actuators

MDPI

Article

Robot Three-Finger Grasping Strategy Based on DeeplabV3+

Qiang Bai [1], Shaobo Li [1,2,*], Jing Yang [2], Mingming Shen [1,3], Sanlong Jiang [1] and Xingxing Zhang [2]

[1] School of Mechanical Engineering, Guizhou University, Guiyang 550025, China; cme.qbai18@gzu.edu.cn (Q.B.); 15985146314@163.com (M.S.); gs.sljiang19@gzu.edu.cn (S.J.)
[2] State Key Laboratory of Public Big Data, Guizhou University, Guiyang 550025, China; jyang23@gzu.edu.cn (J.Y.); XingxingZHANG_Star@163.com (X.Z.)
[3] College of Mechanical and Electrical Engineering, Guizhou Normal University, Guiyang 550025, China
* Correspondence: lishaobo@gzu.edu.cn

Abstract: Researchers all over the world are aiming to make robots with accurate and stable human-like grasp capabilities, which will expand the application field of robots, and development of a reasonable grasping strategy is the premise of this function. In this paper, the improved deeplabV3+ semantic segmentation algorithm is used to predict a triangle grasp strategy. The improved model was trained on the relabeled Cornell grasp datasets and tested on self-collected datasets. Compared with the existing rectangular grasp strategy, the proposed algorithm and triangle grasp strategy have achieved outstanding performance in stability, accuracy, and speed. Finally, based on the ROS platform, this paper deploys the trained model and verifies the real effect of the trained grasping strategy prediction model, and achieves excellent grasping effect.

Keywords: semantic segmentation; grasp strategy; triangle; SPP; robot

Citation: Bai, Q.; Li, S.; Yang, J.; Shen, M.; Jiang, S.; Zhang, X. Robot Three-Finger Grasping Strategy Based on DeeplabV3+. *Actuators* **2021**, *10*, 328. https://doi.org/10.3390/act10120328

Academic Editor: Daniele Leonardis

Received: 18 October 2021
Accepted: 9 December 2021
Published: 12 December 2021

1. Introduction

Researchers all over the world aim to make robots achieve intelligent, human-like grasp capabilities, which will expand the application field of robots and create huge economic and social value. With the rapid development of deep learning and the improvement of camera sensor hardware, research on object recognition and location based on machine vision has made great progress [1], but there is less research on object grasp point detection, and it is mainly focused on rectangular grasp strategy [2–6]. Ian Lenz et al. [3] proposed a two-layer cascaded deep learning network to predict grasping strategy: The first deep network was used to quickly exclude the impossible grasp options; the second filtered the grasp strategy based on the first network and output the optimal value. The improved CNN proposed by Joseph Redmon et al. [6] had strong constraint processing, so that the model only needed to traverse the image once to achieve accurate grasp strategy, which greatly improved the running speed of the model. Sulabh Kumra et al. [7] used a residual network and a unique skip connection structure: the model achieved excellent object feature extraction and good results on the Cornell grasp dataset. Douglas Morrison [5] et al. proposed a GG-CNN model, which could directly generate grasp strategies from pixel-level depth images, overcoming other deep learning models which rely on sampling and classifying individual grasp candidates, resulting in long calculation times. D Avella [8] et al. designed a custom soft robotics end-effector and integrated it into a complete and autonomous robot grasping system in order to overcome the limitations of existing robots' grasping of objects in messy environments. The method achieved a success rate of 74.66% on objects of different difficulties, and had good generalization for new environments. From the above existing research results, it is found that research on object grasping points is relatively limited, and mainly concentrated in the field of rectangular grasping. This is because the existing research on grasping aims at relatively regular and simple objects (such as cylinders and cuboids), and a rectangular grasping strategy can be realized by a two-finger gripper with low cost. However, with the deepening of grasping research

and the improvement of people's expectations for robot grasping, the existing research methods are no longer suitable for objects with complex structures (such as GARAGE KIT and children's toys). In comparison to a human-like grasping experience, the two-finger rectangular grasping strategy demonstrates poor generalization and is unstable. In order to solve the above problems, in this paper, we researched object grasp point detection from the level of pixels. First, a semantic segmentation algorithm was used to accurately segment the object from the image. The semantic segmentation algorithm used in this paper was deeplabV3+ [9], proposed by Google in 2018, which achieved excellent results on VOC12. Second, the improved semantic segmentation algorithm was used to generate a triangle grasp strategy. The triangle grasp strategy has three grasp points, which greatly improves the stability of the grasp and is suitable for irregular objects.

The paper is divided into six parts: The first chapter summarizes and analyzes the advantages and disadvantages of the existing research results of object grasp strategy, and puts forward the innovation of this paper; the second chapter introduces the principle of the algorithm; the third chapter describes the improvement process, significance of the model, and the setting of experimental parameters; the fourth chapter describes the training process of the model; the fifth chapter comments the experimental results; and the sixth chapter is the conclusion.

2. Principal Analysis

Precise semantic segmentation is the premise of grasp strategy generation. Google's deeplabV3+ algorithm not only inherits the excellent performance of the previous algorithm on multi-scale objects, but also solves the problems of low prediction accuracy and boundary information loss caused by feature map resolution degradation resulting from multiple down-sampling of DCNN.

2.1. Atrous Convolution Kernel

Inserting a specific number of zeros into an ordinary convolution kernel is called atrous convolution, which is one of the core innovations of the deeplab algorithm. Multiple atrous revolution of different specifications can be connected in parallel to achieve excellent multi-scale information extraction ability. Receptive field is an important parameter of the semantic segmentation algorithm, and its main task is to calculate the number of nodes in the current layer that can feel the previous layer, which has an impact on the accuracy and speed of the model.

As shown in Figure 1a, after the stride of one-dimensional convolution is changed from 2 to 1, the receptive field clearly changes, and the receptive field of both is 3 when the output is (0, 1, 2, 3) and (0, 1, 2, 3, 4, 5, 6). It is found that (0, 1, 2, 3) in pool 1 corresponds to (0, 2, 4, 6) in pool 2, which shows that when the receptive fields are the same, other nodes are added to the model, which can make the feature map of the latter layer denser and allow it to contain more information.

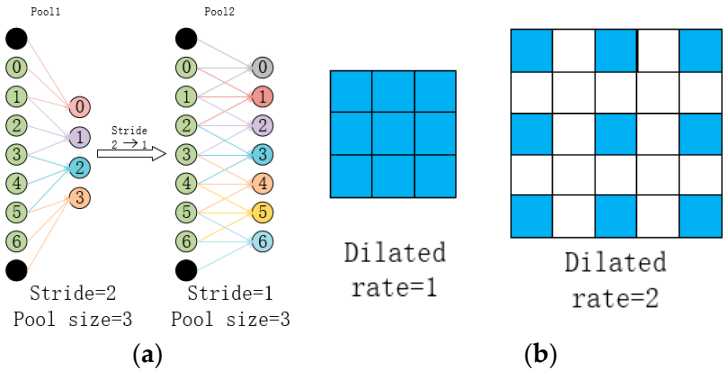


Figure 1. Receptive field and atrous convolution (**a**) the structure of receptive field; (**b**) atrous convolution kernel.

In order to keep the receptive field unchanged, the ordinary solid convolution layer filters need to be expanded, which is the origin of the atrous convolution algorithm. After using atrous convolution, the size of filter k changes as follows:

$$k = k + (k-1)(hole\_size - 1) \tag{1}$$

where k is the size of the original filter, *hole_size* is the dilated rate. Assuming that the original filter size is 3 and the dilated rate is 2, according to the formula (1), the expanded filter size is 5, and the expanded part is filled with 0 (Figure 1b).

In addition, the calculation of the model receptive field starts from the first layer of the input layer and then calculates in turn. Therefore, the receptive field size of the output feature map of the first layer convolution layer is equal to the size of the filter. The calculation formula of the model receptive field is as follows:

$$S_n = S_{n-1} * s \tag{2}$$

$$RF_n = RF_{n-1} + (k_n - 1) * S_{n-1} \tag{3}$$

where S_n is the strides of n layers in front of the network, and s is the stride of the current layer. RF_n is the receptive field (RF) of the upper layer, RF_{n-1} is the receptive field of this layer, k_n is the size of convolution kernel, and the effect of padding on the receptive field can be ignored. The receptive field of the n-th layer of the model can be quickly calculated through the above formula, which can not only improve the overall control, but also help to judge the rationality of the model structure.

2.2. Encoder and Decoder

In order to make the model have better multi-scale information fusion capability, deeplabV3+ combines the spatial pyramid pool (SPP) (Figure 2a) and encoder–decoder module used in deeplabV3 (Figure 2b), and then proposes a new atrous convolution encoder–decoder structure (as shown in Figure 2c), which well realizes the balance of accuracy and speed. The encoder part (Figure 2c left) realizes the extract of high-level feature semantic information through continuous down-sampling, and takes into account the semantic information of different size objects through the SPP, which greatly improves the multi-scale information perception ability. The decoder module (Figure 2c right) restores the image to the original size by up-sampling while retaining the boundary information, which solves the problem of boundary information loss in the semantic segmentation model.

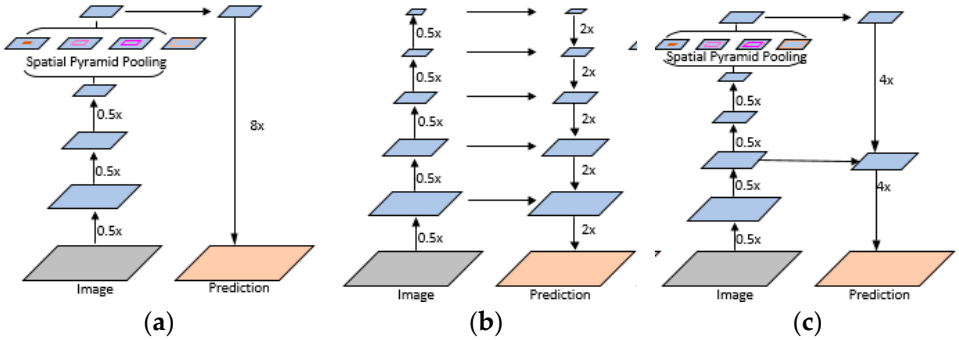


Figure 2. Schematic diagram of decoder–encoder (**a**) spatial pyramid pooling; (**b**) encoder-decoder structure; (**c**) atrous spatial pyramid pooling.

2.3. Overall Structure of DeeplabV3+

Figure 3 shows the overall structure of deeplabV3+ in detail. The encoder is used to extract the semantic information contained in the high-level features of the image. The atrous convolution module not only realizes image feature extraction and multi-scale context information acquisition, but also replaces the down-sampling module, which

makes the output stride of the feature map 16. The main function of the decoder module is to extract the boundary information from the low-level features of the image.

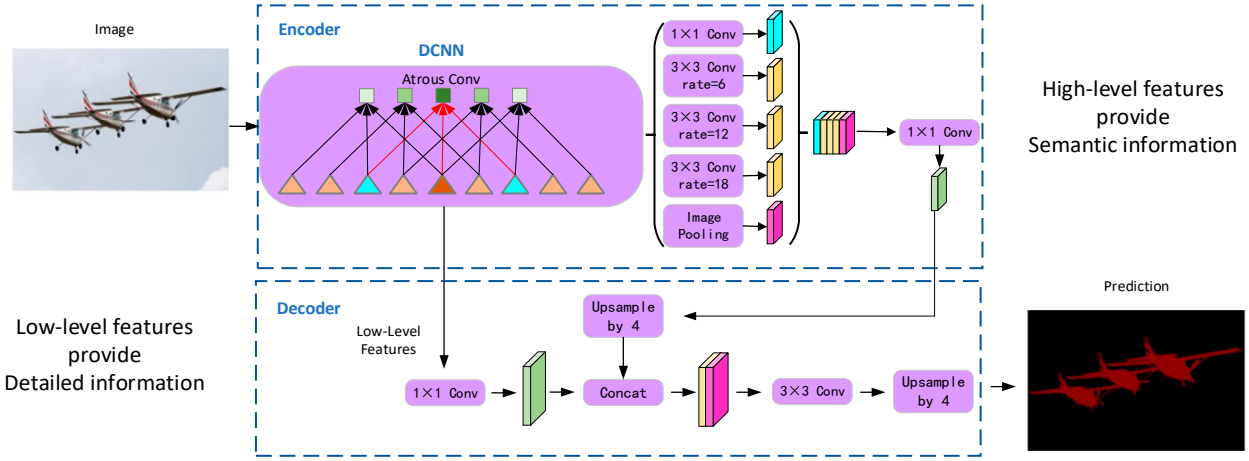


Figure 3. The overall structure of deeplabV3+.

3. Improvement of Model Structure

Deeplab algorithm is mainly used in the field of semantic segmentation, so it needs to be improved before used in the research of object grasp strategy. Firstly, the detection of grasping points is a multi-output task, including grasp confidence value, grasp angle and grasp width. As shown in Figure 4, after input RGB, the model needs to output three values: grasp confidence value, grasp angle and grasp width, and realize the output of the optimal grasping strategy based on them. Secondly, it can be seen from Figure 3 that the output of the model needs to be modified to three before the output of the grasp strategy can be realized and Figure 5 shows the improved model structure in detail.

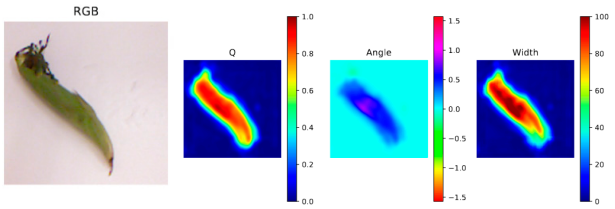


Figure 4. Basic parameters of grasp.

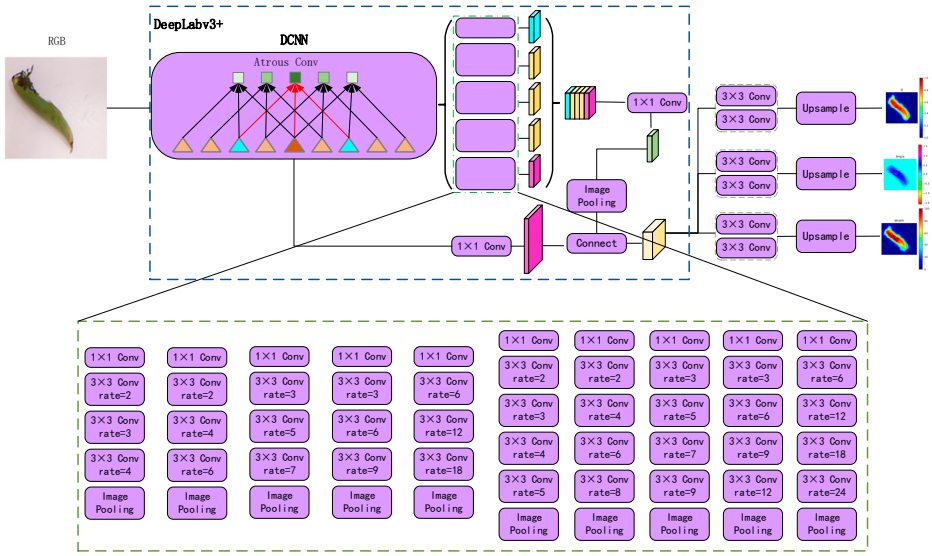


Figure 5. Improved deeplabV3+ structure.

As shown in Figure 2, SPP is the core of the algorithm. It not only plays the role of feature extraction, but also as the core innovation of deeplab. At present, there is no systematic mathematical theoretical basis for SPP structure, but the setting of atrous value and the number of convolution kernels are the key factors to determine the performance of SPP, therefore, this paper will optimize the model from these two aspects. After summarizing the relevant references [10–15], it is found that the current research has personalized and transformed the atrous value and the number of convolution kernels for their respective fields, so as to improve the matching degree between the model and the application field. After specific analysis, the following principles are summarized:

(1) The SPP with large and small atrous value is adopted on high-resolution image and low-resolution image, respectively, which can alleviate the grid effect and ensure the ability of the model to obtain multi-scale object information at the same time;
(2) Sawtooth structure and loop structure have their own advantages and disadvantages. In principle, the complexity is reduced as much as possible while ensuring the performance of the model;
(3) The difference between atrous values should not jump too large, which basically meets the distribution of arithmetic sequence.

The above analysis shows that it is necessary to design different specifications of SPP according to different tasks. Therefore, this paper designs and tests a variety of SPP structures in order to develop a model more suitable for triangle grasping strategy prediction. On the other hand, in order to modify the output to three indicators for evaluating the grasp strategy, the last convolution layer and up-sampling layer of the model are deleted (as shown in Figure 5), and the deeplabV3+ is used as a feature extractor. When images are input into the model, the feature extractor will output several feature maps of specific size, and these feature maps are input into the grasp predictor, the grasp confidence value, grasp angle and grasp width will be output. The prediction part consists of two $3 * 3$ convolution kernels and up-sampling. The predictor predicts the center position of the triangle by determining whether the location of the point can be grasped, and the grasp angle and width predictor is used to refine the size and direction of the triangle.

The prediction of grasping confidence value is a binary classification problem: we use the softmax cross-entropy as the loss function. The grasping angle is a multi-object and multi-classification problem: we use the sigmoid cross-entropy as the loss function. The prediction of the grasping width is a regression problem, and we use the mean square error function as the loss function.

Kullback–Leibler (KL) divergence can be used to measure the difference between two distributions, which can be described as formula (4).

$$D_{KL}(p \parallel q) = \sum_{i=1}^{n} p(x_i) \log\left(\frac{p(x_i)}{q(x_i)}\right) \tag{4}$$

where n is all the possibilities of the event, p is the label value, and q is the predicted value. The smaller the value of D_{KL}, the closer the q and p distribution are.

According to formula (4), we can get formula (5):

$$\begin{aligned} D_{KL}(p \parallel q) &= \sum_{i=1}^{n} p(x_i) \log(p(x_i)) - \sum_{i=1}^{n} p(x_i) \log(q(x_i)) \\ &= -H(p(x)) + \left[-\sum_{i=1}^{n} p(x_i) \log(q(x_i))\right] \end{aligned} \tag{5}$$

where $-H(p(x))$ represents the entropy of P.

Based on formula (5), we can get the cross entropy:

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) \log(q(x_i)) \tag{6}$$

Then, softmax cross-entropy can be described as the following formulas:

$$S_j = \frac{e^{a_j}}{\sum_{k=1}^{T} e^{a_k}} \tag{7}$$

$$L = -\sum_{j=1}^{T} y_i \log s_j \tag{8}$$

where L is the loss value and S_j is the j-th value of the softmax output vector S.

Sigmoid cross-entropy can be described as the following formula (9):

$$\begin{aligned} & Z * (-\log(sigmoid(x))) + (1 - Z) * (-\log(1 - sigmoid(x))) \\ & = Z * (-\log(1/(1 + \exp(-x)))) + (1 - Z) * (-\log(\exp(-x)/(1 + \exp(-x)))) \\ & = Z * \log(1 + \exp(-x)) + (1 - Z) * (-\log(\exp(-x) + \log(1 + \exp(-x)))) \\ & = Z * \log(1 + \exp(-x)) + (1 - Z) * (x + \log(1 + \exp(-x))) \\ & = (1 - Z) * x + \log(1 + \exp(-x)) \\ & = x - x * Z + \log(1 + \exp(-x)) \end{aligned} \tag{9}$$

where x is the predicted value and Z is the label value.

Mean square error function can be described as the following formula (10):

$$MSE(y, y') = \frac{\sum_{i=1}^{n} (y_i - y_i')^2}{n} \tag{10}$$

where y_i is the label value and y_i' is the predicted value.

In order to compare the accuracy of different SSP structures in grasping strategy, this paper designed different structure and parameter experiments to systematically train the model and comprehensively compare its performance, as shown in Table 1.

Table 1. Experimental parameters distribution.

SPP_Number	**4**				**5**				
Dilated_rate1-2-3-4	1-2-4-6	1-3-5-7	1-3-6-9	1-6-12-18	1-2-3-4-5	1-2-4-6-8	1-3-5-7-9	1-3-6-9-12	1-6-12-18-24

4. Experiment

4.1. Introduction to Grasp Strategy

As shown in Figure 6, the triangle grasp representation with fixed orientation is inspired by human grasping behavior. The human will first consider a reasonable grasping position before determining the grasping posture of the hand. In the plane image, the location of the grasping object is represented by pixel coordinates, and then the hand posture in the grasping process is mapped to the grasping angle and the unfolding width of the gripper. Therefore, a real and reasonable plane grasp representation should take pixels as the core, including grasp angle and width. As shown in Figure 6a, the triangle grasp strategy is expressed as:

$$G = (x, y, \omega, \theta, d) \tag{11}$$

where (x, y) is the center coordinate of the height of the bottom edge of the triangle, ω is the height of the triangle, θ is the angle between the height of the triangle and the horizontal line, and d is the width of the bottom edge of the triangle. Compared with the traditional rectangular grasping method [2–6], the triangular grasping representation has two advantages: It has higher grasping stability on complex contour objects; Deeplab algorithm can provide more accurate label and prediction accuracy.

(**a**) (**b**)

Figure 6. Schematic diagram of triangle grasp strategy (**a**) principle; (**b**) physical display [16].

4.2. Introduction of Datasets

The grasp datasets used in this paper is Cornell grasp dataset, which has 240 objects, 885 RGB images and depth images. Because the three finger grasp strategy in this paper is quite different from the rectangular grasp strategy, this paper only uses the image data, and the dataset relabeled by Wang Dexin et al. [17] is used for the training of the model (as shown in Figure 7).

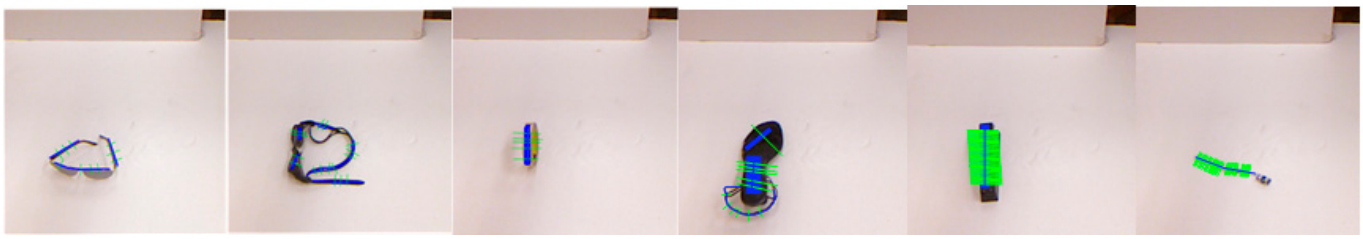


Figure 7. Relabeled triangle grasp strategy.

Figure 7 shows some visualized label data. For the convenience of labeling, the triangle is not labeled directly, but represented by a fixed length segment with direction and endpoint. All points in the blue area are grasping points. A green line is drawn with each grasping point as the end point and the included angle between the green line and the horizontal line is the grasping angle. The length of the green line is half of the grasping width. For cases where symmetrical grasping is not possible, only one green line is drawn, and two green lines in opposite directions represent symmetrical grasping.

4.3. Experimental Environment

The training environment is the 64 bit of Ubuntu 18.04, which adopts the pytorch deep learning framework. The hardware configuration is: core i9-9900x, RAM 128 GB, NVIDIA GeForce RTX2080Ti*2.

5. Result and Discussion

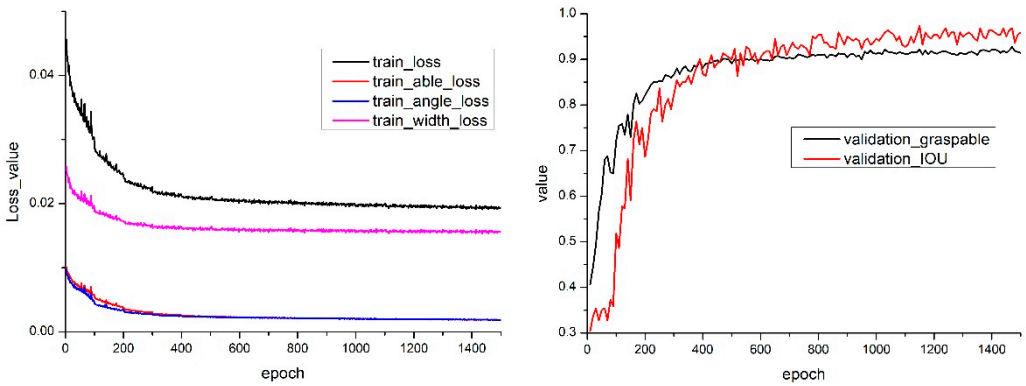
This chapter will verify the improved grasp strategy generation algorithm proposed in Chapter 3. Figure 5 shows the improved model structure, and the experimental parameters are shown in Table 1. The pytorch framework was developed by the Facebook team and opened source on GitHub in 2017. Pytorch has the advantages of simple framework, easy to use, support for dynamic calculation graph and high memory utilization.

5.1. Training Process Analysis

There are few studies on triangular grasp strategy, and even less on the generation of grasp strategy by improving deeplabV3+. Based on the above situation, this paper sets a set of traditional solid convolution kernels ablation experiment as a baseline to facilitate comparison with other experiments.

It is found from the curves in Figure 8 that the loss value of the model decreases gradually with the increase of the number of iterations, and there is no gradient explosion or violent oscillation, which shows that the training of the model is normal. The overall loss value is calculated as follows:

$$\text{Train\_loss} = \text{train\_able\_loss} + \text{train\_angle\_loss} + \text{train\_width\_loss} \quad (12)$$



(**a**) (**b**)

Figure 8. Experimental results of traditional solid convolution kernel (**a**) Overall loss curve; (**b**) Performance curve on test set.

From the formula (12), it is found that the overall loss is composed of three parts: confidence value loss, angle loss and width loss. The loss value of width is significantly higher than the other two items. This is because the determination of grasp width needs to convert the pixel width in the image into metric width, so there is a large error in model training.

According to the 10 groups of parameter structures in Table 1, this paper makes a systematic experiment on the model, and visualizes the training and test curves of the six groups of models with the best performances, as shown in Figures 9 and 10. Figure 9 is the model performance curve under four convolution kernels and it can be seen from the loss curve that the training process of the model is relatively stable. By observing the prediction curve (Figure 9b,d,f), it is found that the accuracy of the model is gradually increasing with the increase of the number of iterations, which shows that the proposed model structure is reasonable and there is no over-fitting. On the other hand, it is found that the coincidence rate of IOU and grasp confidence value is not high, and the IOU value is gradually higher than the grasp confidence value with the increase of iterations. This is because the grasp confidence value is generated on the basis of IOU, so it is normal that it is slightly lower than the IOU.

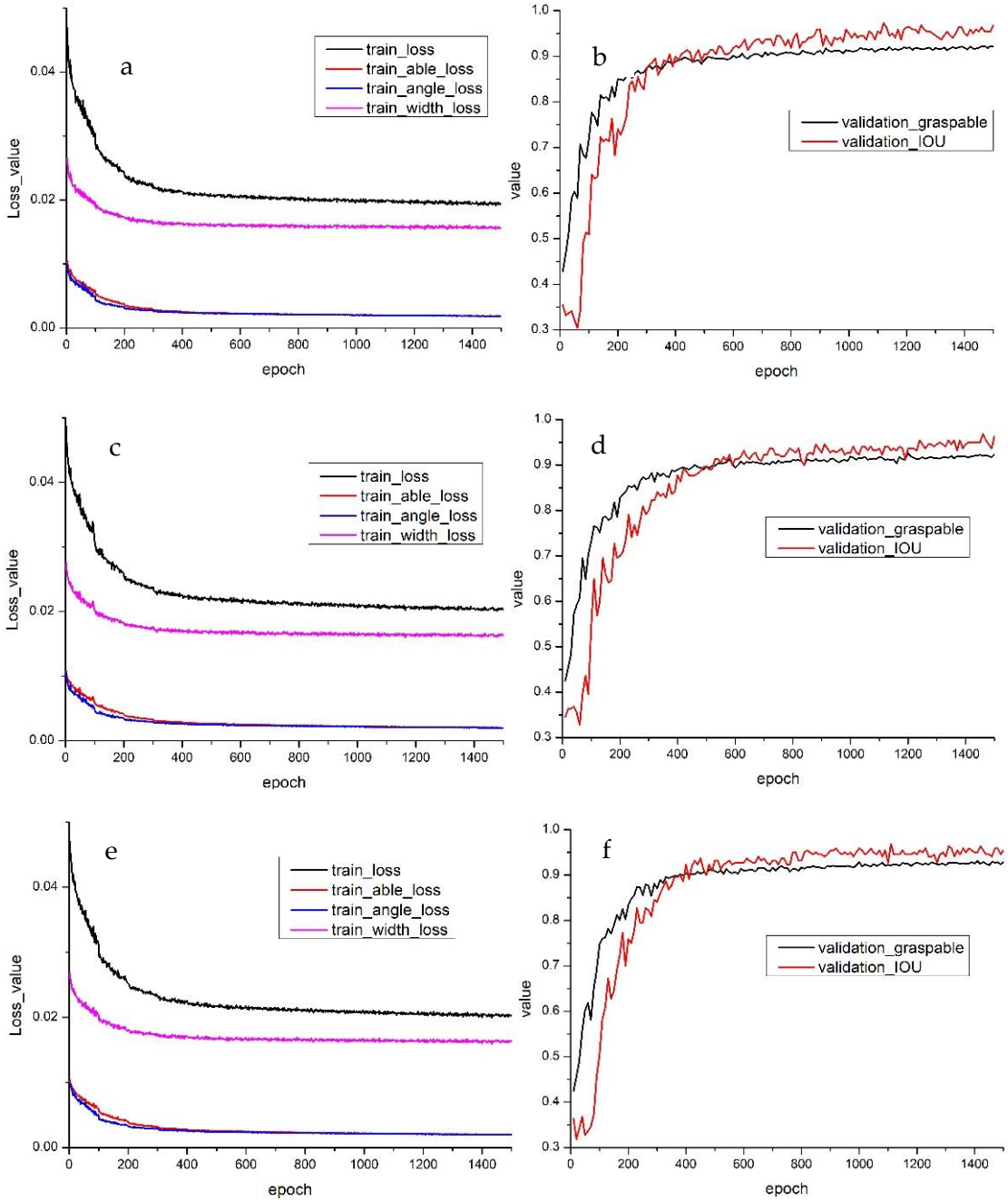


Figure 9. Training and prediction curves of 4 convolution kernel: (**a**) loss curve with dilated rate of 1-2-3-4; (**b**) performance prediction curve with dilated rate of 1-2-3-4; (**c**) loss curve with dilated rate of 1-2-4-6; (**d**) performance prediction curve with dilated rate of 1-2-4-6; (**e**) loss curve with dilated rate of 1-3-6-9; **(f)** performance prediction curve with dilated rate of 1-3-6-9.

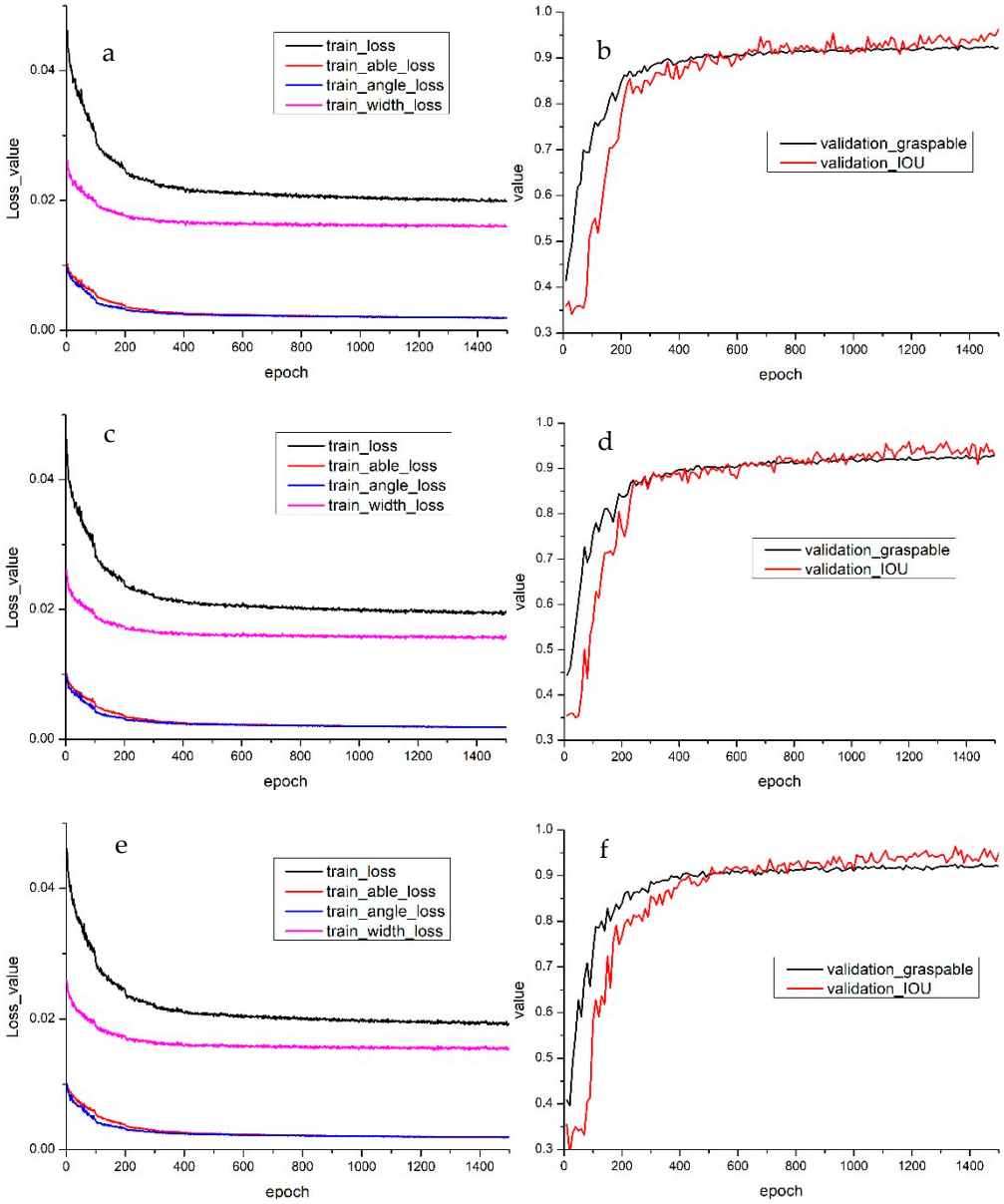


Figure 10. Training and prediction curves of five convolution kernel (**a**) loss curve with dilated rate of 1-2-3-4-5; (**b**) performance prediction curve with dilated rate of 1-2-3-4-5; (**c**) loss curve with dilated rate of 1-2-4-6-8; (**d**) performance prediction curve with dilated rate of 1-2-4-6-8; (**e**) loss curve with dilated rate of 1-3-5-7-9; (**f**) performance prediction curve with dilated rate of 1-3-5-7-9.

Figure 10 is the model performance curve under 5 convolution kernels, and the overall loss curve is similar to Figure 9, which shows that there is no over-fitting phenomenon with the increase of model complexity. With the increase of the number of iterations, the prediction curve (Figure 10b,d,f) is also gradually increasing, but the overall accuracy is similar to or even slightly insufficient with the structure of four convolution cores, which indicates that the improvement of complexity does not further improve the performance of the model. In addition, under the five convolution kernels, the coincidence degree of IOU and grasping confidence value curve increases significantly, which shows that with the increase of model complexity, the influence of IOU on grasping confidence value increases, while the influence of other parameters decreases.

In order to more objectively evaluate the performance of the model under different parameter structures, the important performance indicators are summarized in Table 2. It can be seen from the data in the Table 2 that the loss values of the models under different

data structures have little difference, but the specific analysis shows that the loss value of 1-3-5-7-9 is the smallest, because this structure has five parallel atrous convolution kernels, the parameter interval is set reasonably and relatively prime. On the other hand, it is found that the IOU and grasp value of the model on the test set are very close, because the model has achieved good performance in IOU and grasp value after training. The last two column of the Table 2 shows the prediction accuracy of the model on the test set, which is the most important index to evaluate the performance of the model. It can be seen that the performance of different models varies significantly and the parameter structure with the best performance is 1-2-4-6, followed by 1-2-3-4-5 and 1-3-5-7-9. The 1-6-12-18 structure adopted by deeplabV3+ paper does not achieve good prediction performance, which shows that this structure is not suitable for the prediction of grasp strategy, and also shows the importance of this study.

Table 2. Performance comparison under different parameter structures.

Structure	Loss	Loss_Able	Loss_Angle	Loss_Width	Valida Tion_IOU	Validation Graspable	Accuracy_ Image_Wise	Accuracy_ Object_Wise
1-2-3-4	0.01923	0.00184	0.00187	0.01552	0.97	0.920935	92.39%	90.25%
1-2-4-6	0.02021	0.00200	0.00202	0.01618	0.97	0.923298	97.83%	97.04%
1-3-6-9	0.02030	0.00198	0.00196	0.01636	0.97	0.928626	94.57%	93.27%
1-2-3-4-5	0.01995	0.00197	0.00195	0.01604	0.96	0.922368	95.65%	94.67%
1-2-4-6-8	0.01973	0.00187	0.00187	0.01600	0.96	0.927294	88.89%	86.76%
1-3-5-7-9	0.01918	0.00191	0.00192	0.01535	0.96	0.919595	95.65%	96.98%
1-6-12-18 [9]	0.02026	0.00201	0.00198	0.01627	0.95	0.919623	92.39%	92.05%

The SPP structure proposed by deeplab algorithm realizes excellent multi-scale information grasp function by using parallel atrous convolution of different dilated rate. The specific principle is as follows:

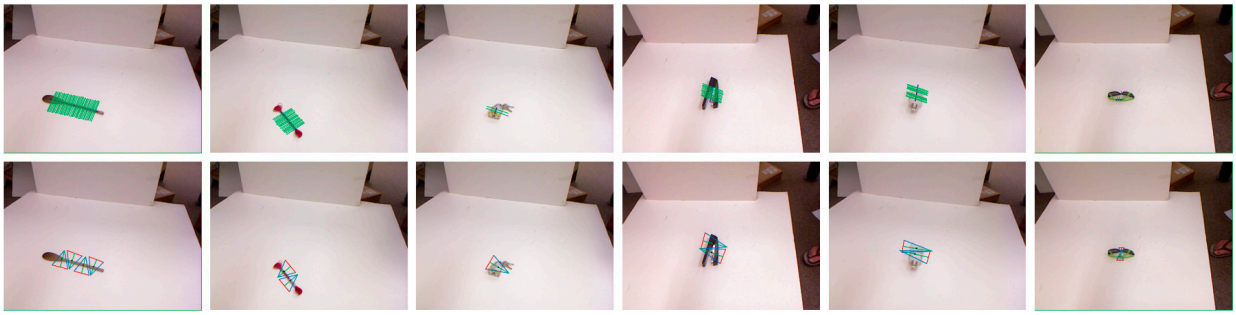
$$y[i] = \sum_{k} x[i + r \bullet k]\omega[k] \tag{13}$$

where r represents the stride of the input signal sampling, that is, the input x is convoluted with the up-sampling filters obtained by inserting r-1 zeros along each spatial dimension between two continuous filters.

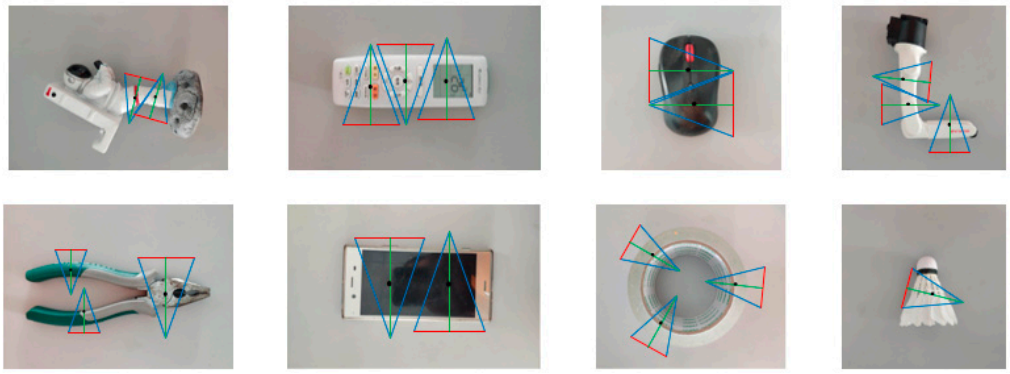
By changing the dilated rate of atrous convolution, model can well control the receptive field of the model and adjust the compactness of the model. From formula (6), it can be seen that large or small dilated rate will affect the model's ability to extract semantic information, resulting in poor performance in the grasp strategy, which also explains the reason why the 1-2-4-6 parameter structure has the best performance.

5.2. Visualization of Data

Figure 11a shows the grasp effect of the model on Cornell datasets, and it can be found that since the triangular grasping strategy is more stable than the rectangular strategy, it is considered a reasonable grasping option even when the grasping position is located at the object edge away from the center of gravity, which greatly improves the generalization performance. Figure 11b shows the prediction results of the model on the self-collected datasets and it is found that the model has good generalization performance on different datasets.



(**a**)



(**b**)

Figure 11. Visualization on different datasets: (**a**) Visualization on Cornell University datasets; (**b**) Visualization on self-collected datasets.

Because most of the current research on grasping strategy is aimed at the rectangular grasping strategy of two fingered dexterous hand, this paper selects the representative research results in this field for horizontal comparison with the results of triangle. It can be seen from the Table 3 that with the rapid development of the algorithm, the accuracy of the grasping strategy is getting higher and higher. Due to there are uncontrollable errors between the prediction accuracy of the algorithm and the execution accuracy of the robot, only by obtaining high accuracy at the model level can the high accuracy of the robot be guaranteed to a certain extent, which also explains the importance of the accuracy of the algorithm. Further analysis shows that the existing grasp strategy research is mainly focus on the rectangular based on CNN, which is completely different from the triangle grasp strategy based on semantic segmentation proposed in this paper. Because the semantic segmentation algorithm can accurately segment the object from the image and generate the grasp strategy, although the triangular grasp strategy is more complex than the rectangular grasp strategy, it still achieves advanced prediction performance. Table 3 show that the method proposed in this paper is superior to the current mainstream methods in accuracy and speed.

Table 3. Performance comparison of different algorithms.

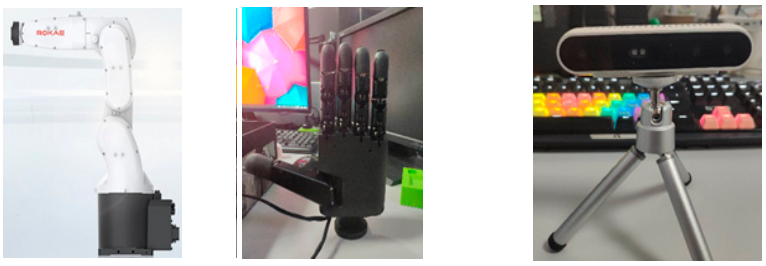
Author	Year	Representation	Datasets	Algorithm	Accuracy (%)		Speed (ms)
					IW	OW	
Yun Jiang [18]	2011	Rectangle	Self-made	Two-step proces	60.5	58.3	5000
Ian Lenz [3]	2013	Rectangle	Cornell grasp dataset	A two-step cascaded system	88.4	88.7	—
Joseph Redmon [6]	2015	Rectangle	Cornell grasp dataset	Single-stage regression	88.0%	87.1%	76
Sulabh Kumra [7]	2017	Rectangle	Cornell grasp dataset	ResNet-50 * 2	89.2	88.9	16
Di Guo [2]	2017	Rectangle	Cornell grasp dataset	Hybrid architecture	93.2	89.1	—
Fu-Jen Chu [19]	2018	Rectangle	Cornell grasp dataset	ResNet-50	96.5	96.1	20
YULIN XU [4]	2019	Oriented diameter circle	Cornell grasp dataset	GraspCNN	96.5%	—	50
Douglas Morrison [5]	2020	Rectangle	Cornell grasp dataset	GG-CNN	88.0%	—	20
Wang Dexin [17]	2020	Triangle	Cornell grasp dataset	SGDN	96.8%	92.3%	19
Ours	**2021**	**Triangle**	Cornell grasp dataset	**SSGP**	**97.83%**	**97.04%**	**19**

* This represents the concatenation of the two ResNet-50 into a whole. Bold: This line is the accuracy of the model proposed in this paper.

5.3. Application Verification

5.3.1. Platform Introduction

Figure 12 shows the hardware platform of the experiment: Figure 12a is a 6-DOF robot for trajectory operation of object grasping; Figure 12b is a five-fingered dexterous hand, in which the five fingers can be controlled independently. Therefore, the thumb, index finger and middle finger are used as the actuator of the triangular grasping strategy; Figure 12c is the depth camera, which is used for image acquisition and object positioning to provide position information for the grasping of the dexterous hand.



(**a**) (**b**) (**c**)

Figure 12. Experimental hardware platform. (**a**) 6-DOF robot; (**b**) dexterous hand; (**c**) depth camera.

5.3.2. System Architecture

Robot object grasp involves a series of operations, such as image acquisition, model loading, trajectory planning, and dexterous hand execution, so it is a relatively complete recognition and control system. Figure 13 shows the experimental framework of this paper in detail, and the picture in the upper left is the experimental environment. The system is mainly composed of four parts: Firstly, the RGB and depth information of the image is collected by the depth camera and uploaded to the ROS platform for processing; secondly, the ROS platform gets the accurate position of the object based on the collected image, and then predicts a reasonable grasp strategy based on the trained model; thirdly, ROS will plan an appropriate trajectory for the robot to approach the object to be grasped; Finally,

ROS sends grasping instructions to the dexterous hand to complete the grasping. The whole process is scheduled by the ROS platform deployed on one computer, which has good application value.

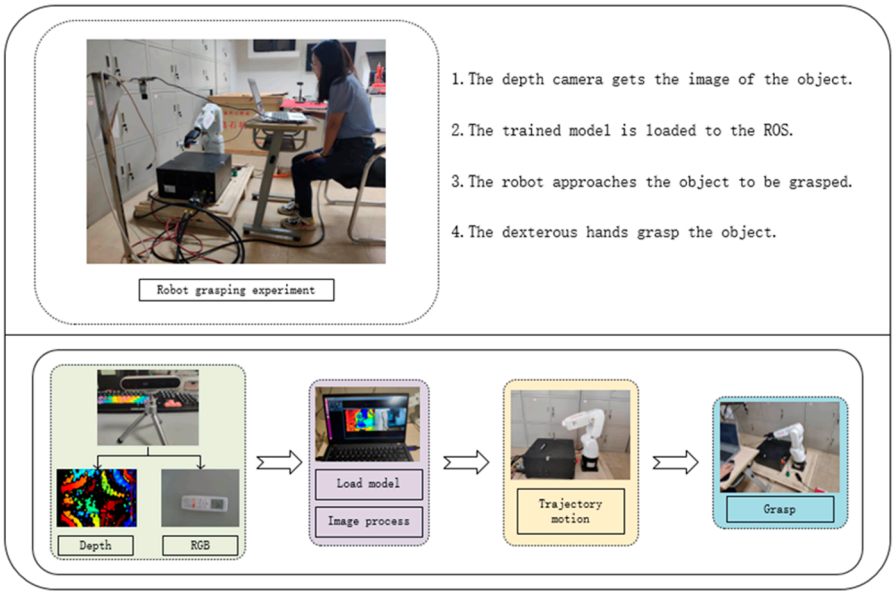


Figure 13. Overall framework of robot grasp.

5.3.3. Result Analysis

In this paper, 10 kinds of common objects (household objects) and five kinds of uncommon objects (adversarial objects) in life are used for 10 repeated grasping experiments respectively to count the success rate (Figure 14). Common objects have regular shapes and no obvious dents or protrusions on the surface (Figure 14a) and uncommon objects are rare objects in life, with irregular shapes and obvious depressions or protrusions on the surface (Figure 14b). Because the traditional two finger gripper has only two force points, once the gripping position is selected at the dents or protrusions, the object will probably slide down, resulting in grasping failure. Therefore, the research of two fingers grasping are mostly common objects.

(**a**) (**b**)

Figure 14. Objects to be grasped. (**a**) Common objects, (**b**) uncommon objects.

The robot carried out a total of 150 grasps, of which 140 were successful, and the success rate was 93.3%. The success rate on common objects were 95%, and the accuracy rate on uncommon objects were 90%.

Figure 15 shows the objects to be grasped for the experiment and the actual grasping effect. It is found that the three-fingered dexterous hand can grasp both common and uncommon objects stably, which shows that the trained model has good generalization performance. However, there are also grasping failure cases, which is mainly due to the sliding after grasping caused by the smooth surface of the object and the unreasonable grasping strategy predicted by the model. In order to further highlight the excellence of the research of this paper, Table 4 comprehensively compares the experimental results of this

paper with the existing research results. It is found that the triangle grasping method based on semantic segmentation proposed in this paper has achieved advanced performance on a variety of different objects.

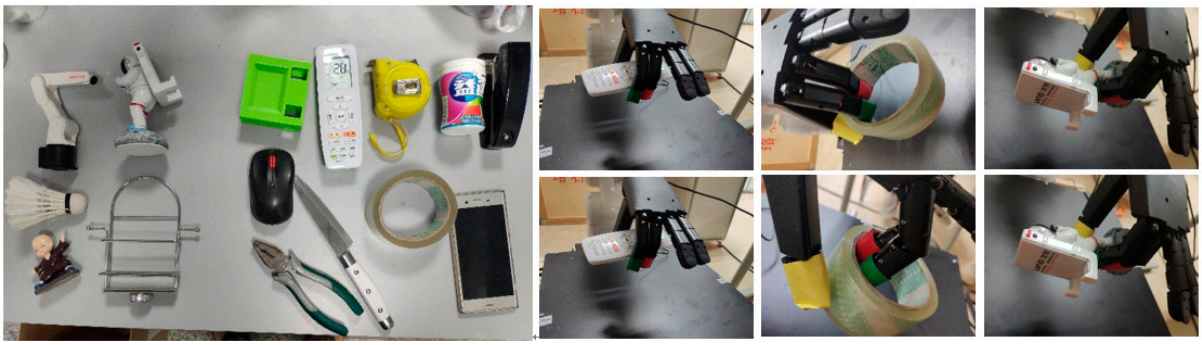


Figure 15. Object to be grasped (**left**) and successfully grasp sample (**right**).

Table 4. Comparison of robot grasping experimental results.

Author	Accuracy on Common Objects (%)	Accuracy on Uncommon Objects (%)	Overall Accuracy (%)	Two/Three Fingers	Year
Ian Lenz [3]	89 (89/100)	-	89	Two fingers (rectangle)	2015
Pinto Lerrel [20]	73 (109/150)	-	73	Two fingers (rectangle)	2015
Na Yong-Ho [21]	72	69	70.5	Two fingers (rectangle)	2017
Chu Fu-Jen [19]	89 (89/100)	-	89	Two fingers (rectangle)	2018
Morrison [5]	92 (110/120)	84 (67/80)	88.5	Two fingers (rectangle)	2019
Shang Weiwei [22]	92 (276/300)	-	92	Five fingers (rectangle)	2020
Ours	95 (95/100)	90 (45/50)	93.3	Three fingers (triangle)	2021

Aiming at the problem of low stability of the traditional rectangular grasp strategy, a triangle grasp strategy based on deeplabV3+ semantic segmentation algorithm is proposed, trained on Cornell grasp dataset and tested on self-collected dataset. Then, the trained model is deployed on the robot platform for application verification.

6. Conclusions

Accurate and reasonable grasp strategy is the premise to achieve object grasp, so it has important research significance. This paper studies the triangle grasp strategy based on deeplabV3+ semantic segmentation algorithm. Through theoretical analysis and experimental verification, it is found that when the SPP structure is changed to 1-2-4-6, the model can achieve 97.83% and 97.04% optimal performance on image_wise and object_wise, respectively, and compared with the traditional rectangular grasp strategy, it not only improves the grasp stability, but also achieves the leading performance in prediction accuracy. This paper broadens the direction for the research of robot grasping and makes a useful exploration.

Object grasping is a complex system engineering, which includes not only the grasping strategy, but also the prediction of grasping posture. Therefore, this paper will proceed from the perspective of depth image and 3D point cloud to carry out three-dimensional reconstruction of the object, and then realize the study of multi-grasp posture. The grasping strategy and grasping posture are integrated to further improve the success rate and stability of grasping.

Author Contributions: Conceptualization, S.L. and Q.B.; methodology, Q.B., S.L. and J.Y.; software, Q.B. and J.Y.; validation, Q.B. and X.Z.; formal analysis, Q.B.; investigation, Q.B.; resources, Q.B.; data curation, Q.B.; writing—original draft preparation, Q.B.; writing—review and editing, S.J. and

M.S.; visualization, Q.B.; supervision, S.L.; project administration, S.L.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially funded by the National Important Project under grant No. 2020YFB1713300, No. 2018AAA0101803 and by Guizhou University Talents Project Nos. GRjH[2020]14 and by Guizhou University Cultivation Project Nos.GDP[2019]22, and by Key Laboratory of Ministry of Education Project Nos.QKHKY[2020]245.

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

References

1. Xie, R.; Liu, J.; Cao, R.; Qiu, C.S.; Duan, J.; Garibaldi, J.; Qiu, G. End-to-End Fovea Localisation in Colour Fundus Images With a Hierarchical Deep Regression Network. *IEEE Trans. Med. Imaging* **2021**, *40*, 116–128. [CrossRef] [PubMed]
2. Guo, D.; Sun, F.; Liu, H.; Kong, T.; Fang, B.; Xi, N. A hybrid deep architecture for robotic grasp detection. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 1609–1614.
3. Lenz, I.; Lee, H.; Saxena, A. Deep learning for detecting robotic grasps. *Int. J. Robot. Res.* **2015**, *34*, 705–724. [CrossRef]
4. Xu, Y.; Wang, L.; Yang, A.; Chen, L.; Ynag, A. GraspCNN: Real-Time Grasp Detection Using a New Oriented Diameter Circle Representation. *IEEE Access* **2019**, *7*, 159322–159331. [CrossRef]
5. Morrison, D.; Corke, P.; Leitner, J. Learning robust, real-time, reactive robotic grasping. *Int. J. Robot. Res.* **2019**, *39*, 183–201. [CrossRef]
6. Redmon, J.; Angelova, A. Real-time grasp detection using convolutional neural networks. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA), Seattle, WA, USA, 26–30 May 2015; pp. 1316–1322.
7. Kumra, S.; Kanan, C. Robotic grasp detection using deep convolutional neural networks. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017; pp. 769–776.
8. Avella, D.S.; Tripicchio, P.; Avizzano, C.A. A study on picking objects in cluttered environments: Exploiting depth features for a custom low-cost universal jamming gripper. *Robot. Comput.-Integr. Manuf.* **2020**, *63*, 101888.
9. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018.
10. Liu, W.; Lee, J. A 3-D Atrous Convolution Neural Network for Hyperspectral Image Denoising. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5701–5715. [CrossRef]
11. Wang, Z.; Wang, Z.; Dong, B.; Chen, C.; Yang, Y.; Yu, Z. Accelerating Atrous Convolution with Fetch-and-Jump Architecture for Activation Positioning. In Proceedings of the 2020 IEEE International Conference on Integrated Circuits, Technologies and Applications (ICTA), IEEE, Nanjing, China, 23–25 November 2020; pp. 151–152.
12. Liu, Y.; Zhu, X.; Zhao, X.; Cao, Y. Adversarial Learning for Constrained Image Splicing Detection and Localization Based on Atrous Convolution. *IEEE Trans. Inf. Forensics Secur.* **2019**, *14*, 2551–2566. [CrossRef]
13. Pan, X.; Li, L.; Yang, D.; He, Y.; Liu, Z.; Yang, H. An Accurate Nuclei Segmentation Algorithm in Pathological Image Based on Deep Semantic Network. *IEEE Access* **2019**, *7*, 110674–110686. [CrossRef]
14. Lv, L.; Li, X.; Jin, J.; Li, X. Image Semantic Segmentation Method Based on Atrous Algorithm and Convolution CRF. In Proceedings of the 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 19–25 October 2019; pp. 160–165.
15. Pan, D.; Yu, F.; Li, C.; Yang, L. Multi-oriented Scene Text Detector with Atrous Convolution. In Proceedings of the 2020 IEEE Information Communication Technologies Conference (ICTC), Nanjing, China, 29–31 May 2020; pp. 346–350.
16. Jie, D.; Yu, H.; Li, C.; Wu, H.; Ni, F. Research of Teleoperation Grasping Control Method Based on Three-fingered Dexterous Hand. In Proceedings of the 2020 IEEE 39th Chinese Control Conference (CCC), Shenyang, China, 27–29 July 2020; pp. 3812–3816.
17. Wang, D. SGDN: Segmentation-Based Grasp Detection Network For Unsymmetrical Three-Finger Gripper. *arXiv* **2020**, arXiv:2005.08222.
18. Jiang, Y.; 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, Shanghai, China, 9–13 May 2011; pp. 3304–3311.
19. Chu, F.-J.; Xu, R.; Vela, P.A. Real-World Multiobject, Multigrasp Detection. *IEEE Robot. Autom. Lett.* **2018**, *3*, 3355–3362. [CrossRef]
20. Pinto, L.; Gupta, A. Supersizing self-supervision: Learning to grasp from 50 K tries and 700 robot hours. In Proceedings of the 2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 3406–3413.
21. Na, Y.-H.; Jo, H.; Song, J.-B. Learning to grasp objects based on ensemble learning combining simulation data and real data. In Proceedings of the 2017 IEEE 17th International Conference on Control, Automation and Systems (ICCAS), Jeju, Korea, 18–21 October 2017; pp. 1030–1034.
22. Shang, W.; Song, F.; Zhao, Z.; Gao, H.; Cong, S.; Li, Z. Deep Learning Method for Grasping Novel Objects Using Dexterous Hands. *IEEE Trans. Cybern.* **2020**, *10*, 1–13. [CrossRef] [PubMed]