machines

MDPI

Article

A Novel Grasp Detection Algorithm with Multi-Target Semantic Segmentation for a Robot to Manipulate Cluttered Objects

Xungao Zhong [1,2], Yijun Chen [1], Jiaguo Luo [1], Chaoquan Shi [1] and Huosheng Hu [3,*]

[1] School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen 361024, China; zhongxungao@163.com (X.Z.); chenyijun@stu.xmut.edu.cn (Y.C.); luo_jiaguo@163.com (J.L.); shichaoquan@s.xmut.edu.cn (C.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
* Correspondence: hhu@essex.ac.uk

Abstract: Objects in cluttered environments may have similar sizes and shapes, which remains a huge challenge for robot grasping manipulation. The existing segmentation methods, such as Mask R-CNN and Yolo-v8, tend to lose the shape details of objects when dealing with messy scenes, and this loss of detail limits the grasp performance of robots in complex environments. This paper proposes a high-performance grasp detection algorithm with a multi-target semantic segmentation model, which can effectively improve a robot's grasp success rate in cluttered environments. The algorithm consists of two cascades: Semantic Segmentation and Grasp Detection modules (SS-GD), in which the backbone network of the semantic segmentation module is developed by using the state-of-the-art Swin Transformer structure. It can extract the detailed features of objects in cluttered environments and enable a robot to understand the position and shape of the candidate object. To construct the grasp schema SS-GD focused on important vision features, a grasp detection module is designed based on the Squeeze-and-Excitation (SE) attention mechanism, to predict the corresponding grasp configuration accurately. The grasp detection experiments were conducted on an actual UR5 robot platform to verify the robustness and generalization of the proposed SS-GD method in cluttered environments. A best grasp success rate of 91.7% was achieved for cluttered multi-target workspaces.

Keywords: robot manipulation; grasp detection; semantic segmentation; cluttered objects

Citation: Zhong, X.; Chen, Y.; Luo, J.; Shi, C.; Hu, H. A Novel Grasp Detection Algorithm with Multi-Target Semantic Segmentation for a Robot to Manipulate Cluttered Objects. *Machines* **2024**, *12*, 506. https://doi.org/10.3390/machines12080506

Academic Editor: Giuseppe Carbone

Received: 19 June 2024
Revised: 17 July 2024
Accepted: 25 July 2024
Published: 27 July 2024

1. Introduction

Robot grasp manipulation has been widely used in industrial assembly, sorting, and human-machine interaction [1–5]. However, robots face a challenge when they conduct grasp manipulations in cluttered environments where the objects are close to each other and have similar shapes and sizes. Therefore, how to improve the success rate of robotic grasping in cluttered environments needs to be solved urgently, and reliable grasp detection algorithms are needed. In general, grasp detection can be divided into the analytical method and the empirical method [6]. The analytical method is to calculate the grasp pose according to the object's 3D geometric model, and the established kinematics model of the manipulator. This method can realize the migration of grasping to a certain extent, but its generalization ability is limited at the modeling levels. As a 3D object model cannot be obtained beforehand, it is difficult to model the physical interaction between the robotic arm and the object [7].

On the other hand, the empirical method does not require a 3D model of the object. It trains a grasping network using data-driven techniques, and then uses the off-line learned model to reason the grasp configuration for novel objects. Yu et al. [8] used a five-dimensional rectangle to represent the grasp detection, Mahler [9] used a point and an angle to represent the structure of the gripper, and Li et al. used a 6D gripper model [10]. However, these methods cannot effectively manage cluttered objects as the robot's vision

systems are unable to accurately identify the contours and positions of individual objects. Besides, the occlusion reduces the effective features available for grasp reasoning and increasing the complexity of the grasping task.

In this paper, we focus on improving robotic grasping performance in cluttered environments. An improved adaptive grasp representation is first proposed based on width prediction. Then, we incorporate a state-of-the-art feature extraction network to contribute to target segmentation in complex cluttered environments. Innovatively, we further enhance the performance of the grasp detection network by incorporating residual and attention modules. The grasp network predicts a distinct grasp configuration for each object, then eliminates the prediction redundancy and improves the precision and efficiency of the grasping tasks. In general, the paper makes the following contributions:

- A high-performance grasp schema SS-GD is proposed by combining a semantic segmentation module with a grasp detection module to effectively reduce prediction redundancy in multi-target grasp pose detection and improve the probability of robots performing robust grasping operations.
- A grasp detection network is proposed based on Mask-D multi-channel synthetic data, aiming to enhance the perception of shape information for candidate objects. The SE attention mechanism is introduced to further strengthen the network's feature extraction capability.
- By leveraging various advanced visual algorithms, we explore the optimal segmentation-grasping cascade combination in diverse cluttered grasping scenarios.
- Experimental results in real-world environments demonstrate that our cascade SS-GD algorithm exhibits superior performance in cluttered scenarios, particularly in environments characterized by severe stacking and background interference.

The rest of the paper is organized as follows. Section 2 briefly overviews some related work. A grasp representation is presented in Section 3. Section 4 proposes a new grasp schema SS-GD based on a cascaded deep network with the combination of a semantic segmentation module and a grasp detection module. The method aims to improve the success rate in robotic grasping tasks. The results and analysis are presented in Section 5 to show the feasibility and performance of the proposed method. Finally, a brief conclusion is given in Section 6.

2. Related Works

In this section, the relevant works in the field of grasping detection are reviewed and the object segmentation algorithms in robot grasping are introduced.

2.1. Grasp Detection

To perform the grasp tasks, a dataset with rich grasp configuration information is needed to improve the neural network training and evaluation. The Cornell [11] dataset and Jacquard [12] dataset are widely used for training grasp networks and their working principles are similar, i.e., by annotating rectangular boxes, the network predicts plane grasp attributes on RGB or RGB-D images. Morrison et al. [13] introduced the Grasp Generative Convolutional Neural Network (GGCNN) to overcome limitations of current deep learning grasping techniques by avoiding discrete sampling of grasp candidates and long computation times, achieving an 83% grasp success rate on a set of previously unseen objects. In 2022, they further proposed GGCNN2, a multi-view approach which improves overall grasp success rate in clutter by 10%. Kumra et al. [14] used generative residual convolutional neural network (GR-ConvNet) to generate antipodal robotic grasping poses for novel objects from n-channel images. They optimized the gradient vanishing problem when training the network through five residual layers, enabling the network to achieve higher accuracy in Cornell and Jacquard dataset validation. Yu et al. [15] developed a grasp network called Squeeze-and-Excitation ResUNet and proved that the Squeeze-and-Excitation module can effectively improve generalization ability on different datasets.

Our objective is to train the network by inputting multimodally fused shape information, enabling the network to predict the grasp poses of unseen objects. This poses more challenges to the network's feature extraction capabilities compared to previous works. Inspired by the above works, we innovatively integrated residual modules with bottleneck layers with SE attention mechanisms based on the work of [14], and validated RGB-D inputs on the Cornell dataset and the Jacquard dataset.

2.2. Semantic Segmentation in Robots' Grasp

The single target grasping algorithm cannot grasp messy, cluttered targets in messy environments in which the objects are blocking each other, which increases the difficulty of grasping detection [16,17]. Araki et al. [18] proposed a multi-task model that can simultaneously conduct object detection, semantic segmentation, and sucker grasping detection. This model can perform accurate object detection and segmentation but cannot deal with scenarios that require the use of an antipodal gripper. Xu et al. [19] proposed a multi-task convolutional network to represent the grasping detection of a manipulator, which was based on semantic segmentation and usable to extend to scenarios with unknown categories. Xie et al. [20] proposed a two-stage object instance-level segmentation network (UOIS-Net), separately leveraging synthetic RGB and synthetic depth for unseen object instance segmentation. The method generated preliminary masks by employing depth maps and regressing center votes in either 2D or 3D. Subsequently, these initial masks undergo refinement using RGB information, and finally, cascade with the 6D grasping network to complete the grasping task. Ainetter et al. [21] proposed a depth-aware Coordinate Convolution algorithm to improve the accuracy of grasping detection and object segmentation. However, they evaluated on a training dataset, but not in actual grasping scenarios. Wang et al. [16] used a vision transformer [22] as the backbone for target detection and applied it to robot grasping tasks. Its global self-attention mechanism was time-consuming, and its vision transformer produced a low-resolution feature map.

To execute high-performance grasping operations within cluttered scenes featuring unknown categories, we have employed Swin Transformer [23] as the backbone in the semantic segmentation network. The aim is to deliver precise target localization and shape information for the subsequent grasping detection network. The integration of shifted windows and multi-head self-attention is instrumental in facilitating global interaction, thereby augmenting the overall performance and generalization capabilities of our approach in the context of multi-object grasping within cluttered environments.

3. Grasp Representation

The grasp configuration in this paper is improved according to the definition of the grasp rectangle [24]. The adaptive grasp width is introduced based on five-dimensional grasping, and redefined below:

$$G_r = (P_r, \phi_r, w_{r-o}, w_{r-c}, Q_r) \tag{1}$$

where $P_r = (x_r, y_r, z_r)$ is the grasp position in the robot coordinate system, ϕ_r is the angle of rotation around the z-axis, and w_{r-o} and w_{r-c} are the open and closed width of two-finger gripper approaching and grasping the object, respectively. Q_r stands for grasping confidence, its scalar value range is [0,1]; the closer to 1, the higher the grasping confidence.

The current robot grasp detection algorithm completely closes the robot gripper when picking up thin or fragile objects, which may damage the objects. Therefore, we introduce the grasping closed width w_{r-c}, which is set according to w_{r-o}; here we have employed $w_{r-c} = \lambda w_{r-o}$. Through the experimental testing, when $\lambda < 0.4$ is used, due to the large closure of the gripper, it may damage the object when picking up thin plastic, paper cups, and other objects. When $\lambda > 0.4$ is used, when picking up heavier objects, it may make the grasping unstable, and the object may fall off. When $\lambda = 0.4$, the gripper is closed to the appropriate width.

Introducing an adaptive grasping width can improve the grasping performance compared to grasping configurations with only the position and rotation angle of the grasping point alone. We detect a pixel-level grasp configuration from RGB image $I = R^{3\times h\times w}$ and depth $D = R^{h\times w}$ with height h and width w, which can be defined as follows:

$$G_i = (P_i, \phi_i, w_{i-o}, w_{i-c}, Q_i) \tag{2}$$

where $P_i = (x_i, y_i)$ is the grasp position in the image coordinates, and f_i is the rotation angle in the camera coordinate, which represents the rotation scalar of each point required to grasp the object of interest; the range is in $[0, \pi]$. w_{i-o} and w_{i-c} are the opening width and closing width, respectively, of the gripper predicted by the network in the image coordinate system. Q_i is the grasp confidence, which is of each point in the image, and its scalar value is between 0 and 1.

The closer the value to 1, the greater the success rate of grasping. The goal is to infer a set of grasp $G = (G_1, G_2, ..., G_k)$ that maximizes the grasping success rate given k groups of the candidate grasp:

$$\{G_i^*\} = \text{argmax prob}(Q_i|I, D, G_i) \tag{3}$$

To execute grasping tasks, the pixel-level grasp detection should be transformed into a gripper configuration. It involves system calibration and a robot moving model:

$$G_r = T_{rc}(T_{ci}(G_i^*)) \tag{4}$$

where T_{ci} represents the conversion function from 2D image coordinates to camera coordinates, and T_{rc} is the conversion from camera coordinates to robot workspace.

4. Principle and Method

Considering the grasp tasks are conducted by using single vision perception, the robot should accurately detect the candidate target and realize an appropriate grasp configuration, especially in grasping similar objects in cluttered environments and improving the success rate of the grasping. Thus, this section presents a grasp method SS-GD for cluttered environments, which is based on the cascaded deep network.

As shown in Figure 1, the semantic segmentation module with Swin Transformer is used for multi-target semantic segmentation. This module consists of a series of hierarchical Transformer blocks to extract multi-scale features from the input image. Then, an accurate segmentation mask of the candidate target is predicted for subsequent grasp detection. The grasp detection module is designed with a Squeeze-and-Excitation Bottleneck and an encoder-decoder network architecture. The SE-Bottleneck module enhances feature representation by adaptively recalibrating channel-wise features based on a global intra-feature relationship. This is achieved through a squeeze operation that generates channel-wise statistics, followed by an excitation operation that rescales the original features accordingly.

The encoder-decoder structure of the grasp detection module is tailored to handle the spatial dimensions of the grasping area. The SE-Bottleneck module's decay rate is set to 16. The encoder compresses the spatial dimensions and extracts high-level semantic information, while the decoder up-samples and expands these dimensions to restore the precise grasping area. The mask of each object, combined with depth information to form an RGB-D input, is fed into the grasp detection module. This module outputs the corresponding grasp configuration for each object, including the grasp quality, grasp angle, and grasp width for each pixel within the mask region. The grasp configuration with maximal quality, as detected by the network, is identified as the optimal grasp position.

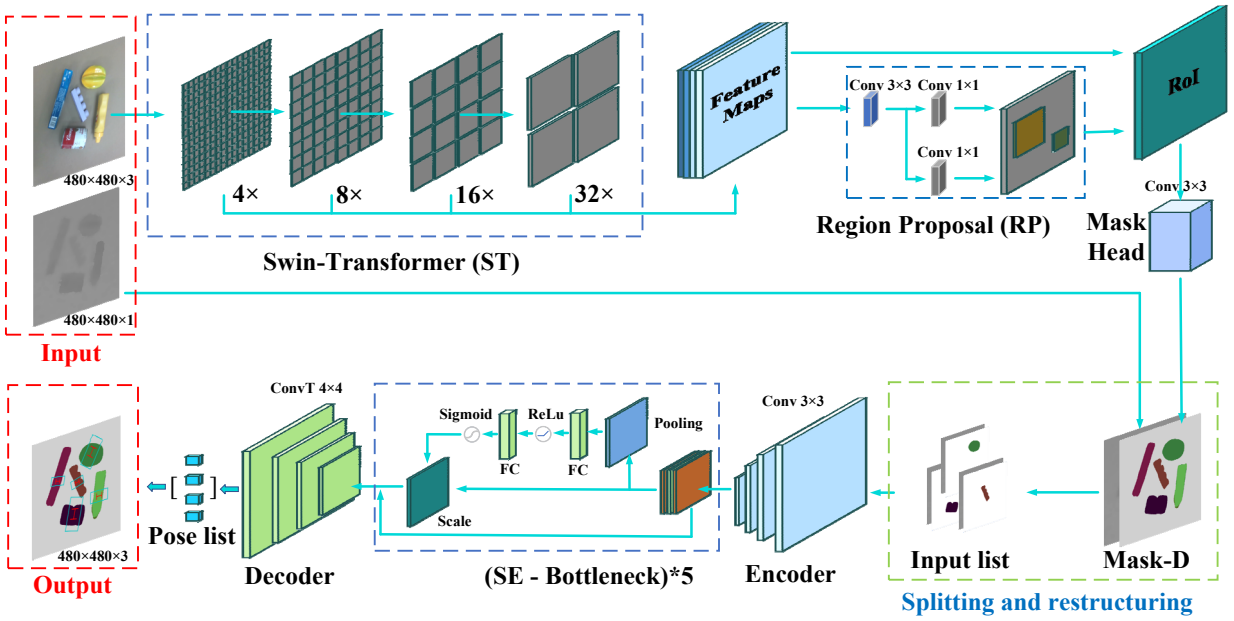

Figure 1. The structure of the proposed SS-GD schema by combining semantic segmentation module and grasp detection module.

4.1. Semantic Segmentation Module

Previous methodologies furnish instance masks alongside category-level semantic labels, which often face challenges in generalizing to novel categories [18,21]. In our approach, we introduce an object proposal algorithm to yield masks relevant to the grasp detection network. The pixel-level accurate segmentation enables a robot to understand the shape and location information of the object to be grasped. In the Swin Transformer [24] architecture, images are divided into a series of hierarchical blocks instead of treating the entire image as a continuous grid. Each block consists of a set of pixels, and these blocks are processed into multi-level feature representations. In each block, instead of fully connecting every pixel, the pixels are divided into multiple windows, and local operations are performed on these windows, thus reducing computational costs.

The information from different blocks is integrated by performing feature fusion at various layers. This hierarchical structure helps the model capture image information at different scales. Specifically, feature maps with different scales are extracted from the collected RGB image by Swin Transformer. These feature maps are fed into the Regional Recommendation Network to generate regional proposals, which are subsequently pooled through region of interest proposals. Finally, the binary mask is output through two layers of convolution. The details of the segmentation network are analyzed below:

(1) Training strategy

In this work, the Swin Transformer is adopted as the semantic segmentation backbone network, and the loss function also follows the Transformer settings. Thus, the loss of the mask is a binary cross-entropy function as follows:

$$L_{mask} = \frac{1}{n}\sum_{i=1}^{n}[-y_i \cdot log(p(x_i)) - (1-y_i) \cdot log(1-p(x_i)] \quad (5)$$

where y_i is the ground truth and $p(x_i)$ is the predicted unit pixel value. n represents the total number of pixels.

The GraspNet-1Billion dataset [25] provides a simulation of object positional information within cluttered scenes, including 88 distinct object types. However, the substantial scale of the 3D dataset imposes a computational resource burden during network training. To address this issue, the dataset is annotated in the COCO label format, facilitating network training, and partitioned into a training set and a test set in a 4:1 ratio. Training

comprised 100 epochs, employing the Adam optimizer with a momentum parameter set to 0.9, and an initial learning rate established at 0.001. This configuration was chosen to balance computational efficiency with effective model convergence.

(2) Evaluation metric

We utilized the mean Intersection over Union (mIoU) to calculate the mean average precision in the GraspNet-1Billion dataset, which serves as the primary evaluation metric. Mask R-CNN [26] and Yolo-v8 were trained with identical parameter configurations, which should ensure a consistent experimental setup, and enable a fair comparison in terms of segmentation performance. After 100 training epochs, we obtained validation results for three segmentation networks, as depicted in Figure 2.

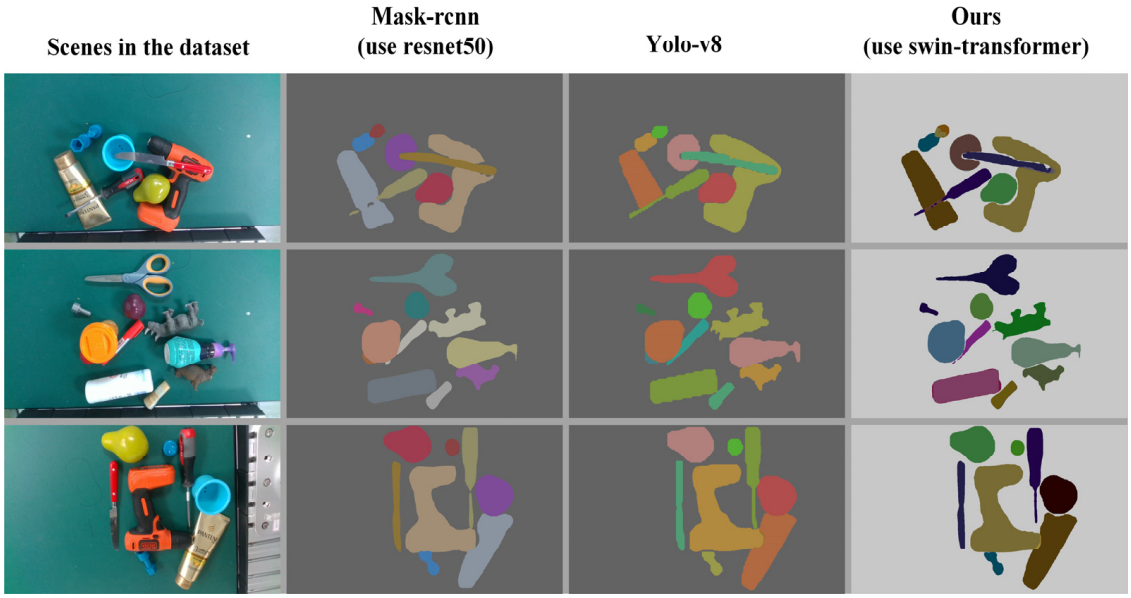

Figure 2. The segmentation results of cluttered targets with different semantic segmentation networks.

The results reveal a notable lack of shape details in the validation outcomes of the Mask R-CNN and Yolo-v8 frameworks. Mask R-CNN provides less accurate descriptions of occluded regions, while the results generated by Yolo-v8 exhibit partial omissions and face challenges in generating a smooth edge. In contrast, our segmentation module demonstrates superior performance in preserving a more comprehensive set of shape information by leveraging the Swin Transformer architecture.

In real-world scenes, as shown in Figure 3, we conducted experiments to further validate the robustness of the segmentation network, involving scenarios with novel objects. For textureless background segmentation tasks, Mask R-CNN segments all targets but introduces some segmentation errors outside the boundaries of the objects. In more complex background scenarios, such as experimental scenes with wrinkled and textured surfaces shown in the last two columns of Figure 3, these disturbances often propagate into the target regions. YOLO-v8 tends to accurately reconstruct object shape information across different experimental scenarios. However, this result does not generalize well to novel objects. In contrast, the Swin Transformer network consistently generates high-quality object segmentation in both scenarios, demonstrating robust performance in real-world environments with diverse scenes and object categories. These highlight the reliable segmentation robustness of our module in complex scenarios.

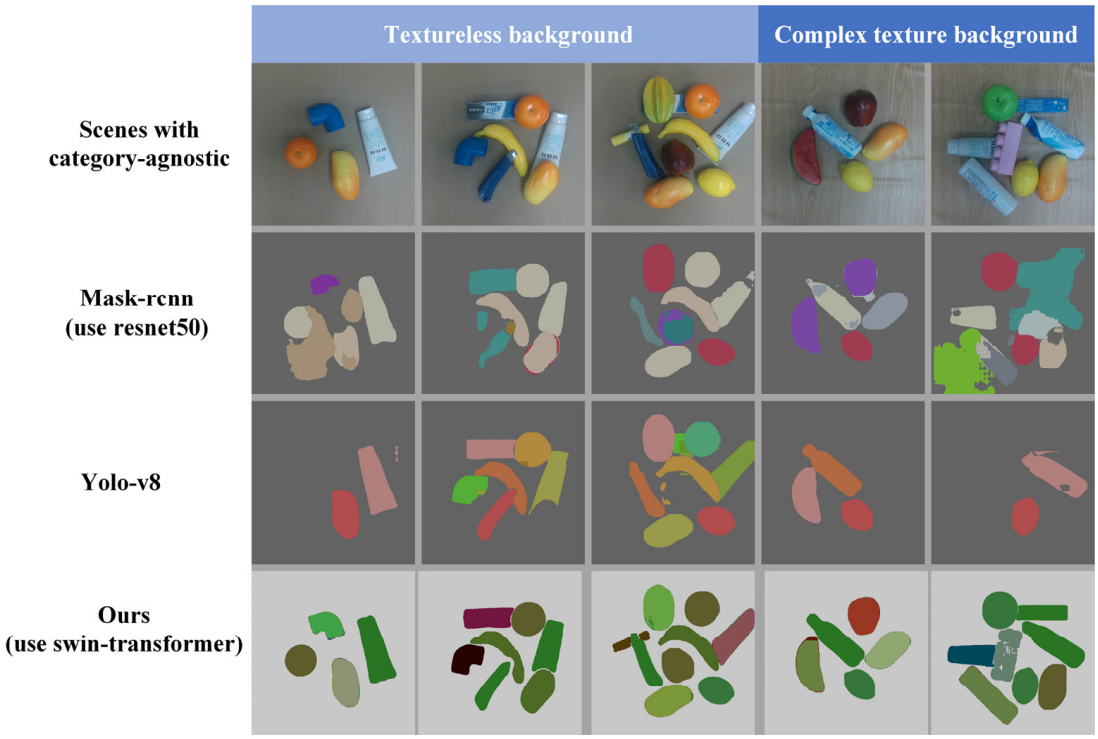

Figure 3. The validation for segmentation networks in different real-world scenarios with novel objects.

4.2. Grasp Detection Module

As shown in Figure 1, the designed grasp detection schema consists of an encoder module, SE-Bottleneck attention mechanism module, and a decoder module. We separate a single target from the RGB mask, and the depth map is based on the binary mask output of the instance segmentation. The segmented mask is combined with a depth map as the input of the grasping detection module to predict the grasp configuration of the candidate object. They are composed of four channels of input information we called Mask-D.

$$Input = concat(mask, depth) \tag{6}$$

Following, the details of the grasp detection network will be analyzed.

Encoder model: The encoder is constructed by four down-sampling convolution modules and used to extract high-level features representations from the processed data. It can also extract the gripper configuration information and map it into the low-dimensional distribution.

SE-Bottleneck: Five SE-Bottleneck modules are employed to dynamically learn inter-channel dependencies. Figure 4 shows the SE-Bottleneck architecture that involves a squeezing operation to reduce dimensionality and an excitation operation for channel-wise dependency modulation. Firstly, the dimension of the input feature map is reduced by a 1×1 convolution layer, and the number of channels is reduced simultaneously, thus reducing the calculation cost of the subsequent convolution layer. Then, the convolution layer uses a 3×3 convolution kernel for feature extraction to cover a larger receptive field, and the number of channels is increased back to the present value through another 1×1 convolution layer to match the output feature map in the identity mapping path, as shown in Equation (7).

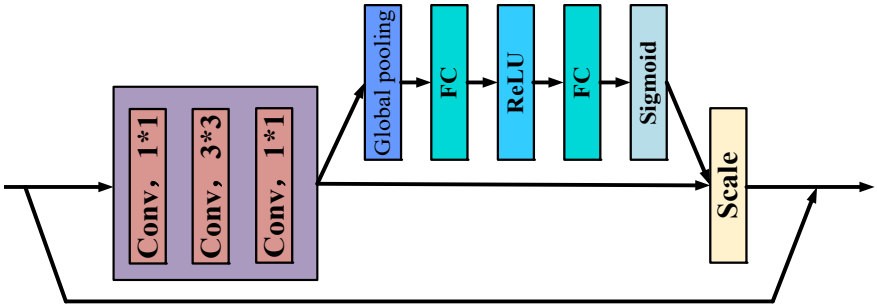

Figure 4. The structure of the SE-Bottleneck attention mechanism module.

Next, we introduce the Squeeze-and-Excitation (SE) module [27] to improve the feature extraction ability of the grasping network. This module first performs the squeeze operation on the feature map to obtain the channel-level global feature information, that is, uses the global average pooling layer to compress the feature parameters, as shown in Equation (8). Finally, the excitation operation is performed on the global features to learn the relationship between different channels, and to multiply the original feature map to obtain the final grasping feature information, as shown in Equations (9) and (10). The feature extraction process is as follows:

$$u_c = v_c * X = \sum_{s=1}^{C'} v_c^s * x^s \tag{7}$$

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{8}$$

$$s = sigmoid(W_2 * ReLU(W_1 z_c)) \tag{9}$$

$$\widetilde{x}_c == s u_c \tag{10}$$

where v_c represents the *c-th* convolutional kernel, $X \in R^{H' \times W' \times C'}$ represents the input, x^s represents the s-th input covered by the current convolutional kernel, $u_c \in R^{H \times W}$ represents the output, z_c is the result of performing global average pooling on the feature $U = [u_1, u_2, ...u_c]$ over spatial dimensions $H \times W$, and W_1 and W_2 represent linear layers. The calculated s here is the core of this module, representing the weights for each channel. $\widetilde{x}_c$ denotes the output after processing through the SE-Bottleneck.

Decoder module: The decoder consists of four up-sampling convolution modules to accurately restore the grasping area due to the grasping area being smaller than the object masks. The decoder is configured with three parallel-configured grasping heads at its summit, and can separately generate the grasp quality, grasp angle, and grasp width of each pixel in each mask region, as well as a feature map encapsulating attributes related to grasping. The network identifies the area with maximal quality as the optimal grasp position.

In contrast to traditional multi-target grasping pose prediction, our network focuses on generating a unique and reliable grasping pose for each object. Masks output by the semantic segmentation network will result in n inputs for the grasping detection module, and each input will generate a unique prediction based on the maximum confidence through the Encoder model, SE-Bottlenecks and Decoder module, which are finally stored in the list. When $(n-1)$ cycles are completed, all these predictions will be read out and a unique visual grasping rectangle will be generated. Subsequently, these grasping rectangular boxes are converted by Equation (4) and participate in the sequencing, and the positions and postures at the highest places will be preferentially sent to the actuator.

(1) Training strategy

In the grasp detection module, we use the Smooth L1 loss function because of its robustness. As shown in Equation (11), Smooth L1 can limit the gradient in two ways, where x is the difference between the predicted value and the ground truth. When the difference between the prediction value and the ground truth is large, the gradient value

will be not be excessively suppressed. When the difference between the predicted value and the ground truth is small, the gradient values will remain sufficiently small without vanishing:

$$Smooth_{L1}(x) = \begin{cases} 0.5x^2, & if|x|<1 \\ |x|-0.5, & otherwise \end{cases} \tag{11}$$

In the prediction task, the loss function configured for grasp training can be defined as follows:

$$Loss = \frac{1}{m}\sum_{i=1}^{m} smooth_{L1}(w_i - w_i^*) \tag{12}$$

where w_i is the predicted value and w_i^* is the ground truth, and m refers to the index of all predicted objects.

(2) Evaluation metric

The Cornell and Jacquard datasets contain rich grasping configuration information, which were used to train our grasping network. The Adam optimizer was employed to optimize and train the network with an initial learning rate set to 0.001, and the ratio of training and test set to 4:1. The network was trained end-to-end for 100 epochs. When the following conditions are satisfied, the network's predicted grasp result is dependable:

$$\begin{cases} \Delta\phi = \left|degrees(\phi_p - \phi_l)\right| < 30^\circ \\ IoU = \left|\frac{P \cap GT}{P \cup GT}\right| > 0.25 \end{cases} \tag{13}$$

where $\Delta\phi$ is the rotation angle difference between the predicted grasping rectangle ϕ_p and the ground truth rectangle ϕ_l. The Intersection over Union (*IoU)* score is a measure of the overlap between the predicted grasping rectangle P and the ground truth rectangle GT.

5. Results and Analysis

5.1. Evaluation on Cornell and Jacquard Datasets

To verify the performance of the proposed grasp detection module SS-GD, the RGB-D images in Cornell and Jacquard are used to evaluate the designed network and to compare between the state-of-the-art methods. The results of IoU are used to score the predicted grasping rectangle and the ground truth rectangle.

As shown in Table 1, the baseline designed without the SE-Bottleneck exhibited a grasping accuracy of 94.3% in Cornell and 93.6% in Jacquard. In contrast, our proposed method, leveraging SE-Bottleneck attention mechanisms, achieved superior grasping detection accuracy, exhibiting a grasping accuracy of 97.8% in Cornell and 94.9% in Jacquard. The significant improvement in grasping detection accuracy highlights the substantial impact of the introduced SE bottleneck attention module. In addition, our proposed network crawling inference time is about 40ms, which meets most scenarios that require real-time detection.

Table 1. Comparison results on Cornell and Jacquard data.

Author	Methods	Cornell	Jaquard
Morrison [24]	GGCNN2	65.0%	84.0%
Depierre [28]	Grasping Regression	95.2%	85.7%
Wang [16]	TF-GRASP	96.7%	94.6%
Kumra [14]	GR-ConvNet	96.6%	94.6%
Song [29]	RPN	95.6%	91.5%
Liu [30]	Q-Net	95.2%	92.1%
Ours	baseline	94.3%	93.6%
Ours	SS-GD	97.8%	94.9%

The visualization verification and evaluation results of grasping pose are shown in Figure 5. The grasp configuration is generated at the location where the grasping quality is

enhanced, effectively verifying the accuracy of the proposed grasping detection method in different object categories.

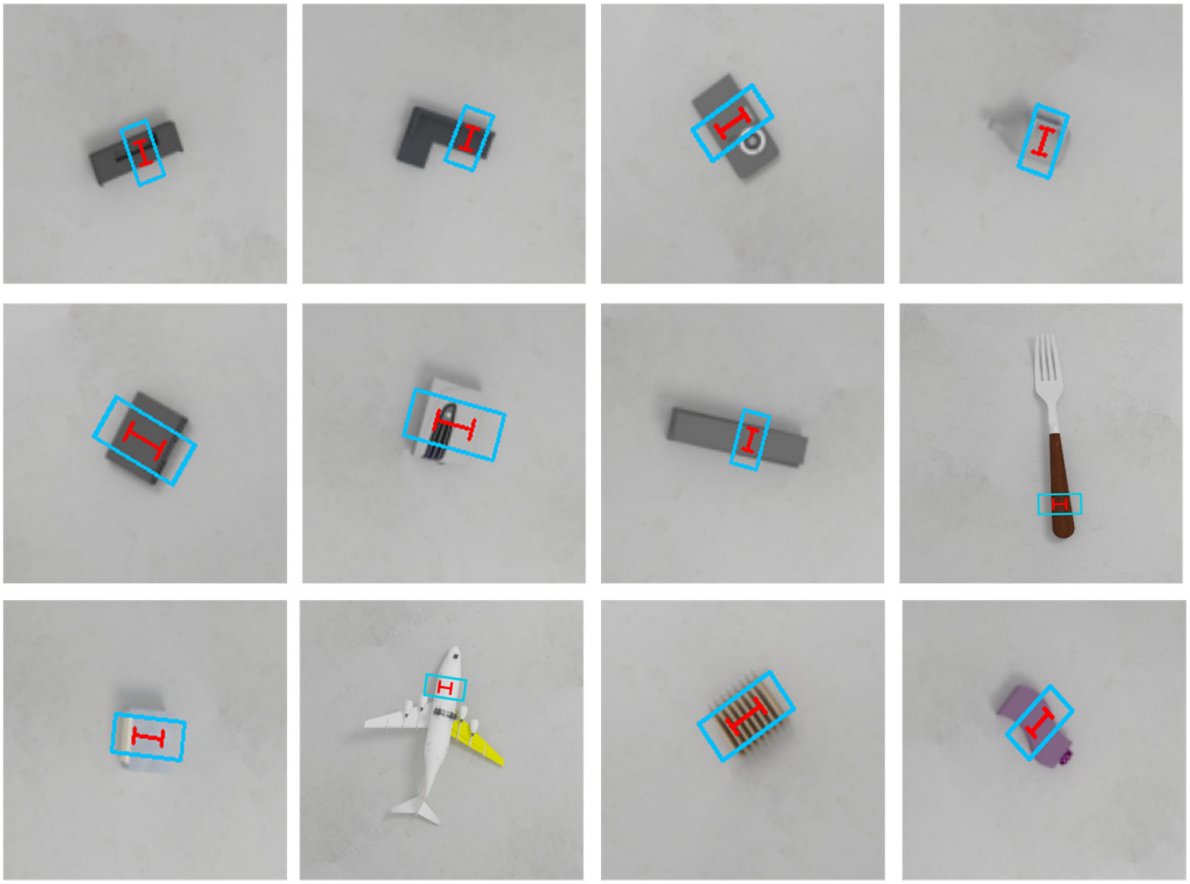

Figure 5. Grasping detection evaluation on the Jacquard dataset. The blue box indicates the pre-grasp position of the gripper, and the red line indicates the closing width of the gripper.

Discussion: Although the self-attention mechanism used for robotic grasping detection tasks exhibits commendable global interaction capabilities, its global self-attention mechanism introduces a considerable time overhead and yields a low-resolution feature map. This, in turn, can result in performance degradation, making grasping of objects in cluttered environments difficult. Our proposed SE-Bottleneck attention mechanism introduced in the grasping detection module can help the robot pay attention to important grasping configuration information, which improves the performance of real grasping tasks.

5.2. Ablation Experiments

The inputs for grasp framework with segmentation network consist of masks and depth maps, namely Mask-D. As depicted in Figure 6, the grasp framework without the involvement of a segmentation network utilizes the conventional RGB-D as inputs. The segmentation-independent grasp framework is distinguished from the segmentation-cascaded grasping framework by different visualization bounding boxes.

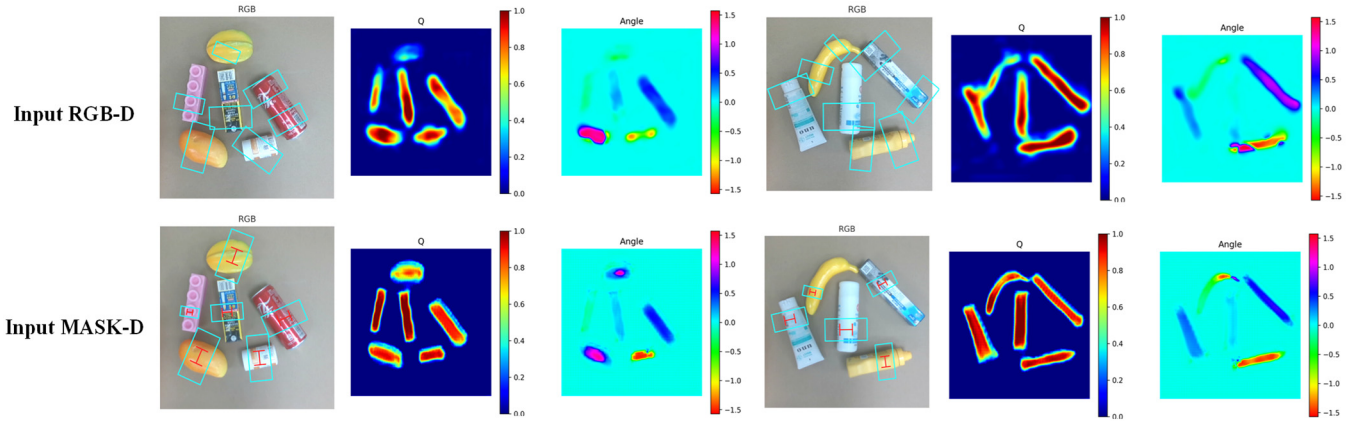

Figure 6. The validation of assistance of masks in grasp detection. The first line represents the input with RGB-D image, while the second line illustrates the input with Mask-D image processed by the segmentation network.

In the context of visualization bounding boxes, the predicted grasp angles by our segmentation-cascaded framework are superior to that of the segmentation-independent framework. This is reflected in the quality maps of the grasping, which illustrates that mask-based input can contribute more uniform quality distribution and facilitate reasonable angle predictions. It is also worth noting that multi-object detection based on global confidence often generates multiple bounding boxes for the same target. These predicted boxes are frequently chaotic, located far from the object's centroid, making it hard for the robot to grasp an object stably. On the contrary, the mask-based input for the cascaded framework can generate unique and reliable predictions for each object and demonstrates the auxiliary role of the two-stage structure in grasp detection.

In the semantic segmentation network, to verify the contribution of Swin Transformer in the network, resnet101 is used as the backbone to conduct ablation experiments. We conducted a training evaluation on the GraspNet-1Billion dataset. Based on the network framework proposed in this experiment, we used two backbone feature networks for comparison and verification. The experimental data are shown in Table 2. When resnet101 is used as backbone, the grasp detection accuracy is 89.8% and the segmentation accuracy is 74.2; when Swin Transformer is used as backbone, the grasp detection accuracy is 92.6% and the segmentation accuracy 78.6%. The experimental results show the necessity and rationality of using Swin Transformer as backbone.

Table 2. The precision of algorithms proposed by different backbone.

Backbone	Grasp Detection Accuracy (IoU%)	Segmentation Accuracy (IoU%)
Resnet101	89.8	74.2
Swin-transformer	92.6	78.6

To verify the necessity of the SE modules, we conducted corresponding ablation experiments, and we took the network without SE modules as the baseline. We trained and evaluated on the Jacquard dataset. As can be seen from Table 3, baseline achieved an accuracy of 86.5%, and the complete grasp detection module achieved an accuracy of 89.4%. The experimental results show that the SE module can improve the accuracy of the network.

Table 3. Comparison results based on the Jacquard dataset.

Method	Grasp Detection Accuracy (IoU%)
Baseline	86.5
SS-GD	89.4

5.3. Comparison Study on Real-World Tasks

To verify the excellence and rationality of the proposed algorithm, SS-GD is compared with the most advanced grasp detection algorithm GR-ConvNet [14] in practical scenarios.

Case 1: grasp detection test in a textureless workspace. As shown in Figure 7, we compared four scenarios in which the number of objects gradually increases to structure a multi-target cluttered environment. It is evident that the grasp configurations predicted by the GR-ConvNet algorithm exhibit certain prediction errors. These errors are also reflected in the predicted quality maps. The primary reason for that is the accuracy issue in multi-object generation; based on the network's multi-object prediction mechanism, the regions with high scores throughout the entire scene are visualized, and these regions may represent the edges of certain objects. In real robot grasping tasks, only poses with the highest confidence are assigned to the executing robot.

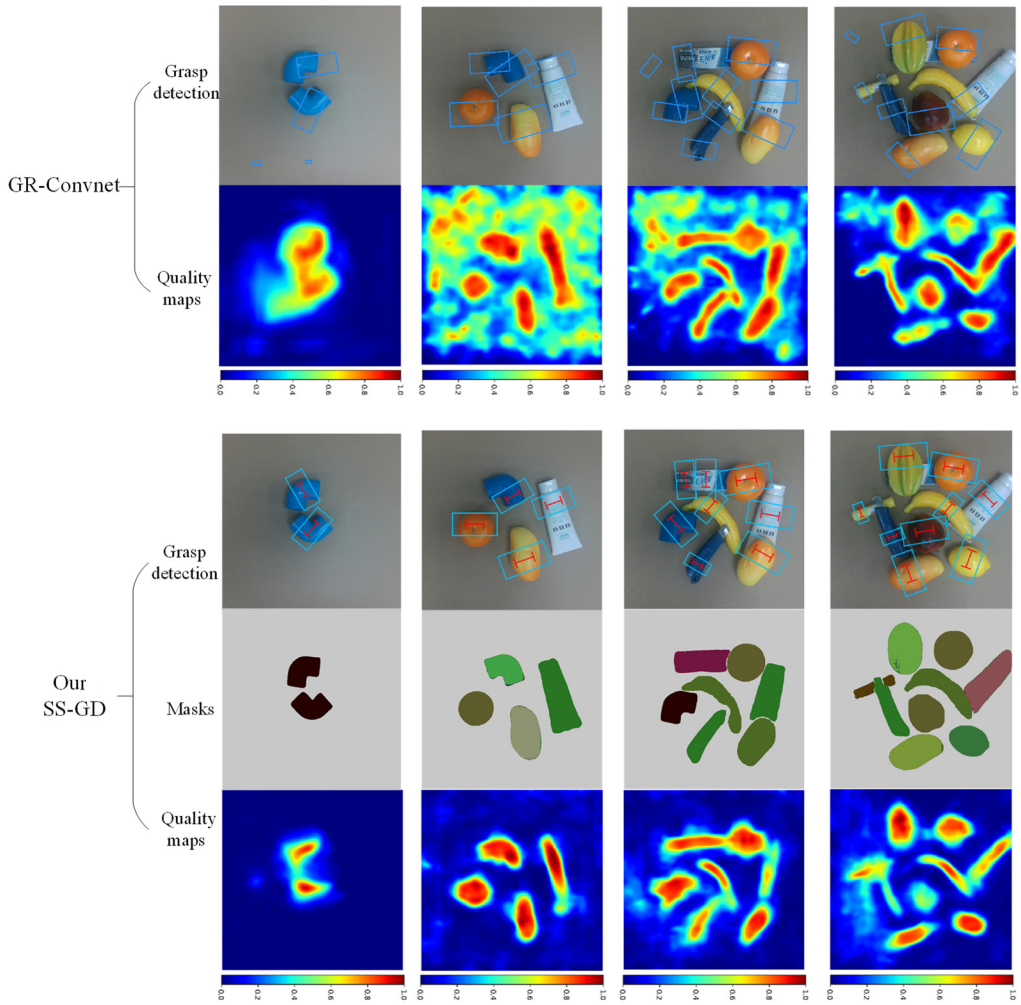

Figure 7. The grasp detection results for case 1.

However, in a global scene, the point with the highest confidence may not necessarily be the most suitable grasping point for a particular object. This could potentially lead to

unstable grasping operations, even though sometimes these operations may be successful. Our SS-GD algorithm utilizes segmentation masks to generate reliable grasp configurations, instilling the maximum confidence for each target. It is worth noting that the predicted grasping configurations almost fall within the geometric center of each object. This enhances the robustness and accuracy of the robot's grasping operations.

Case 2: As depicted in Figure 8, the grasp detection is performed in a novel workspace, which differs from the training dataset. It is evident that the GR-ConvNet algorithm [14] is adversely affected by the cluttered background, as the textures and wrinkles in the background are mistakenly identified as graspable objects. The cause of this phenomenon is that the grasp reasoning network, during training, extracts texture features of grasping objects, while the shape information of objects is neglected. This results in a well-performing network during training validation but a suboptimal performance in predicting in novel environments. In contrast, our SS-GD algorithm demonstrates robust performance by extracting explicit shape features provided by the segmentation network, offering accurate predictions for reasonable grasp configurations, especially in compact targets and novel multi-target cluttered environments.

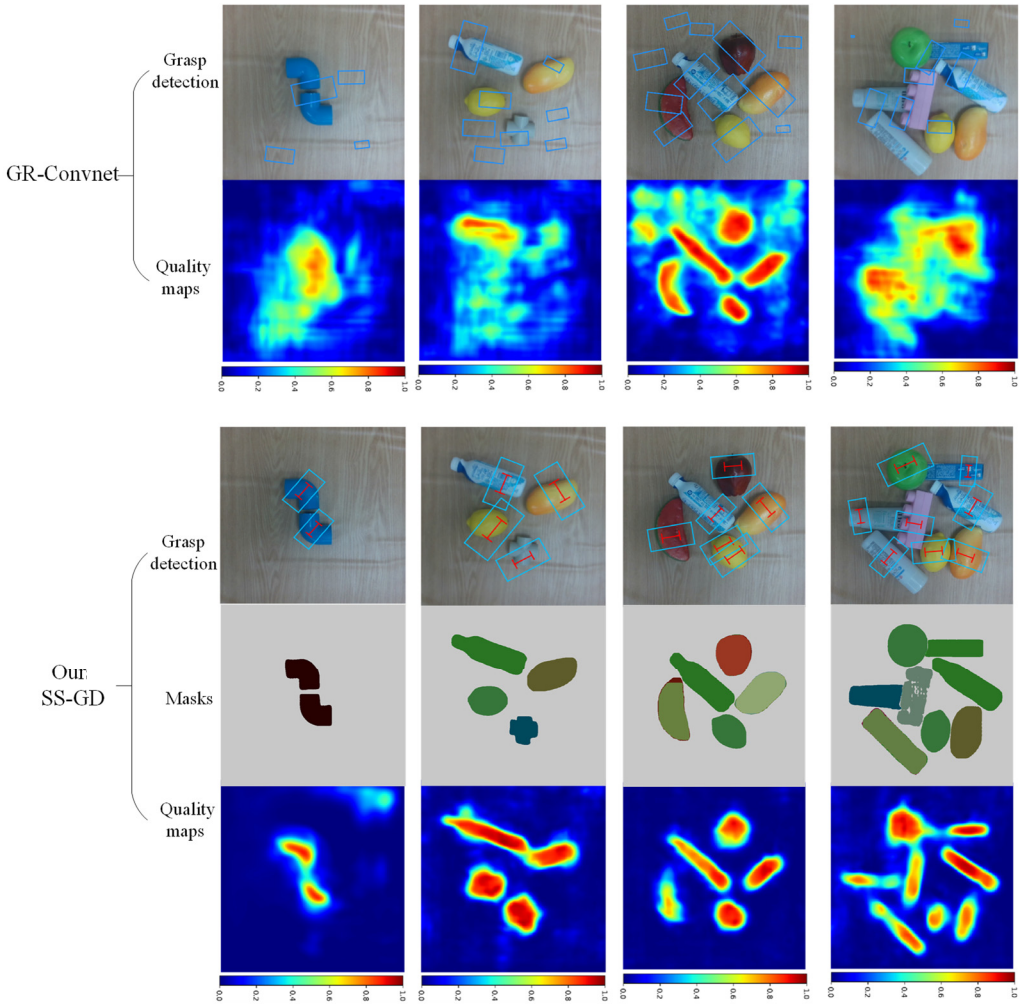

Figure 8. The grasp detection results for case 2.

Discussion: According to the above test results, the proposed SS-GD can accurately predict the grasp position and generate the corresponding grasp configuration at the local maximum. The accuracy of the semantic segmentation module at pixel level can better help the robot understand the position and shape information of objects. In the grasping

detection module, the attention mechanism is introduced to improve the feature extraction ability of the network model and then improve the accuracy of grasp detection.

5.4. Grasping Test on Robotic Manipulator

In this section, we define a grasp task with about 10 objects as light stacking, and a task with about 20 objects as heavy stacking. As shown in Figure 9, we use the UR5 6-DoF manipulator, Robotiq 2f-85 gripper, and Realsense D435i binocular depth camera to conduct the grasping experiment. The network model is constructed in the popular PyTorch platform, and trained using an RTX 3090 GPU with 24G memory. In addition, the test objects comprised 30 categories, including outdoor sports equipment, fruits, and industrial products. The proposed method was practically evaluated in stacked scenarios, where the robot engaged in a loop of grasping and picking until the objects were successfully cleared from the stack. In the robot grasping test, the calculation of grasp success rate SR is as follows:

$$SR = \frac{SG}{SG + FG} \tag{14}$$

where SG is the number of successful grasps, and FG is the number of failed grasps.

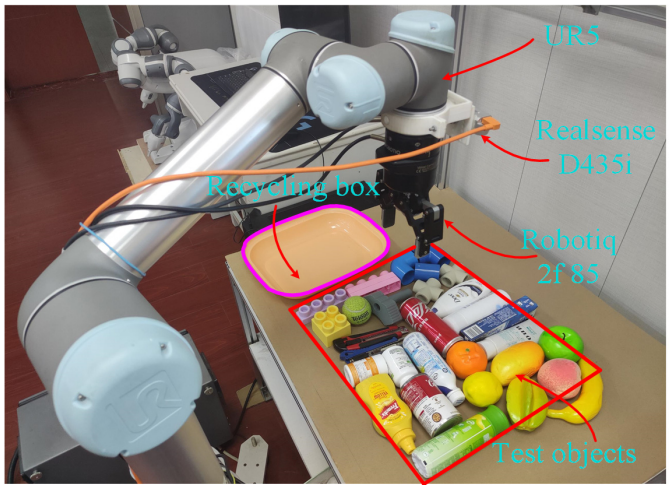

Figure 9. The UR5 robotic experimental platform and the test objects.

Figure 10 shows the experimental results as a showcase to demonstrate the effectiveness of our proposed approach. The first line illustrates the final predicted grasping configurations generated by the grasping detection module. These configurations are carefully sorted and filtered to eliminate redundancies resulting from multiple predictions. The second line displays the output of the masks by the semantic segmentation module, providing rich shape information and capturing stacked hierarchical relationships for the robot. The third line depicts the real robot executing grasping and picking-up tasks, meticulously adhering to the predicted grasping pose and closure degree through matrix transformations.

Our experiments were conducted in 10 stacked environments, comprising 100 grasping tasks. To clear the workspaces, the system attempted 109 grasps, in which the number of successful grasps was 100 and the number of failed grasps was 9; thus, the grasp success rate was 91.7%. From information acquisition to the robot's response, the entire system's response time does not exceed 1 s.

As summarized in Table 4, in contrast to multi-object cluttered scenarios, our proposed algorithm was evaluated on stacked scenarios and demonstrated outstanding performance in terms of grasp success rate. SS-DG significantly improves the performance of grasp accuracy compared with the existing methods. Although the objects were stacked, the proposed SS-DG effectively deal with those challenges. Thus, the RIGNet algorithm displayed the highest grasp success rate among the state-of-the-art methods.

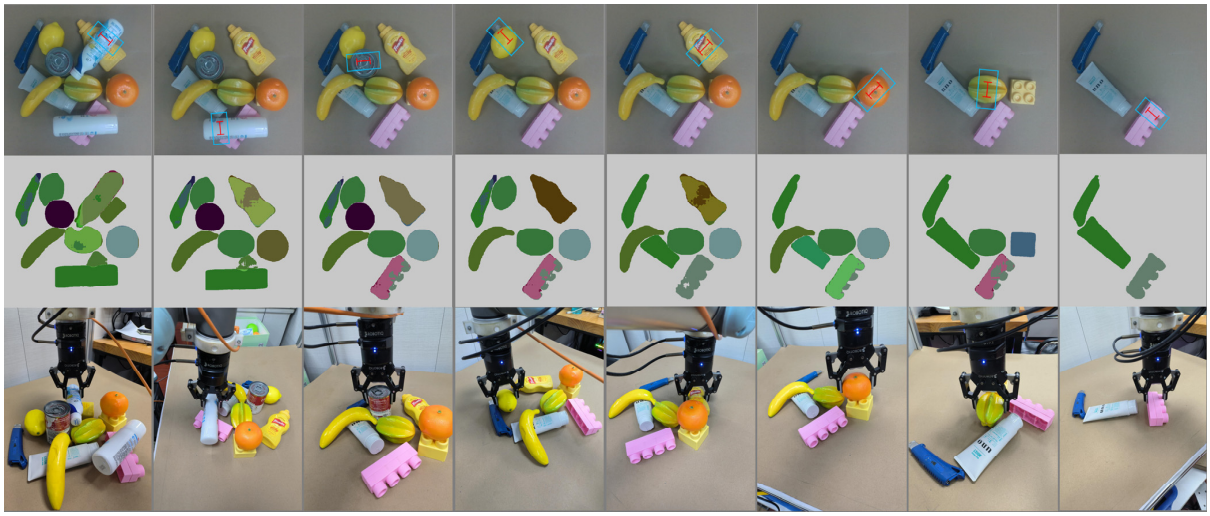

Figure 10. The robot grasping experiment of a lightly stacked scene.

Table 4. Comparison results with other algorithms.

Methods	Schema	Input	Grasp Success Rate (%)
Morrison [24]	GGCNN2	D	87.0
Liu [30]	Q-Net	RGB-D	90.2
Asif [17]	GraspNet	RGB-D	86.4
Zhang [31]	ROI-GD	RGB	83.7
Zhang [32]	MECNN	RGB	90.6
Park [33]	SMTNet	RGB-D	86.1
Ours	SS-GD	RGB-D	91.7

To assess the algorithm's robustness in handling complex grasping scenarios, we conducted more tests in scenarios containing 19 heavily stacked objects. As shown in Figure 11, after five rounds of testing, the algorithm achieved a grasp success rate of 79.8%. In these tests, the system made a total of 119 grasp attempts, resulting in 95 successful grasps and 24 failures. The results indicate that, despite the increased number of objects and corresponding growth in mask complexity, the robot's grasping effectiveness remains unaffected. Notably, in heavily stacked scenes, performance degradation is primarily attributed to a decrease in mask quality output by the segmentation network when dealing with adjacent objects. Additionally, as the robot approaches the target object for grasping, limitations arise when the gripper encounters obstruction from other objects, leading to insufficient space for parallel gripper maneuvering.

Figure 11. The robot grasping experiment for a heavily stacked scenario.

6. Conclusions

This paper proposes a novel deep network schema called SS-DG, which integrates semantic segmentation and grasp detection modules to enhance the success rate of robotic grasping in cluttered environments. Key innovations include the following:

(1) Incorporation of Swin Transformer: This component significantly improves object detection in scenarios with occlusion or stacking. By enabling the robot to disregard protrusions formed by stacked objects, Swin Transformer helps avoid unstable grasps that might otherwise occur from focusing on convex shapes created by the stacking of items.

(2) Introduction of the SE Attention Mechanism: This mechanism enhances the grasp detection network by predicting precise grasp poses for each object. It achieves this by combining object masks and depth maps, which helps prevent the generation of multiple detection boxes for a single object.

Extensive experiments were conducted, including grasping tests under light and heavy stacking conditions using the UR5 robot platform. The best grasping success rate achieved was 91.7%. Future work will involve using a dual-arm robot to grasp objects that exceed the range of a single gripper, thus extending the applicability of the SS-DG schema to tasks requiring fine manipulation.

Our research has a certain value in scenarios where robots are required to grasp stacked objects, and can significantly improve the grasp accuracy, but there are still some limitations that need to be overcome by further research. First, the current algorithm still faces some challenges in grasping objects with extreme overlap or occlusion. Secondly, the performance of the algorithm in processing objects with high reflection or low texture needs to be improved. In addition, the real-time performance of the algorithm also needs to be further optimized in more complex practical applications.

Author Contributions: Methodology, X.Z.; software, X.Z., J.L. and C.S.; validation, J.L. and C.S.; formal analysis, Y.C.; writing—original draft preparation, X.Z., J.L., C.S. and H.H.; writing—review and editing, Y.C., X.Z. and H.H.; visualization, Y.C. and X.Z.; supervision, H.H.; project administration, X.Z. and H.H.; funding acquisition, X.Z. and H.H. 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.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

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

References

1. Wang, J.; Lin, X.; Yu, H. Poat-net: Parallel offset-attention assisted transformer for 3D object detection for autonomous driving. *IEEE Access* **2021**, *9*, 151110–151117. [CrossRef]
2. Pan, X.; Xia, Z.; Song, S.; Li, L.E.; Huang, G. 3D object detection with Point former. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Kuala Lumpur, Malaysia, 19–25 June 2021; pp. 7463–7472.
3. Wang, C.; Li, C.; Han, Q.; Wu, F.; Zou, X. A Performance Analysis of a Litchi Picking Robot System for Actively Removing Obstructions, Using an Artificial Intelligence Algorithm. *Agronomy* **2023**, *13*, 2795. [CrossRef]
4. Ye, L.; Wu, F.; Zou, X.; Li, J. Path planning for mobile robots in unstructured orchard environments: An improved kinematically constrained bi-directional RRT approach. *Comput. Electron. Agric.* **2023**, *215*, 108453. [CrossRef]
5. Wu, Z.; Tang, Y.; Hong, B.; Liang, B.; Liu, Y. Enhanced precision in dam crack width measurement: Leveraging advanced lightweight network identification for pixel-level accuracy. *Int. J. Intell. Syst.* **2023**, *2023*, 9940881. [CrossRef]
6. Bohg, J.; Morales, A.; Asfour, T.; Kragic, D. Data-driven grasp synthesis—A survey. *IEEE Trans. Robot* **2014**, *30*, 289–309. [CrossRef]
7. He, Z.; Wu, C.; Zhang, S.; Zhao, X. Moment-based 2.5-D visual servoing for textureless planar part grasping. *IEEE Trans. Ind. Electron.* **2018**, *66*, 7821–7830. [CrossRef]

8. Yu, S.; Zhai, D.H.; Xia, Y. CGNet: Robotic Grasp Detection in Heavily Cluttered Scenes. *IEEE/ASME Trans. Mech.* **2023**, *28*, 884–894. [CrossRef]
9. 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.
10. Li, Y.; Kong, T.; Chu, R.; Li, Y.; Wang, P.; Li, L. Simultaneous semantic and collision learning for 6-DOF grasp pose estimation. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Prague, Czech Republic, 27 September–1 October 2021; pp. 3571–3578.
11. Jiang, Y.; Moseson, S.; Saxena, A. Efficient grasping from RGB-D images: Learning using a new rectangle representation. In Proceedings of the IEEE International Conference on Robotics and Automation, Shanghai, China, 9–13 May 2011; pp. 3304–3311.
12. Depierre, A.; Dellandréa, E.; Chen, L. Jacquard: A large-scale dataset for robotic grasp detection. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 3511–3516.
13. Morrison, D.; Corke, P.; Leitner, J. Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach. Robotics: Science and Systems (RSS), May 2018. Available online: https://arxiv.org/abs/1804.05172 (accessed on 15 May 2018).
14. Kumra, S.; Joshi, S.; Sahin, F. Antipodal robotic grasping using generative residual convolutional neural network. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Las Vegas, NV, USA, 24 October–24 January 2021; pp. 9626–9633.
15. Yu, S.; Zhai, D.-H.; Xia, Y.; Wu, H.; Liao, J. SE-ResUNet: A novel robotic grasp detection method. *IEEE Robot. Automat. Lett.* **2022**, *7*, 5238–5245. [CrossRef]
16. Wang, S.; Zhou, Z.; Kan, Z. When transformer meets robotic grasping: Exploits context for efficient grasp detection. *IEEE Robot. Autom.* **2022**, *7*, 8170. [CrossRef]
17. Asif, U.; Tang, J.; Harrer, S. GraspNet: An Efficient Convolutional Neural Network for Real-time Grasp Detection for Low-powered Devices. In Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18), Stockholmsmässan, Sweden, 13–19 July 2018; pp. 4875–4882.
18. Araki, R.; Onishi, T.; Hirakawa, T.; Yamashita, T.; Fujiyoshi, H. MT-DSSD: Deconvolutional single shot detector using multi-task learning for object detection, segmentation, and grasping detection. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Paris, France, 31 May–31 August 2020; pp. 10487–10493.
19. Xu, R.; Chu, F.J.; Tang, C.; Liu, W.; Vela, P.A. An affordance keypoint detection network for robot manipulation. *IEEE Robot. Autom.* **2021**, *6*, 2870–2877. [CrossRef]
20. Xie, C.; Xiang, Y.; Mousavian, A.; Fox, D. Unseen object instance segmentation for robotic environments. *IEEE Trans. Robot.* **2021**, *37*, 1343–1359. [CrossRef]
21. Ainetter, S.; Böhm, C.; Dhakate, R.; Weiss, S.; Fraundorfer, F. Depth-aware object segmentation and grasp detection for robotic picking tasks. *arXiv* **2021**, arXiv:2111.11114.
22. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
23. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Kuala Kuala Lumpur, Malaysia, 19–25 June 2021; pp. 10012–10022.
24. Morrison, D.; Corke, P.; Leitner, J. Learning robust, real-time, reactive robotic grasping. *Int. J. Robot. Res.* **2022**, *39*, 183–201. [CrossRef]
25. Fang, H.S.; Wang, C.; Gou, M.; Lu, C. Graspnet-1billion: A large-scale benchmark for general object grasping. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 11444–11453.
26. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2961–2969.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
28. Depierre, A.; Dellandréa, E.; Chen, L. Scoring grasp ability based on grasp regression for better grasp prediction. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 4370–4376.
29. Song, Y.; Gao, L.; Li, X.; Shen, W. A novel robotic grasp detection method based on region proposal networks. *Robot. Comput. Integr. Manuf.* **2020**, *65*, 101963. [CrossRef]
30. 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.* **2022**, *71*, 1–10. [CrossRef]
31. 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 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 4768–4775.

32. Zhang, H.; Lan, X.; Bai, S.; Wan, L.; Yang, C.; Zheng, N. A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Macau, China, 3–8 November 2019; pp. 6435–6442.
33. Park, D.; Seo, Y.; Shin, D.; Choi, J.; Chun, S.Y. A single multi-task deep neural network with post-processing for object detection with reasoning and robotic grasp detection. In Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), Xi'an, China, 30 May–5 June 2021; pp. 7300–7306.