avoiding the issues of gradient vanishing and exploding. Within the residual block structure, there are two convolution modules. Convblock1 includes a 3×3 convolution, batch normalization, and Rectified Linear Unit (ReLU) activation. The second convolution module repeats the operations of the first convolution module, followed by a 3×3 convolution and batch normalization.

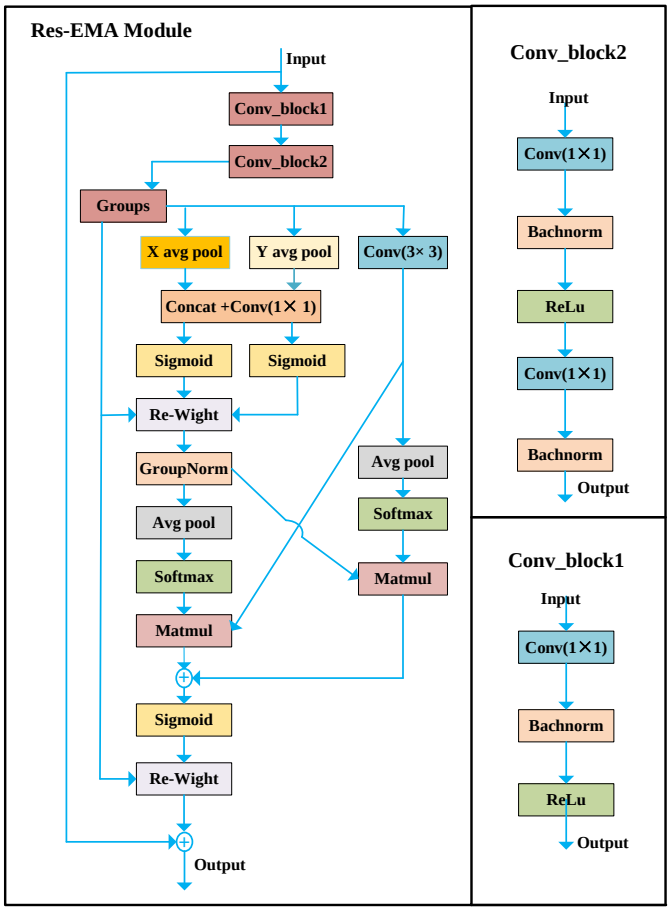


Figure 3. The Res-EMA module.

The second part consists of Efficient Multi-Scale Attention (EMA). Firstly, the features after convblock2, denoted as $F \in \mathbb{R}^{C \times H \times W}$, are divided into G sub-features in the channel dimension. This grouping aims to learn distinct grasp semantic information, thereby

enhancing the model's expressive capacity. The grouping process can be described by Equation (8).

$$F = [F_0, F_i, \ldots, F_{G-1}], F_i \in \mathbb{R}^{C//G \times H \times W} \tag{8}$$

Subsequently, similar to the operations in Coordinate Attention [15], given the input features, 1D global average-pooling is applied along the horizontal and vertical directions separately to encode spatial position information. Through this operation, positional information is encoded into the feature map to enhance sensitivity to spatial positions, highlighting feasible grasping regions. The 1D global average-pooling along the horizontal direction, denoted for the c-th channel over the spatial dimension H, can be described by Equation (9).

$$z_c^H(H) = \frac{1}{W} \sum_{0 \le i \le W} f_c(H, i) \tag{9}$$

Similarly, the 1D global average pooling along the vertical axis W for the c-th channel can be described by Equation (10):

$$z_c^W(W) = \frac{1}{H} \sum_{0 \le j \le H} f_c(j, W) \tag{10}$$

where c represents the number of input channels and H and W denote the spatial dimensions of the input features. The aforementioned operations are consistent with Coordinate Attention [15]. The shared components with Coordinate Attention are named the 1×1 branch, while the branch with the 3×3 kernel is named the 3×3 branch.

After feature encoding along the two image height directions, a shared 1×1 convolution is employed. Following the decomposition of the output of the 1×1 convolution into two vectors, the sigmoid function is utilized to fit a 2D binomial distribution on the linear convolution. Feature interaction across channels is achieved by using multiplication to aggregate two channel-wise attention maps within each group, facilitating comprehensive feature representation. The 3×3 branch employs a 3×3 convolutional operation to capture inter-channel feature interactions, thereby expanding the feature space.

The output of the 1×1 branch is aggregated with the output of the 3×3 branch across different spatial dimensions, enabling a more comprehensive aggregation of grasp-related features. Following the output of the 1×1 branch, 2D global average pooling is applied to encode global spatial features and model long-range dependencies. The 2D global average pooling method can be described as Equation (11)

$$z_c = \frac{1}{H \times W} \sum_{j}^{H} \sum_{i}^{W} f_c(i, j) \tag{11}$$

After 2D global average pooling, a Softmax function is applied to fit a linear transformation. Through matrix multiplication, a spatial attention map is obtained, collecting spatial information features at different scales. Similarly, in the 3×3 branch, 2D global average pooling is used to encode global spatial information. Another Softmax and matrix multiplication produce a second attention map, preserving precise spatial position information. Finally, the two sets of weights are aggregated, and a sigmoid function is employed to highlight the global context for all pixels.

4.2. Feature Fusion Pyramidal Module

The encoder progressively deepens the network, extracting higher-level semantic information. However, this intensification accompanies a reduction in feature resolution, potentially neglecting and gradually losing certain spatial features. Consequently, crucial grasp-related information may be compromised during the up-sampling stage, limiting the overall quality of grasp detection. To tackle this challenge, Liu et al. [29] effectively reduced feature information loss by employing spatial pyramid pooling. Nevertheless, a

large dilation rate (rate = 6, 12, 18) in this context may result in suboptimal information utilization. To overcome these concerns, we introduce the Feature-Augmented Pyramid Module (FAPM), strategically designed with a small dilation rate. The FAPM acts as a crucial bridge between the encoder and decoder, effectively mitigating the aforementioned issues and contributing to improved grasp detection accuracy.

As illustrated in Figure 4, the Feature Fusion Pyramid Module (FFPM) incorporates a multi-scale feature extraction process. Initially, the features are input into four parallel feature pyramids, each utilizing atrous convolution with different dilation rates of 1, 3, 3, and 6, respectively. This unique structure enhances information utilization while maintaining a broad receptive field. To enable comprehensive feature interaction, the extracted features from the pyramids are concatenated and added together. This process ensures a holistic representation of features across different scales. Subsequently, feature extraction is finalized through a convolution module coupled with a skip connection. This convolution module comprises a 3×3 convolution operation, batch processing, and Rectified Linear Unit (ReLU) nonlinear activation. The intricate design of the FFPM promotes effective multi-scale feature extraction, fostering enhanced information integration within the grasping network.

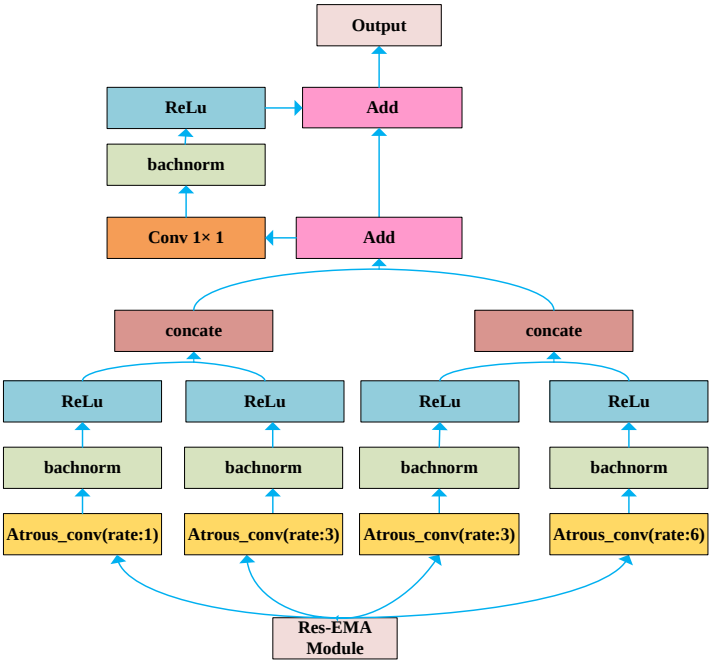


Figure 4. The feature Fusion Pyramidal Module.

4.3. Loss Function

In this paper, we treat the grasp detection problem as a regression task, framing it as the optimization of the minimum error between the predicted grasp and its corresponding label. To achieves this, we employ the L1 loss function, which has the inherent advantage of diminished sensitivity to outliers, encompassing instances that are notably large or small. This characteristic renders the L1 loss function robust and stable for our training model. The mathematical representation of the L1 loss function is denoted as Equation (12).

$$L1\ \text{loss} = \begin{cases} 0.5x^2, & \text{if } |x| < 1 \\ |x| - 0.5 & \text{otherwise} \end{cases} \tag{12}$$

In each prediction task, the loss for grasping position prediction can be expressed as Equation (13).

$$L_p = -\frac{1}{N}\sum_{i}^{N} L1\ \text{loss}\left(p_i - \hat{p_i}\right) \tag{13}$$

where p_i and $\hat{p}_i$ represent the ground truth and predicted values of the grasp quality score, respectively.

The loss for prediction of the grasping angle can be expressed as Equations (14) and (15).

$$L_{\cos 2\theta} = -\frac{1}{N}\sum_{i}^{N} L1\ \text{loss}\left(\cos 2\theta_i - \cos 2\hat{\theta}_i\right) \tag{14}$$

$$L_{\sin 2\theta} = -\frac{1}{N}\sum_{i}^{N} L1\ \text{loss}\left(\sin 2\theta_i - \sin 2\hat{\theta}_i\right) \tag{15}$$

where $\cos 2\theta$ and $\cos 2\hat{\theta}_i$ represent the ground truth and predicted values of the grasp angle score, respectively. Similarly, $\sin 2\theta_i$ and $\sin 2\hat{\theta}_i$ are also the ground truth and predicted values of the grasp angle score.

While the loss for the predicted grasp width can be expressed as Equation (16).

$$L_w = -\frac{1}{N}\sum_{i}^{N} L1\ \text{loss}\left(w_i - \hat{w}_i\right) \tag{16}$$

where w_i and $\hat{w}_i$ represent the ground truth and predicted values of the grasp width score, respectively.

Thus, the total loss for grasping detection can be expressed as the function in Equation (17):

$$L_{total} = L_p + L_w + L_{\cos 2\theta} + L_{\sin 2\theta} \tag{17}$$

5. Experimental Setup and Materials

5.1. Experiment Platform

The robotic grasping platform is depicted in Figure 5. The system comprises a UR5 robotic arm, a Realsense D435i camera, and a parallel gripper (Robotiq-2F85). The RealSense D435i is mounted at the end of the robotic arm, and the camera is connected to the computer. Additionally, the UR5 robotic arm is connected to the controller along with the parallel gripper (Robotiq-2F85), and they are connected to the PC via an Ethernet cable. Communication between the PC and the controller is achieved through sockets. We use the hand–eye calibration method [30] to obtain the transformation matrix and the camera's intrinsic parameters; the process is achievable within the Robot Operating System (ROS). PyTorch serves as the deep learning framework for both model training and prediction, which runs on an NVIDIA RTX3090 GPU (Nvidia, Santa Clara, CA, USA) with Ubuntu 18.04. For optimization, we employed the popular Adam optimizer with a learning rate set at 0.001. Additionally, we adopted a data split ratio of 9:1 for training and testing purposes.

We randomly selected a variety of unknown objects for grasping tests, which included household items and adversarial objects. The household items encompassed 16 different categories, such as mice, adhesive tapes, charging cables, screwdrivers, fruit models, medications, toiletries, beverages, etc. These items exhibited diverse colors, shapes, and sizes. To comprehensively validate the robot's grasping generalization capability, we randomly selected adversarial objects characterized by abstract geometric properties and uncertain surface features. The test objects are shown in Figure 5. In real-world grasping experiments, objects are randomly placed on a tabletop to create experimental conditions of single-object, multi-object, and cluttered scenes. The main challenge lies in the uncertainty in object shapes, sizes, placements, and quantities. Therefore, the model trained on the Jacquard dataset was used to infer grasp configurations in these scenes.

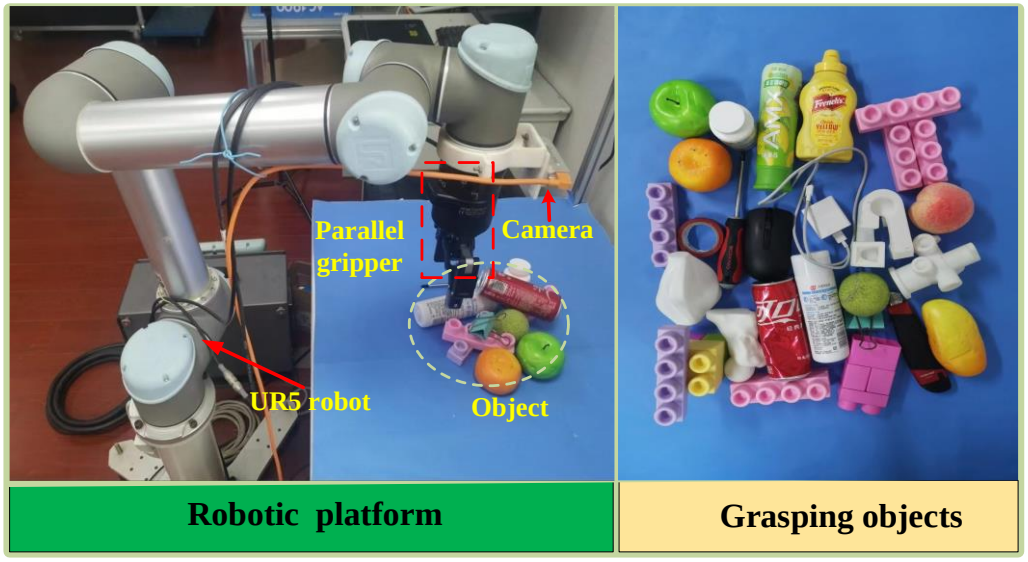


Figure 5. Robotic platform and grasping objects.

Specifically, the robot moved to an initial position approximately 0.5 m away from the tabletop. The camera captured a depth image, which was then input into FAGD-Net. The model generated a best grasp configuration, which included grasp position, angle, and width. Then, the grasp configuration with the highest confidence was transformed into the robot coordinate system. Finally, the robot adjusted the gripper pose based on the predicted grasp position and angle, approached the target, and finally closed the gripper to grasp the object based on the predicted width. A successful grasp was recorded when the object was grasped and lifted at least 10 cm.

5.2. Dataset and Processing

The Jacquard Grasping Dataset [33] is a comprehensive resource in the field of robotic grasping, offering rich grasp-related information. It comprises 54,000 RGB-D images with a resolution of 1024 × 1024 pixels. The dataset includes 1,181,330 unique grasp annotations for training purposes and requires no pre-processing.

The Cornell Grasping Dataset [34] is a renowned dataset in robotic grasping. It comprises 1035 RGB-D images capturing 240 distinct real-world objects, with images at a resolution of 640 × 480 pixels. Among these, there are 5110 positive grasps and 2909 negative grasps annotated within the dataset.

However, the Cornell Grasping Dataset is relatively small in scale. To address this limitation, we pre-processed the data by augmenting it with random crops, zooms, and rotations.

5.3. Evaluation Index

The evaluation metrics utilized in this study are widely recognized and have been extensively applied to both the Cornell and Jacquard datasets. Our evaluation criteria consist of two conditions as expressed in Equation (18). Firstly, the Intersection over Union (IoU) between the labeled grasp and the predicted grasp must be less than 25%. Additionally, the angle deviation between the grasping direction of the predicted grasp and the actual label should not exceed 30°.

$$\begin{cases} \mathrm{A}\left(\mathrm{A}_p, \mathrm{A}_\mathrm{L}\right) < 30^\circ \\ \frac{\left|\mathrm{G}_p \cap \mathrm{G}_\mathrm{L}\right|}{\mathrm{G}_p \cup \mathrm{G}_\mathrm{L}} > 0.25 \end{cases} \tag{18}$$

where A_p is the grasping prediction angle, A_L is the grasping label angle, G_p is the grasping prediction, G_L is the grasping label, $\mathrm{G}_p \cap \mathrm{G}_\mathrm{L}$ represents the intersection of the grasping

prediction and the grasping label, and $G_p \cup G_L$ represents the union of the grasping prediction and the grasping label.

6. Results and Analysis

6.1. Grasp Detection in the Jacquard Dataset

In this task, the Jacquard was used, with 90% of the images used for training and the remaining 10% for grasping testing. We resized the images to 240 × 240, and during network training, the batch size was set to 8.

In Figure 6, we visualize the detection results on the Jacquard dataset, which include representations of robot grasping points and grasping line segments, along with a heatmap indicating the confidence of grasping for each pixel and the corresponding grasping angles and widths. In experimental testing, our model achieves an accuracy of 96.5% on the Jacquard dataset. It is obvious that the grasp configuration predicted by the model is easily grasped by the robot. Additionally, our model is lightweight with approximately 260 K parameters compared with the 1.8 million parameters of GR-CNN [6]. This is because GR-CNN repetitively stacks residual blocks in the middle of the network to enhance the model's ability to extract grasp-related features, but this approach comes at the expense of efficiency and parameter count.

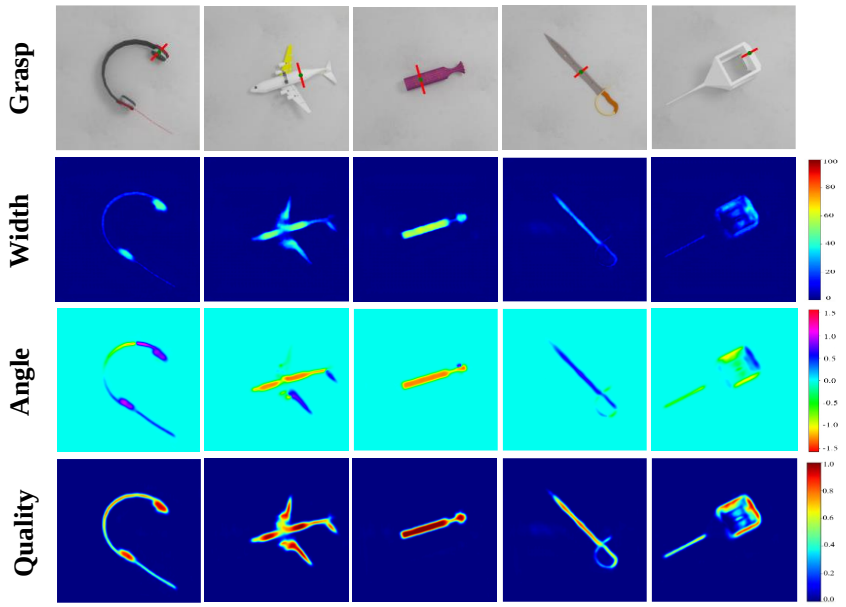


Figure 6. Grasping detection results in the Jacquard dataset.

As recorded in Table 1, our method outperforms works [4,6,33] by 12.5%, 1.9%, and 22.3%, respectively. These methods have relatively simple network architectures and do not emphasize the extraction of grasp-related features sufficiently. For instance, GG-CNN [4] only uses ordinary convolutions for down-sampling features and then up-samples them to output grasp configurations. Furthermore, our model outperforms the latest QQGNN [22] by 1.5%. Although QQGNN employs depth-wise separable convolution to make the grasping model lightweight, this approach may limit grasp-related feature learning as it only allows each channel to interact with its corresponding convolution kernel during the depth-wise convolution stage while neglecting inter-channel correlations.

Table 1. Comparison of results in the Jacquard dataset.

Authors	Method	Year	Accuracy (%)
Depierre et al. [33]	Jacquard	2018	74.2
Zhou et al. [35]	FCGN	2018	91.8
Morrison et al. [4]	GG-CNN	2020	84.0
Kumra et al. [6]	GR-CNN	2020	94.6
Cao et al. [12]	RSE-Net	2021	94.8
Ainetter et al. [36]	Det_Seg_Refine	2021	92.5
Liu, D. et al. [19]	Cascaded Net	2021	92.1
Wang et al. [10]	TF-Grasp	2022	94.6
Yu et al. [11]	SE-Res-Unet	2022	95.7
Cao et al. [21]	Efficient-Grasp	2022	95.6
Zhou Z et al. [14]	AAGDN	2023	96.2
Fu et al. [22]	QQCNN	2024	95.0
Ours	FAGD-Net	2024	96.5

Some other methods adopt attention mechanisms to efficiently extract relevant grasp features. Our model outperforms the methods in [10–12,14], by 1.9%, 0.8%, 1.7%, and 0.3%, respectively. The work [10] uses the self-attention mechanism of transformers to obtain global features. However, this approach may inadvertently discard certain local features, thus resulting in suboptimal performance. On the other hand, the works [11,12] employ the Squeeze-and-Excitation (SE) attention mechanism. While it is effective in isolating important channel features, it overlooks spatial information within the network and lacks the ability to handle multi-scale features. Although AAGDN [14] focuses on spatial position information, it neglects the importance of interactions among spatial positions as a whole. Moreover, the limited receptive field of 1×1 convolution kernels hinders the modeling and utilization of local cross-channel interactions and contextual information. In contrast, our network addresses these issues through the Res-EMA module, which effectively adjusts the importance of feature channels while preserving precise spatial information within the channels. As a result, our grasp detection network achieves an accuracy of 96.5% on the Jacquard Grasping Dataset.

6.2. Grasp Detection in the Cornell Dataset

To mitigate potential overfitting during model training in Cornell dataset, we integrated Dropout regularization and employed the Adam optimizer with a learning rate set at 0.001 while maintaining a batch size of 8.

In Figure 7, we visualize the grasping detection results, where the first row displays the detected grasp points and grasp lines. It is evident that each grasp configuration is allocated within the graspable region. While the subsequent rows show heatmaps of grasp configurations, the heatmaps indicate the grasp configuration for each pixel in the image, with higher confidence grasp positions appearing closer to red. As a result, our model achieves a grasping accuracy of 98.9%. In addition, we conducted inference speed tests in the Cornell dataset, which only required 23 milliseconds of inference time per image, thus ensuring real-time for robot applications.

In Table 2, we present a comparison of the test results among other algorithms in Cornell dataset. Compared with the well-known GG-CNN [4], our method exhibits a performance improvement of 25.9%. Our method also outperforms GR-CNN [6] by 1.2%. The main reason for the improvement lies in the other algorithms' relatively simple network architectures and limited capability in feature extraction. Finally, our algorithm surpasses the state-of-the-art algorithms [10,11] by 0.9% and 0.7%, respectively. It is evident that our model is also advanced.

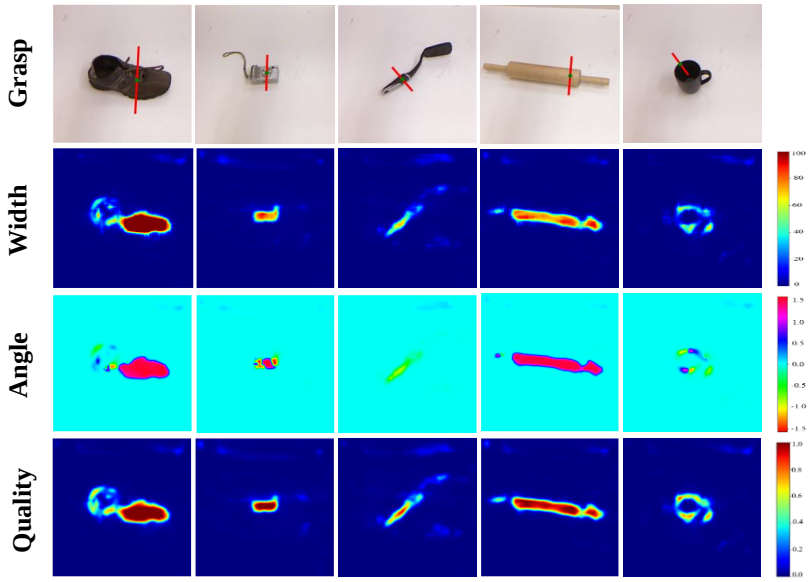


Figure 7. Grasping detection results in the Cornell dataset.

Table 2. Comparison of results in the Cornell dataset.

Method	Year	Accuracy (%)	Speed (ms)
SAE [9]	2015	73.9	1350
GG-CNN [4]	2020	73.0	19
ROI-GD [8]	2019	93.6	40
GR-CNN [6]	2020	97.7	20
Cascaded Net [19]	2021	95.2	-
RSE-Net [12]	2021	96.4	5
Det_Seg_Refine [36]	2021	98.2	63
TF-Grasp [10]	2022	98.0	42
Efficient-Grasp [21]	2022	97.8	6
SE-Res-Unet [11]	2022	98.2	25
QQCNN [22]	2024	97.7	17
Ours	2024	98.9	23

6.3. Ablation Experiment

6.3.1. FFPM and Res-EMA Ablation Tests

To assess the impact of each module on network performance, we conducted ablation studies using the Jacquard dataset, in which four sets of different tests were carried out.

In test 1, our network did not incorporate any additional modules. In test 2, we added the Feature Fusion Pyramid Module (FFPM) without including the Residual Efficient Multi-Scale Attention (Res-EMA) module. In test 3, we included the Res-EMA module without the FFPM. Finally, in test 4, both the Res-EMA module and FFPM were integrated into the network.

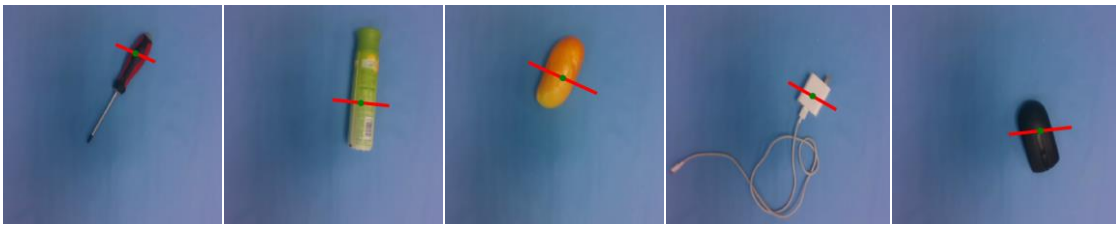
Table 3 presents the results of the model ablation experiment. Comparing test 2 with test 1, the addition of the FFPM resulted in a 1.1% increase in grasping detection, demonstrating its effectiveness in enhancing performance. Similarly, comparing test 3 with test 1, the inclusion of the Res-EMA module led to a 1.2% improvement. In test 4, where both modules were utilized simultaneously, a 2.1% increase in performance was observed compared with test 1, suggesting that the combined use of both modules yielded the best grasping detection performance.

Table 3. The results of the ablation study.

	Test 1	Test 2	Test 3	Test 4
FFPM	×	✓	×	✓
Res-EMA	×	×	✓	✓
Accuracy (%)	94.4	95.5	95.6	96.5

Figure 8 illustrates the samples of grasping detection results with or without the Res-EMA module. It can be observed that the addition of the Res-EMA module leads to the more accurate detection of grasp positions.

(a) Results With the Res-EMA Module



(b) Results Without the Res-EMA Module

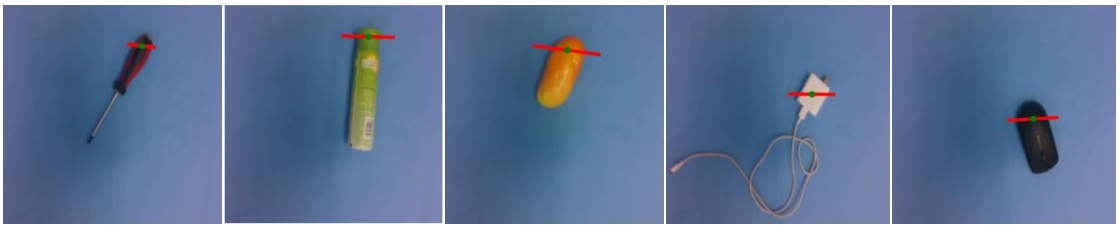


Figure 8. The Res-EMA module ablation study. (**a**) The results with the Res-EMA module. (**b**) The results without the Res-EMA module.

6.3.2. Comparison among Various Attention Mechanisms

To further validate the advantages of our proposed attention mechanism Res-EMA, we incorporated attention mechanisms such as Residual Squeeze-and-Excitation (Res-SE) and Residual Coordinate Attention (Res-CoA) [11,14] into our model in a similar manner. Additionally, we include a baseline module without any attention mechanisms, denoted as the residual block. The positions of the Res-CoA and Res-SE modules within the network are depicted in Figure 9.

The training was conducted for 50 epochs using Jacquard dataset in a consistent environment. After each epoch, we evaluated the grasping performance on approximately 5499 previously unseen images. To ensure fairness, all models were trained using identical network architectures. The experimental results are presented in Figure 10. Comparing the performance of the residual block, Res-CoA, Res-SE, and our Res-EMA modules, it is evident that only our Res-EMA module achieves a peak accuracy of 96.5%, signifying its substantial advantage. The Res-EMA module demonstrates its superior capability in focusing on grasp-related features. This module significantly enhances the detector's ability to accurately identify the optimal grasping region.

In Figure 11, the first column displays the grasping detection results with our Res-EMA module, the second column shows the results with Res-CoA module, the third column presents the results with the Res-SE module, and the last column illustrates the detection results of the baseline module with residual blocks without any attention mechanism.

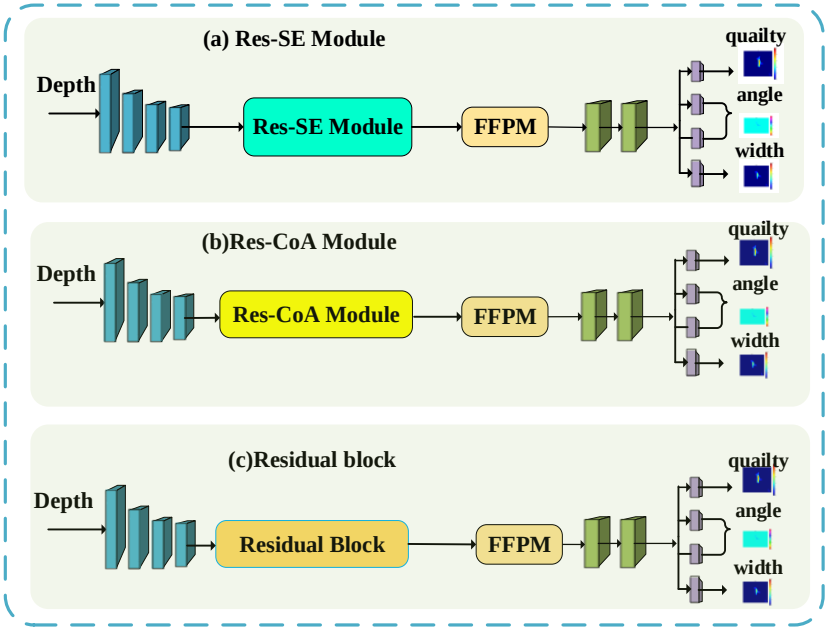


Figure 9. Different attention mechanisms.

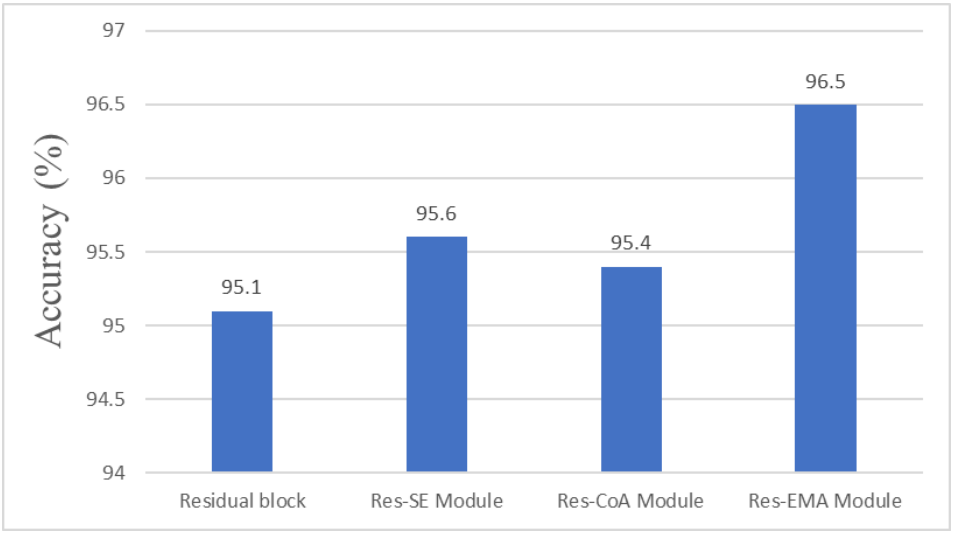


Figure 10. Experimental results of various attention mechanisms.

One can observe that the grasping produced by adding the Res-EMA module is more accurate and facilitates easier object grasping by the robot. While the other attention modules also contribute to effective grasping, their precision is comparatively lower. Among them, the model with only residual blocks performs the worst. In the last column of the fourth row in the Figure 11, representing a small clamp, the detected grasp configuration is invalid. The reason is that this model fails to extract effective grasp-related features, while the model with the Res-EMA module efficiently extracts spatial location features to ensure precise grasp positions. Moreover, extracting features at multiple scales benefits the robot in grasping objects of different sizes.

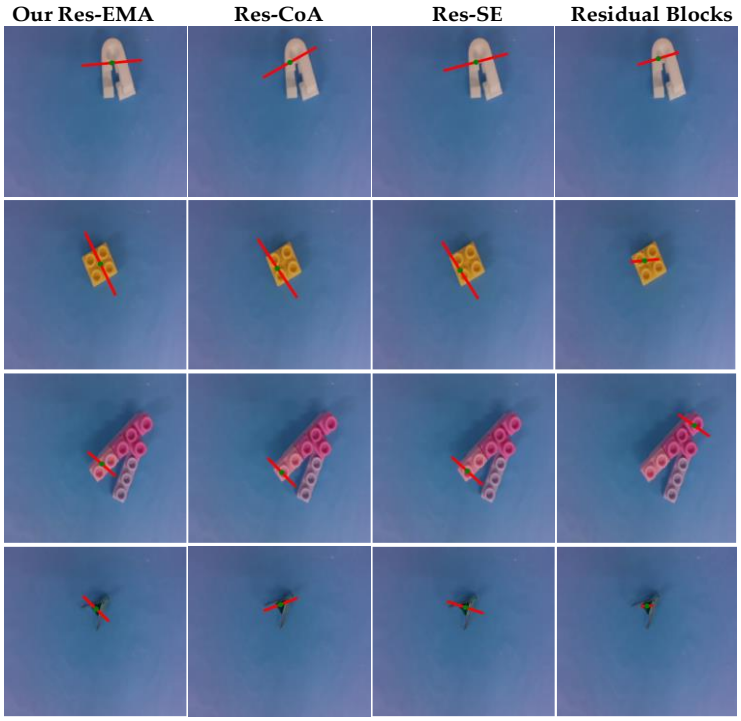


Figure 11. Grasping detection results with different attention mechanisms. The first to fourth columns depict grasping detection using our Res-EMA module; Res-CoA module; Res-SE module, and residual blocks, respectively.

6.3.3. Comparison among Different Feature Fusion Methods

We conducted experimental evaluations on the Feature Fusion Pyramidal Module (FFPM). As the encoder progressively reduces the resolution during the encoding process, certain grasp-related features may be neglected and gradually lost. This can lead to the failure to recover crucial information during decoder up-sampling, resulting in suboptimal performance. Common approaches to address this issue include feature addition and feature concatenation, where low-level features are integrated into the decoder through skip connections to compensate for relevant features disregarded during the encoding process. While feature addition and concatenation are effective, we argue that compared with the proposed FFPM, our FFPM can more efficiently preserve grasp-related features. The mechanisms of feature addition and concatenation are illustrated in Figure 12.

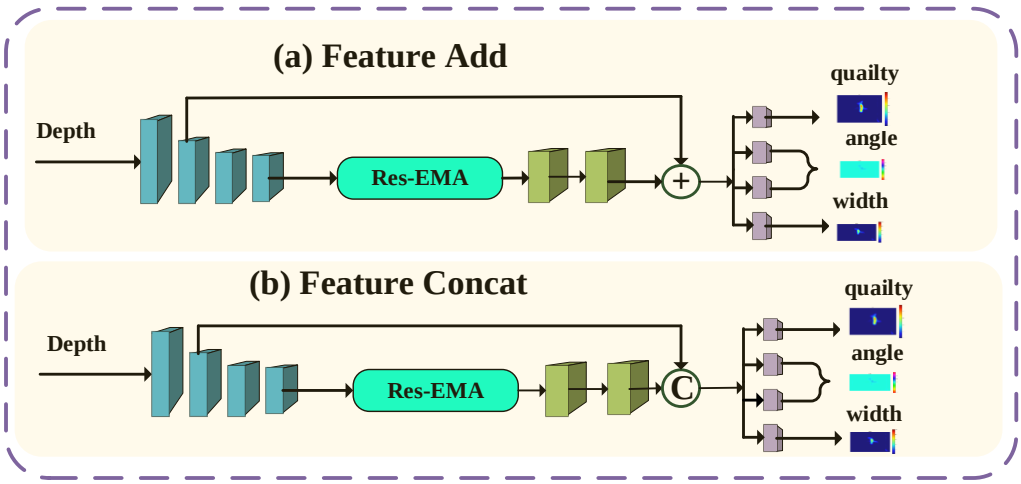


Figure 12. Different feature fusion methods.

We conducted training for 50 epochs on the Jacquard dataset, evaluating the model on the test set after each epoch. The highest accuracy achieved during evaluation is depicted in Figure 13. We compared three methods, including the FFPM, feature addition, and feature and concatenation, with the FFPM demonstrating superior performance, achieving an accuracy of 0.965. In contrast, feature addition and feature concatenation exhibited mediocre performance, indicating that the FFPM excels in preserving grasp-related features.

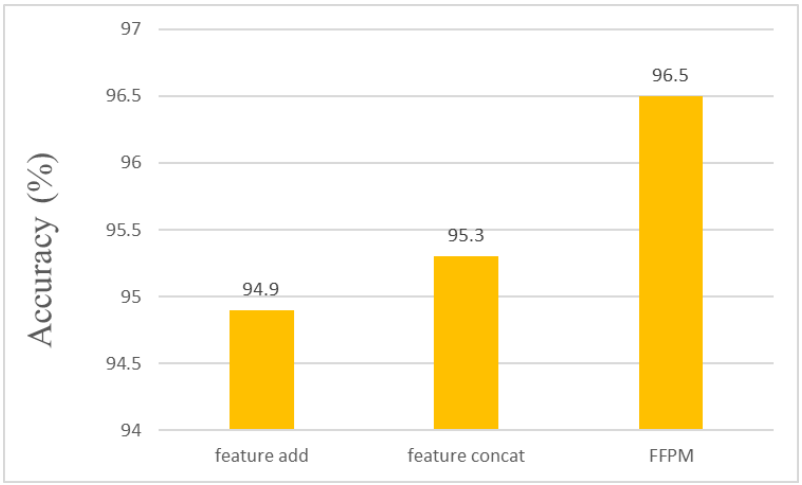


Figure 13. Experimental results of different feature fusion methods.

In Figure 14, we visualize the grasp detection results using different feature fusion methods. It is evident from the figure that the addition of the FFPM results in more precise grasp configurations. Specifically, the grasp positions generated by the FFPM closely align with the object's center of mass, facilitating more effective grasping. The key effectiveness of the FFPM lies in its role as a bridge between the encoder and decoder, effectively preserving grasp-related features and mitigating the loss of relevant grasp features caused by the deepening of the encoding process.

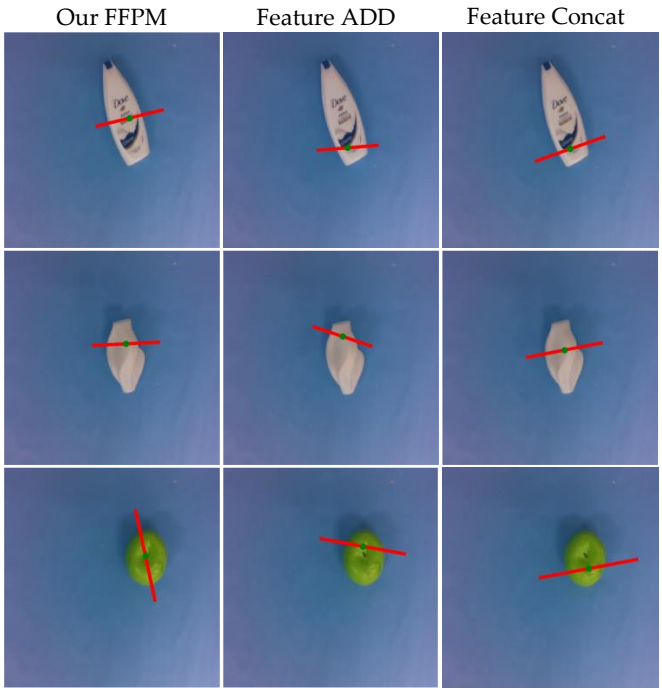


Figure 14. Grasping detection results with different feature fusion methods. The first to third columns depict grasping detection using our FFPM, feature addition, and feature concatenation, respectively.

6.4. Robotic Grasping

6.4.1. Single-Object and Multi-Object Scenes

In the robotic grasping task, we first conducted a single-object grasping test on 16 unknown household objects, varying in shape, size, and color. Each object was randomly positioned in different locations within the robot's workspace. In total, we performed 160 grasping attempts, out of which 152 were successful, resulting in a grasping success rate of 95.0%. The experimental results demonstrate the effectiveness of the proposed FAGD-Net in grasping unknown household objects. We compared the success rates of our method with those of classical and the state-of-the-art algorithms, as shown in Table 4. The comparison indicates that our proposed method also achieves advanced success rates in single-object grasping tasks.

Table 4. Results on single objects.

Work	Household Objects
	Grasping Success Rates (%)
Morrison et al. [4]	92.0
Lenz et al. [9]	89.0
F. Chu [16]	89.0
Liu et al. [19]	94.6
Tian et al. [23]	94.0
Chen et al. [37]	93.5
Li et al. [38]	92.0
Lilai et al. [39]	91.5
Ours	95.0

Furthermore, we conducted grasping tests on adversarial objects with abstract shapes and irregular surfaces. We attempted 90 grasps, out of which 82 were successful, resulting in a success rate of 91.1%. These tests demonstrate that our grasping detection model performs well even on objects with peculiar shapes. Figure 15 depicts a schematic illustration of grasping in a single-object scenario, showcasing the results of grasping detection for both household objects and adversarial objects, followed accurate grasping by the robot. Additionally, we visualize the grasp configurations as heatmaps to intuitively present the feature maps. The grasp heatmap demonstrates the model's effective perception of graspable regions in single-object scenes, with varying confidence levels of grasp configurations distributed across each pixel of the single object. Finally, the robot executes the grasp configuration with the highest confidence level. These results show that FAGD-Net is able to effectively grasp previously unseen objects in single-object scenarios, also showcasing its robust grasping performance.

To assess the effectiveness of our grasp detection model in multi-object scenarios, we randomly selected objects to construct a scene with multiple targets, in which each object was randomly placed on a tabletop. During each trial, the robot identified and executed the grasp with the highest confidence level. A total of 120 grasping attempts were conducted, resulting in 112 successful grasps and achieving an average success rate of 93.3%. The robot demonstrated high-performance grasping even in multi-object scenarios, facilitated by the model's pixel-level learning, which effectively extracts relevant grasp features and enables the model to generalize well even in the presence of multiple objects. The grasp detection results in a multi-object scenario are depicted in Figure 16, where different colors represent varying confidence levels. We visualize the feature maps of the multi-object scenario to observe the distribution of grasp configurations at each pixel. In the grasp heatmap, it can be observed that the model effectively generalizes to multi-object scenarios, perceiving grasp configurations for each pixel. Finally, the robot selects the grasp configuration with the highest confidence for execution.

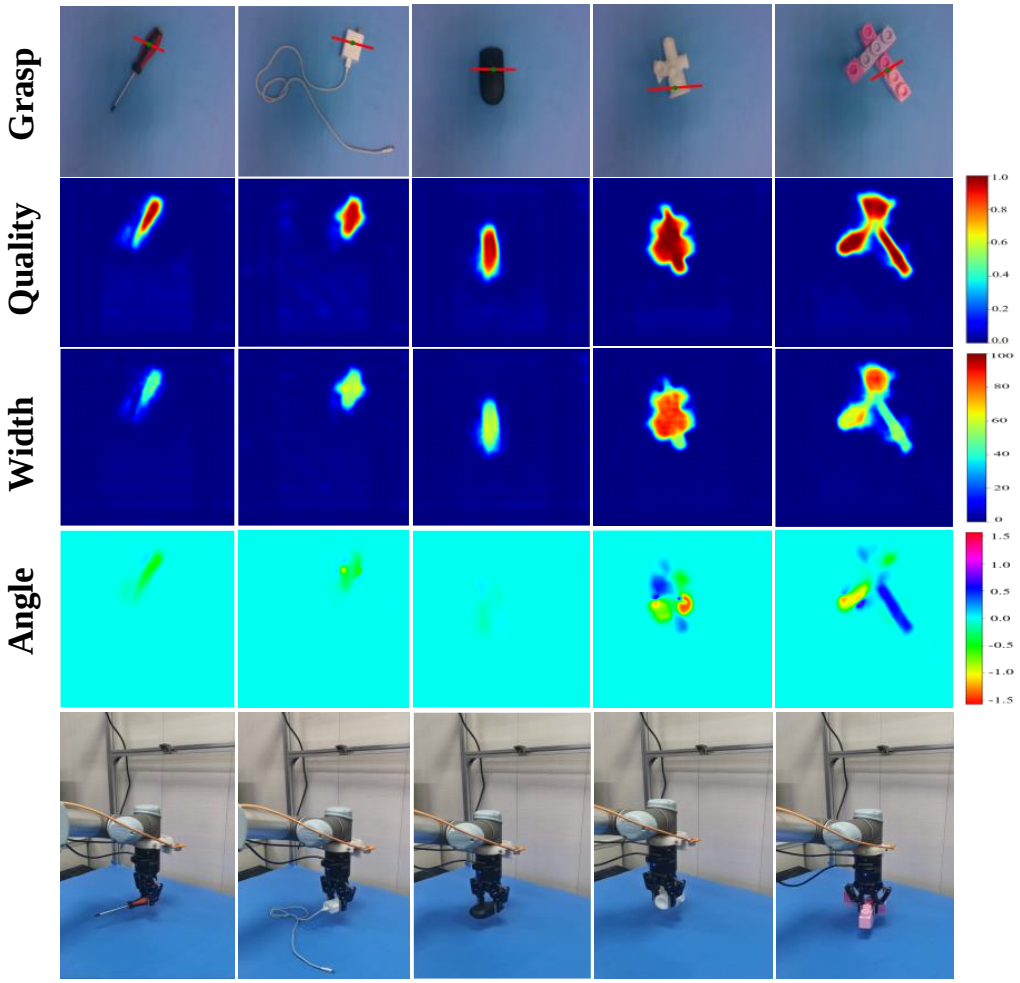


Figure 15. Single-object grasping experiment.

6.4.2. Robot Grasping in Cluttered Scenes

To further validate the effectiveness of our grasp detection algorithm in challenging cluttered environments, we conducted grasping tests by randomly placing 10 to 15 different objects to create a cluttered scenario. The robot executed grasping tasks in these cluttered environments. As shown in Figure 17, in 200 attempts at cluttered grasping, our system achieved an impressive success rate of 91% (182/200). This can be attributed to the model's pixel-level grasp configuration learning, enabling the extraction of grasp-relevant features and facilitating generalization. Thus, the model can effectively infer grasp configurations in cluttered scenes, and the robot also executes the optimal grasp based on the inference results. The experimental results demonstrate the effectiveness of our grasp detection model even when the targets are cluttered.

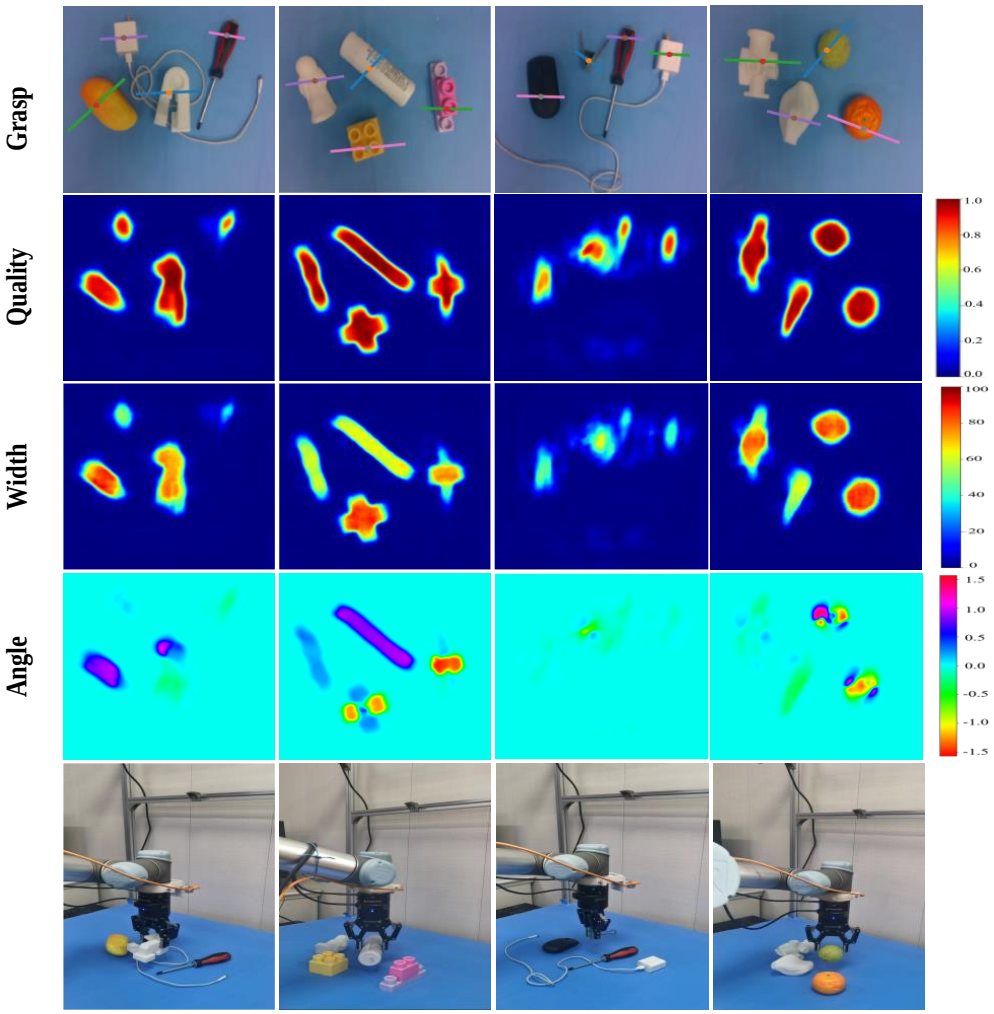


Figure 16. Grasping in multi-object scenes.

Furthermore, Table 5 presents a comparative analysis of different algorithms' performances when handling the objects in cluttered scenes. It is clear that the proposed method outperforms the existing similar methods, which effectively grasps unknown objects in cluttered settings. The main reason for the suboptimal performance of these methods lies in the insufficiency of feature extraction. The network design in [4] is overly simplistic, leading to insufficient expressive capability of the model. The researchers in [8] utilized the Region of Interest (ROI) feature extraction method from Fast R-CNN. However, the ROI extractor resized each candidate box to a fixed size, which could potentially lead to information loss or deformation, thus impacting subsequent feature extraction. In the studies [19,40], a cascading model approach was utilized, where errors may accumulate gradually during the concatenation process, potentially resulting in larger final output errors compared with our approach with single-model output [41]. Although the grasp success rate is close to ours, their feature extraction network utilized six dense blocks, each containing convolutional modules, whereas our backbone extraction network only employed four convolutional modules [42] and two branches in the grasping network, with one branch dedicated to generating bounding boxes. However, if the detector fails to detect objects, the grasp will become ineffective, leading to performance degradation. Our model overcomes the shortcomings of the aforementioned methods.

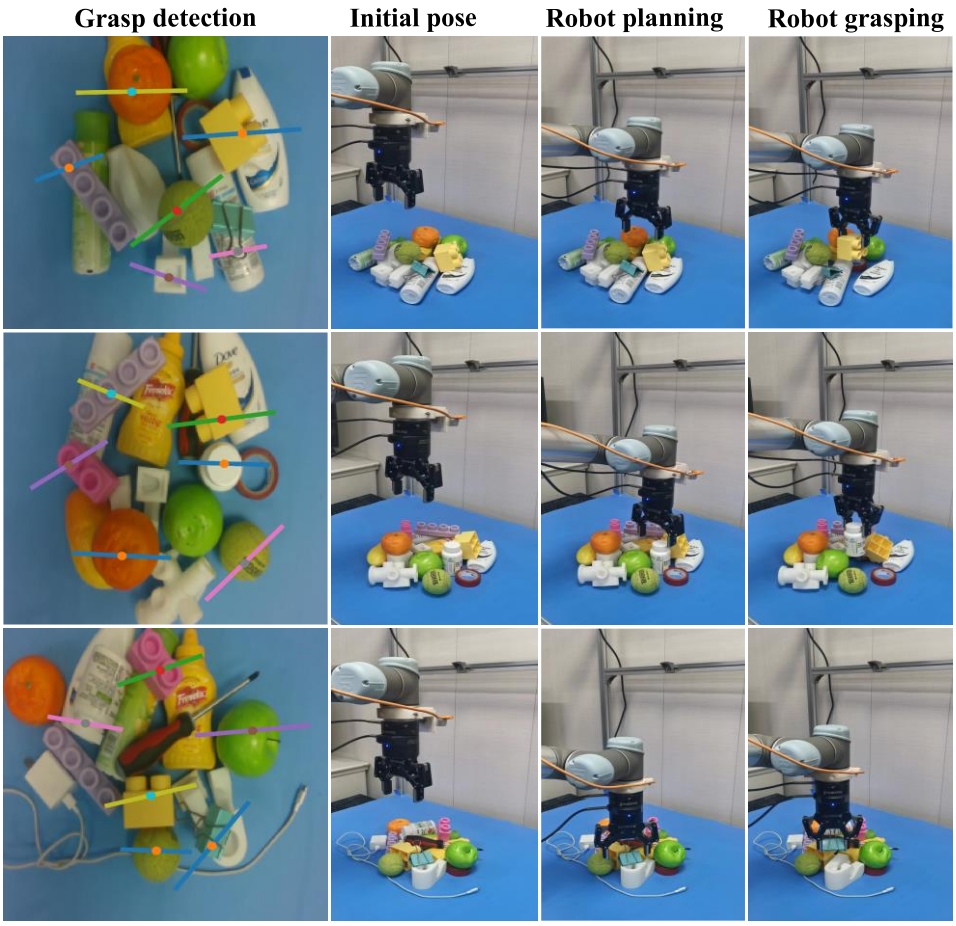


Figure 17. Robot grasping in different cluttered scenes.

Table 5. Results in cluttered scenarios.

Method	Objects in Cluttered Scenarios
	Accuracy (%)
Morrison et al. [4]	87
Zhang et al. [8]	84
Liu et al. [19]	90.2
Yu et al. [40]	90
Asif et al. [41]	90
Li et al. [42]	87
Ours	91.0

7. Discussion

We conducted a comprehensive experiment to evaluate the effectiveness and advancement of our grasp detection algorithm, FAGD-Net. Firstly, we achieved accuracies of 96.5% and 98.9% on Jacquard and Cornell grasp dataset, respectively, showcasing the robustness of our approach. By comparative analysis with existing algorithms, we further highlight the advantages of FAGD-Net.

Subsequently, we conducted ablation studies to validate the contributions of the proposed Res-EMA module and FFPM. Additionally, we introduced Res-SE and Res-CoA modules based on Squeeze-and-Excitation (SE) and Coordinate Attention (CoA) mechanisms from previous grasp works [13,14]. Through comparisons with Res-EMA and Res-block, we emphasized the effectiveness of the Res-EMA module. Furthermore, we compared feature concatenation and fusion methods within the model to highlight the advantages of our FFPM.

Finally, real-world robot grasping experiments were conducted, achieving success rates of 95.0% for single-object scenes, 93.3% for multi-object scenes, and 91.0% for cluttered scenes.

In the tests, we conducted grasping experiments using diverse objects that varied in size, orientation, and type. FAGD-Net demonstrated strong generalization capability. In some other environments, our FAGD-Net can also be applicable, such as on tabletops with textured backgrounds and in low-light environments. The main reason is that textured backgrounds and low-light conditions do not severely affect depth information acquisition.

Although FAGD-Net demonstrated satisfactory grasping capabilities, some potential issues were identified. Occasional collisions with other objects occur during the grasping process, leading to the accidental removal of the object from the scene. Additionally, instances of grasping multiple objects simultaneously were observed, as depicted in Figure 18. There are two main types of failed grasping as follows: (1) when the robot approaches the object to be grasped, the gripper is blocked by cluttered objects and does not have enough space for the parallel-jaw gripper. (2) The grasped target slips from the parallel-jaw gripper because the center of gravity is unstable. Further improvements and optimizations are required to address these challenges and enhance the overall performance of the algorithm.

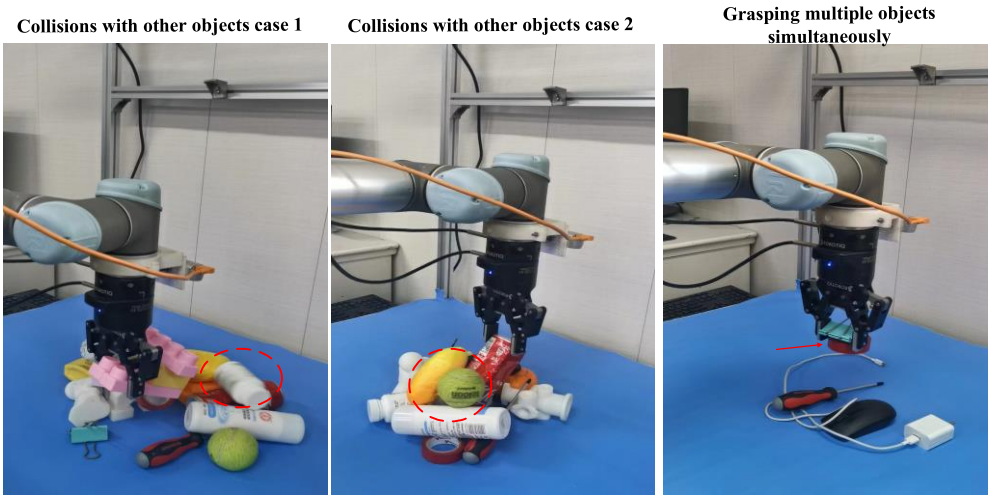


Figure 18. Suboptimal grasping results.

It is worth noting that we consider the grasp configuration that infers the highest confidence from the image. The attributes of the target, such as a certain object's density and rigitity, are not accounted for in the model. The current focus of this study is primarily on objects grasping and placing. We will further study and improve upon them in the future.

Additionally, in this study, the grasped objects have different shapes, making them more suitable for grasp configurations using parallel grippers. In some scenarios, using vacuum/magnetic grippers may be more efficient, such as for flat surfaces or the objects are easily deformed.

8. Conclusions

This work proposes a high-performance grasp detection model named FAGD-Net for predicting optimal grasp configuration. The model contains two core modules, Res-EMA module and FFPM. The Res-EMA module adjusts the importance of feature channels while preserving accurate spatial information features, and the FFPM effectively addresses the issue of overlooked or lost grasp-related features during the down-sampling process in the encoder. We conducted experiments on public grasp datasets and robot experimental platforms to validate the approach. Our proposed model achieves accuracy rates of 98.9% and 96.5% on the Cornell and Jacquard datasets, respectively. Moreover, we conducted robot grasping experiments in various scenes, achieving success rates of 95.0% for single-object

scenes, 93.3% for multi-object scenes, and 91.0% for cluttered scenes. The experimental results demonstrate the effectiveness and advancement of the proposed model.

Author Contributions: Conceptualization, X.Z.; methodology, X.L.; software, X.Z. and X.L.; validation, X.L., Q.L., and X.Z.; formal analysis, T.G. and Y.S.; data curation, X.Z. and X.L.; writing—original draft preparation, X.Z., X.L., and Q.L.; writing—review and editing, X.Z. and X.L.; visualization, X.L. and X.Z.; supervision, H.H.; project administration, X.Z. and Q.L.; funding acquisition, X.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 61703356, in part by the Natural Science Foundation of Fujian Province under Grant 2022J011256 and 2020J01285, and in part by the Xiamen Natural Science Foundation (3502Z20227215).

Data Availability Statement: Data is contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hu, Y.; Wu, X.; Geng, P.; Li, Z. Evolution strategies learning with variable impedance control for grasping under uncertainty. *IEEE Trans. Ind. Electron.* **2019**, *66*, 7788–7799. [CrossRef]
2. Li, G.; Li, N.; Chang, F.; Liu, C. Adaptive Graph Convolutional Network with Adversarial Learning for Skeleton-Based Action Prediction. *IEEE Trans. Cogn. Dev. Syst.* **2022**, *14*, 1258–1269. [CrossRef]
3. Solowjow, E.; Ugalde, I.; Shahapurkar, Y.; Aparicio, J.; Mahler, J.; Satish, V.; Goldberg, K.; Claussen, H. Industrial Robot Grasping with Deep Learning using a Programmable Logic Controller (PLC). In Proceedings of the 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), Hong Kong, China, 20–21 August 2020; pp. 97–103. [CrossRef]
4. Morrison, D.; Corke, P.; Leitner, J. Learning robust, real-time, reactive robotic grasping. *Int. J. Robot. Res.* **2020**, *39*, 183–201. [CrossRef]
5. Mahler, J.; Liang, J.; Niyaz, S.; Laskey, M.; Doan, R.; Liu, X.; Ojea, J.A.; Goldberg, K. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv* **2017**, arXiv:1703.09312.
6. Kumra, S.; Joshi, S.; Sahin, F. Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2020. [CrossRef]
7. Teng, Y.; Gao, P. Generative Robotic Grasping Using Depthwise Separable Convolution. *Comput. Electr. Eng.* **2021**, *94*, 107318. [CrossRef]
8. Zhang, H.; Lan, X.; Bai, S.; Zhou, X.; Tian, Z.; Zheng, N. ROI-based Robotic Grasp Detection for Object Overlapping Scenes. In Proceedings of the 2019 IEEE International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4768–4775. [CrossRef]
9. Lenz, I.; Saxena, A. Deep learning for detecting robotic grasps. *Int. J. Robot. Res.* **2015**, *34*, 705–724. [CrossRef]
10. Wang, S.; Zhou, Z.; Kan, Z. When transformer meets robotic grasping: Exploits context for efficient grasp detection. *IEEE Robot. Autom. Lett.* **2022**, *7*, 8170–8177. [CrossRef]
11. Yu, S.; Zhai, D.-H.; Xia, Y.; Wu, H.; Liao, J. SE-ResUNet: A novel robotic grasp detection method. *IEEE Robot. Autom. Lett.* **2022**, *7*, 5238–5245. [CrossRef]
12. Cao, H.; Chen, G.; Li, Z.; Lin, J.; Knoll, A. Residual squeeze-and-excitation network with multi-scale spatial pyramid module for fast robotic grasping detection. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13445–13451. [CrossRef]
13. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
14. Zhou, Z.; Zhu, X.; Cao, Q. AAGDN: Attention-Augmented Grasp Detection Network Based on Coordinate Attention and Effective Feature Fusion Method. *IEEE Robot. Autom. Lett.* **2023**, *8*, 3462–3469. [CrossRef]
15. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
16. Chu, F.; Xu, R.; Vela, P.A. Real-world multiobject, multigrasp detection. *IEEE Robot. Autom. Lett.* **2018**, *3*, 3355–3362. [CrossRef]
17. Yan, Y.; Tong, L.; Song, K.; Tian, H.; Man, Y.; Yang, W. SISG-Net: Simultaneous Instance Segmentation and Grasp Detection for Robot Grasp in Clutter. *Adv. Eng. Inform.* **2023**, *58*, 102189. [CrossRef]
18. Suwoyo, H.; Hidayat, T.; Jia-nan, F. A Transformable Wheel-Legged Mobile Robot. *Int. J. Eng. Contin.* **2023**, *2*, 27–39.
19. Liu, D.; Tao, X.; Yuan, L.; Du, Y.; Cong, M. Robotic Objects Detection and Grasping in Clutter based on Cascaded Deep Convolutional Neural Network. *IEEE Trans. Instrum. Meas.* **2021**, *71*, 1–10. [CrossRef]
20. Zhang, H.; Zhou, X.; Lan, X.; Li, J.; Tian, Z.; Zheng, N. A real-time robotic grasping approach with oriented anchor box. *IEEE Trans. Syst. Man Cybern. Syst.* **2021**, *51*, 3014–3025. [CrossRef]

21. Cao, H.; Chen, G.; Li, Z.; Feng, Q.; Lin, J.; Knoll, A. Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation. *IEEE/ASME Trans. Mechatron.* **2022**, *28*, 1384–1394. [CrossRef]
22. Fu, K.; Dang, X. Light-Weight Convolutional Neural Networks for Generative Robotic Grasping. *IEEE Trans. Ind. Inform.* **2024**, *10*, 3353841. [CrossRef]
23. Tian, H.; Song, K.; Li, S.; Ma, S.; Yan, Y. Lightweight Pixel-Wise Generative Robot Grasping Detection Based on RGB-D Dense Fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
24. Li, W.; Lambert-Garcia, R.; Getley, A.C.M.; Kim, K.; Bhagavath, S.; Majkut, M.; Rack, A.; Lee, P.D.; Leung, C.L.A. AM-SegNet for additive manufacturing in situ X-ray image segmentation and feature quantification. *Virtual Phys. Prototyp.* **2024**, *19*, e2325572. [CrossRef]
25. Ma, H.; Han, G.; Peng, L.; Zhu, L.; Shu, J. Rock thin sections identification based on improved squeeze-and-Excitation Networks model. *Comput. Geosci.* **2021**, *152*, 104780. [CrossRef]
26. Qi, J.; Liu, X.; Liu, K.; Xu, F.; Guo, H.; Tian, X.; Li, M.; Bao, Z.; Li, Y. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease. *Comput. Electron. Agric.* **2022**, *194*, 106780. [CrossRef]
27. Shaar, F.; Yılmaz, A.; Topcu, A.E.; Alzoubi, Y.I. Remote Sensing Image Segmentation for Aircraft Recognition Using U-Net as Deep Learning Architecture. *Appl. Sci.* **2024**, *14*, 2639. [CrossRef]
28. Fan, Z.; Liu, K.; Hou, J.; Yan, F.; Zang, Q. JAUNet: A U-shape Network with Jump Attention for Semantic Segmentation of Road Scenes. *Appl. Sci.* **2023**, *13*, 1493. [CrossRef]
29. Liu, Y.; Bai, X.; Wang, J.; Li, G.; Li, J.; Lv, Z. Image Semantic Segmentation Approach Based on DeepLabV3 Plus Network with an Attention Mechanism. *Eng. Appl. Artif. Intell.* **2024**, *127*, 107260. [CrossRef]
30. Tsai, R.Y.; Lenz, R.K. A new technique for fully autonomous and efficient 3D robotics hand/eye calibration. *IEEE Trans. Robot. Autom.* **1989**, *5*, 345–358. [CrossRef]
31. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
32. He, K.; Zhang, X.; Ren, S.; Sun, J. Identity mappings in deep residual networks. In *Computer Vision—ECCV 2016, 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016, Proceedings, Part IV*; Springer: Cham, Switzerland, 2016; pp. 630–645.
33. Depierre, A.; Dellandréa, E.; Chen, L. Jacquard: A Large Scale Dataset for Robotic Grasp Detection. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 3511–3516. [CrossRef]
34. Yun, J.; Moseson, S.; Saxena, A. Efficient grasping from RGBD images: Learning using a new rectangle representation. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation (ICRA), Shanghai, China, 9–13 May 2011; pp. 3304–3311. [CrossRef]
35. Zhou, X.; Lan, X.; Zhang, H.; Bai, S.; Tian, Z.; Zhang, Y.; Zheng, N. Fully convolutional grasp detection network with oriented anchor box. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 7223–7230.
36. Ainetter, S.; Fraundorfer, F. End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 13452–13458.
37. Chen, L.; Huang, P.; Li, Y.; Meng, Z. Edge-dependent efficient grasp rectangle search in robotic grasp detection. *IEEE/ASME Trans. Mechatron.* **2020**, *26*, 2922–2931. [CrossRef]
38. Li, Y.; Huang, P.; Ma, Z.; Chen, L. A Context-Free Method for Robust Grasp Detection: Learning to Overcome Contextual Bias. *IEEE Trans. Ind. Electron.* **2021**, *69*, 13121–13130. [CrossRef]
39. Laili, Y.; Chen, Z.; Ren, L.; Wang, X.; Deen, M.J. Custom Grasping: A Region-Based Robotic Grasping Detection Method in Industrial Cyber-Physical Systems. *IEEE Trans. Autom. Sci. Eng.* **2022**, *20*, 88–100. [CrossRef]
40. Yu, Y.; Cao, Z.; Liu, Z.; Geng, W.; Yu, J.; Zhang, W. A Two-Stream CNN with Simultaneous Detection and Segmentation for Robotic Grasping. *IEEE Trans. Syst. Man Cybern. Syst.* **2022**, *52*, 1167–1181. [CrossRef]

41. Asif, U.; Tang, J.; Harrer, S. Densely supervised grasp detector (DSGD). In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; pp. 8085–8093.
42. Li, T.; Wang, F.; Ru, C.; Jiang, Y.; Li, J. Keypoint-based robotic grasp detection scheme in multi-object scenes. *Sensors* **2021**, *21*, 2132. [CrossRef]