# SS-ARGNet: A Novel Cascaded Schema for Robots' 7-DoF Grasping in Adjacent and Stacked Object Scenarios

Xungao Zhong, *Member, IEEE*, Jiaoguo Luo, Xunyu Zhong, Huosheng Hu, *Senior Member*, *IEEE*, Qiang Liu

***Abstract*—Multiple objects in tightly adjacent and stacked configurations pose significant challenges for robotic systems in achieving reliable grasping, as existing algorithms often struggle to distinguish stacked objects and are prone to causing disruptions to the original scene due to improper grasping postures and collisions. To address these challenges, we developed a category-agnostic segmentation and cascaded 7-DoF pose prediction approach for adjacent and stacked objects grasping, using a single vision image. Specifically, a Stacked Segmentation Network (SS-Net) was tailored based on transformer and region proposal modules to achieve robust mask prediction, thereby accurately localizing candidates within the scene. Simultaneously, the Attention Residual Grasping Network (ARG-Net) was proposed to estimate the 7-DoF pose of individual targets, employing a new collision-free strategy to avoid interference between the gripper and the candidates. The integrated SS-Net and ARG-Net (SS-ARGNet) schema significantly enhances robotic performance in practical applications, achieving grasp completion rates of 92.8% and 89.1% for adjacent scenarios, and 87.4% and 84.9% for stacked scenarios, for similar and unknown objects, respectively, with a grasp response time of less than 0.9 seconds.**

***Index Terms*— 7-DoF pose prediction, stacked scenarios, category-agnostic segmentation, cascaded schema, robot grasping**

## I. Introduction

Robust grasping serve as the foundation for robots to develop higher-level embodied-AI (Artificial Intelligence) skills. In recent years, the significant progress has been made in robotic grasping manipulation using various deep learning methods, especially, particularly those aimed at improving the accuracy of grasp detection [1]-[2]. However, robots still struggle to generalize performance beyond scene-limited training samples, especially in scenarios where targets are densely adjacent and stacked, making it challenging for robots to demonstrate the same robustness as in single-object scenarios [3]-[4].

Due to the complexity of stacked scenarios and the diversity of target objects, existing model-based and data-driven algorithms often struggle to determine an appropriate grasp pose for specific targets in stacked environments [5]-[6]. Model-based approaches rely heavily on the accuracy of matching physical models, and pose measuring is significantly hindered when dealing with objects unfamiliar to the trained network. In contrast, data-driven methods do not require pre-established physical models, but instead directly predict grasping configurations based on scene features [7]-[9]. For example, these algorithms predict grasp confidence for each pixel through heatmaps to assess grasp quality. However, because grasp points are predicted across the entire scene, it is challenging to ensure these grasp predictions lead to robust performance in real robotic manipulation.

In scenarios with occlusion or stacking, previous grasping networks struggle to effectively distinguish between targets within point cloud representations. Overlapping point clouds from different objects significantly disrupt grasp pose predictions. Additionally, the close proximity of objects within the scene increases the likelihood of collisions between the robotic system and non-target items, compromising the integrity of the operational environment. As a result, achieving high-performance grasp detection for objects in adjacent and stacked configurations remains a significant unresolved challenge [10]-[13].

As illustrated in Fig. 1, this paper presents a new technique for 7-DoF (degrees of freedom) pose prediction, addressing scenarios where objects may overlap and are closely adjacent. To achieves this, we developed a modular prediction schema using the SS-Net and ARG-Net cascaded architecture, referred to as SS-ARGNet. In this system, the segmentation module SS-Net has been specifically trained on a re-annotated dataset simulating stacked scenarios, enabling it to generate precise masks for each candidate. In addition, the designed grasp detection model, ARG-Net, integrates both channel-wise and

Manuscript received xx, xxx; revised xx, xxxx; accepted xx, xxxx. 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, in part by the Innovation Foundation of Xiamen under Grant 3502Z20206071. (Corresponding author: Xungyu Zhong.)

X.G. Zhong and J.G. Luo are with the School of Electrical Engineering and Automation, Xiamen University of Technology, Xiamen 361024, China, and also with the Xiamen Key Laboratory of Frontier Electric Power Equipment and Intelligent Control, Xiamen 361024, China (e-mail: zhongxungao@163.com; luo_jiaguo@163.com).

X.Y. Zhong is with the School of Aerospace Engineering, Xiamen University, Xiamen 361005, China (e-mail: zhongxunyu@xmu.edu.cn).

H.S. Hu is with the School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, U.K. (e-mail: hhu@essex.ac.uk).

Q. Liu is with the School of Engineering Mathematics and Technology, University of Bristol, BS8 1TW Bristol, U.K., and also with the Department of Psychiatry, University of Oxford, OX2 6NN Oxford, U.K. (e-mail: qiang.liu@bristol.ac.uk).

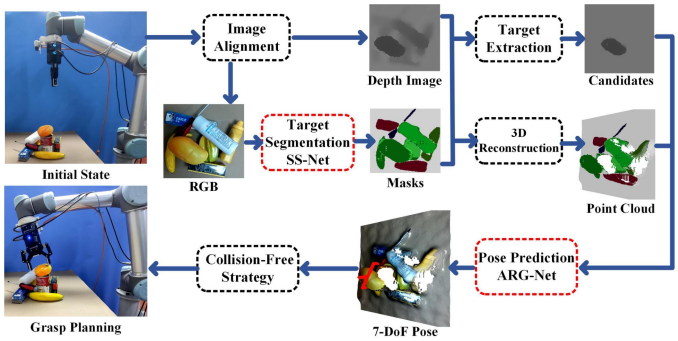



**Fig.1.** Overview of the proposed cascaded modular schema for robot grasp pose prediction in adjacent and stacked object scenarios.

spatial attention mechanisms alongside bottleneck residual modules. These enhancements significantly improve robust feature extraction by allowing the network to focus on correlated features while suppressing less relevant data points.

Furthermore, we introduce a novel method for 7-DoF spatial pose prediction that incorporates a collision-free strategy by assessing the height values of stacked targets and detection optimal grasp widths. This predictive approach aims not only to ensure successful grasping but also to prevent potential collisions between the robotic gripper and candidates in complex arrangements. The key contributions of this work can be summarized as follows:

1) A tailored instance segmentation model, SS-Net: Developed for category-agnostic segmentation in autonomous robotic grasping, SS-Net implements robust mask assignment for each candidate in a stacked state, addressing the challenges of grasp pose prediction for tightly adjacent and stacked objects.

2) A new grasp detection model, ARG-Net: Constructed with multiple residual modules incorporating channel-wise and spatial attention mechanisms, this model enhances grasp feature extraction and compresses network parameters. On the public Jacquard dataset, ARG-Net trained using single-view depth images, achieves a validation grasp accuracy of 96.2%, outperforming other state-of-the-art algorithms to the best of our knowledge.

3) The SS-ARGNet modular prediction schema: Designed by cascading SS-Net with ARG-Net, this system reliably and accurately predicts 7-DoF grasp poses using spatial representation. With the additional collision-free strategy, it effectively improves the grasp completion rate for both similar and unknown objects in adjacent and stacked arrangements. The response time of the grasp system is less than 0.9 seconds in real robotic grasp tasks, meeting real-time requirements.

The remaining sections of this paper are organized as follows. Section II reviewed the relevant literature on robotic grasping. In Section III, we propose a stacked object segmentation and grasp detection method using a cascaded schema, including 7-Dof spatial prediction. Section IV presents the evaluation of the segmentation network, grasping network, and cascaded schema. Real robot experiments are discussed in Section V, with a statistical analysis of grasp completion rates in different scenarios. Finally, Section VI provides a brief conclusion and outlines future work.

## II. Related Work

The relevant literature on deep learning and robotic grasp detection in stacked scenes is reviewed in the following aspects:

Grasping rectangles effectively represent the grasping properties of parallel grippers. Mahler et al. [14] proposed the representative Dex-Net 2.0, agrasp quality estimation network where the Grasping Quality CNN was trained with over 50,000 labeled grasp instances to predict grasp quality from depth images. Morrison et al.[10] introduced the Generative Grasping CNN (GG-CNN), which predicts the grasp quality and angle of each pixel in an image in real-time at a frequency of 50Hz, eliminating the time-consuming discrete sampling of candidate grasps used in previous methods. Kumra et al.'s GR-ConvNet [13] included a residual module utilizing RGB-D multi-channel input to enhance feature acquisition, reducing feature loss from continuous convolution in the feature map. Wang et al.[15] and Shi et al.[16] incorporated encoder and decoder modules in a full CNN to predict pixel-level global maximum confidence grasp points. However, these global maximum confidence methods make it difficult to distinguish boundaries between multiple targets and often lead to grasping collisions or unintentionally picking up multiple objects.

Dong et al.[17] and other researchers [18] trained multi-task networks to enable robots to grasp specific targets in stacked scene. This approach heavily relies on the accuracy of object recognition, which significantly decreases as occlusion increases. For category-agnostic object grasping, Danielczuk et al. [19] proposed SD Mask R-CNN, which uses synthetic deep image training networks to predict masks with different instances. However, the segmentation model was not validated in real robot grasping experiments.

To address severe occlusion and stacking challenges, Wen et al.[20] proposed a grasping framework for industrial objects, training on synthetic class level labels in a simulated environment. To bridge the Sim-to-Real gap, Zeng et al. [21] developed a point-by-point 6-DoF pose estimation network PPR-Net++, which mapped points to centroid space, then performed clustering and voting. However, these methods are limited to specific objects and do not generalize well to category-agnostic objects.

Sundermeyer et al. proposed a class-agnostic model [22] trained with 17 million simulations, which was later extended to real-world application. This model segmented areas of interest around target objects using RGB-D images, reducing prediction dimensions by sampling contact points. Murali et al. [23] designed a cascading framework that generated multiple grasp poses from segmented point clouds, though their method primarily addressed partial occlusion and requires further validation in stacked scenes.

With the increasing popularity of robot grasping application scenarios [24]-[25], the new grasping methods gradually developed from 2D grasping rectangular box to 6-DoF. Liang et al.[26] proposed a lightweight 3D point cloud grasping evaluation network Pointnet-GPD. This method enabled the network to capture complex geometric structures even in sparse point clouds, but the unfiltered clutter made it difficult for the network to eliminate the interference of multiple

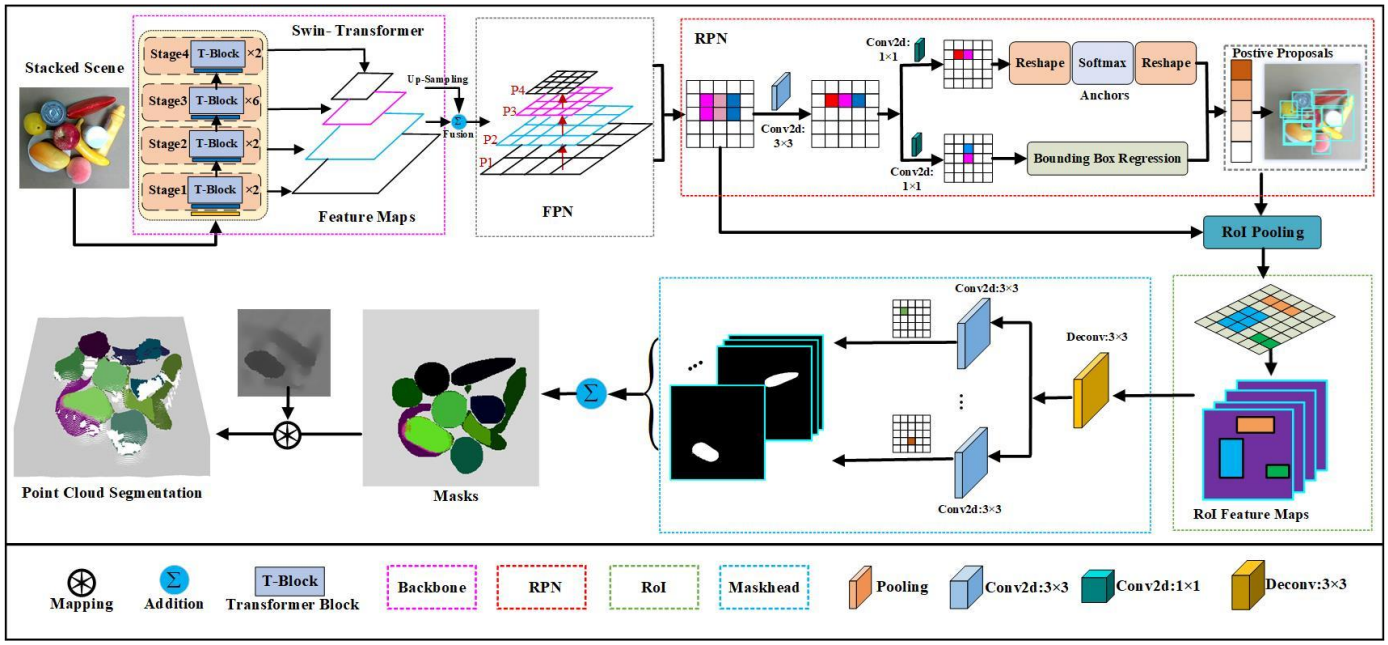



**Fig.2.** The Stacked Segmentation Network (SS-Net) based on Swin-Transformer and Regional Proposal Network (RPN), the SS-Net model generates instance masks and mapping them into the point cloud space.

targets grasping. Li et al. [27] proposed an SSC 6-DoF pose estimation model Pointnet++, was used to encode features and formalize the pose estimation in clutter into a multi-task learning problem that can perform 6-DoF grasping tasks. But the success and completion rates of grasping tasks are significantly deteriorate in stacked environments.

In this paper, we introduce a segmentation network specifically tailored for stacked objects, trained on a newly re-labeled Graspnet-1Billion grasping dataset to predict precise masks for category-agnostic object in adjacent and stacked scenes. By measuring the maximum grasp confidence within the region of a particular object's region, this approach prevents grasps from being predicted at the object's edge or between multiple objects. Additionally, to improve the accuracy of 7-Dof grasp pose prediction and reduce inference latency introduced by the cascaded segmentation network, we designed a lightweight residual grasp network with integrated channel and spatial attention structures in bottleneck layers. This design reduces inference time and enhances prediction accuracy.

## III. Proposed Methods

The proposed cascaded schema, SS-ARGNet, takes RGB-D input and iteratively generates 7-DoF grasp pose for a candidate target. This grasp system enables the independent training and evaluation of SS-Net and ARG-Net, while the modular schema allows each module to be modified without affecting the overall system.

### *A. Stacked Segmentation model SS-Net*

Traditional segmentation networks provide instance masks and category semantic labels, but they often face challenges when extended to new categories. This limitation can severely degrade the performance of the grasp prediction module, particularly in multi-object grasping tasks. Additionally, in these methods, the parallel outputs of object segmentation and grasp configuration are independent of each other, meaning the shape information from the mask is not effectively utilized to improve grasp prediction performance.

In this work, we introduce a category-agnostic object recommendation algorithm and cascade the object segmentation with grasp prediction. The cascaded system autonomously segments masks of stacked objects from a single-view image, and generates unique identifiers for each candidate, without relying on category information. The segmentation algorithm for unknown categories is designed with strong feature learning and discrimination capabilities, enabling the robot to better understand the shape and position of objects to be grasped through pixel-level masked segmentation information.

Thus, the proposed category-agnostic Stacked Segmentation Network (SS-Net), as shown in Fig.2, is constructed using regional proposal techniques, with the Swin-transformer [28] structure as the backbone for target feature extraction. The feature pyramid network (FPN) is then built by up-sampling and fusing the extracted feature maps. In FPN, smaller feature maps (such as P4) are used to segment larger objects, while larger feature maps (such as P1) retain more details of the original image, providing additional texture information for small object segmentation. Next, the Region Proposal Network (RPN), a category-agnostic target detector based on sliding windows, generates anchor boxes for possible targets in the corresponding region of each pixel. These anchor boxes intercept the relevant area in each feature map through the Region of Interest (RoI), which are then used in subsequent convolution operations to generate the mask for each target.

### *B. Target Separation Algorithm*

The raw image of 224 × 224 size is commonly used as input for grasping inference in existing methods. However, this approach introduces challenges, such as difficulty in detecting small objects, which is particularly evident in multi-target or

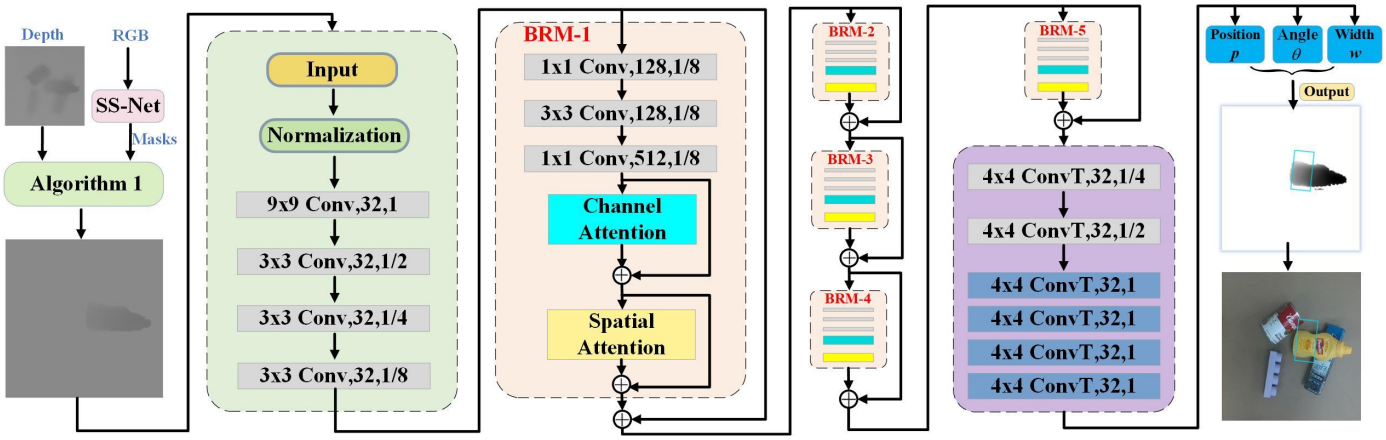



**Fig.3.** The structure of Attention Residual Grasping Network (ARG-Net) for grasp configuration detection.

**Algorithm 1:** the separation of candidates from the scene

**Input:** $M=\{m_1,m_2,...,m_i\}$ and $D(x,y)$, where $M$ represents a set of binarization masks, and $D$ stands for the scene depth image.

**Output:** $SD=\{D_1,D_2,...,D_i\}$, the set of single target depth maps.

$SD \leftarrow \varnothing$
while $M \neq \varnothing$ do
  $m_i \leftarrow m_i / 255$
  $S_i(x,y) \leftarrow m_i \cdot D(x,y)$
  $S_i(j) \leftarrow reshape(S_i(x,y), shape(x \times y))$
  for $j \in S_i$ do
    if $j > mean(S_i(j))$ or $j = 0$ then
      $j \leftarrow mean(S_i(j))$
    end if
  end for
  $D_i(x,y) \leftarrow reshape(S_i, shape(x,y))$
  $SD \leftarrow SD \cup \{D_i(x,y)\}$
end while
**return** $SD$

stacked scene. Thus, this work uses super-resolution down-sampling to crop color and depth maps at a resolution of 1280 ×720, then down-samples them to a size of 240 × 240. This greatly improves the field of view and preserves more details from the original image.

Additionally, to provide robust inputs for the subsequent grasp prediction model, median filtering and multi-frame mean smoothing algorithms are applied to the original depth map. Given an input frame $F_i(x,y)$, where $i$=1,2,...,$n$ with a size of $n$×$n$ and a sampling window $S_{xy}$, the formula used to obtain the robust depth image $D(x,y)$ is:

$$D(x,y)=\frac{1}{n}\sum_{i=1}^{n} median\left(F_i(a,b)\,|\,(a,b)\in S_{xy}\right) \quad (1)$$

As shown in Algorithm 1, a predicted mask combined with a robust depth map is used to partition the depth region $D_i(x,y)$ of the *i-th* object in a stacked scene, thereby converting the scene's multiple-object depth data into each individual single-object depth maps. Given the input variable $M=\{m_1,m_2,...,m_i\}$ is a set of binary masks obtained from SS-Net, and $D(x,y)$ is a robust depth image of the scene, the specific process is as follows:

1) Firstly, apply the mask to the depth image to obtain the depth region $S_i(x,y)$ of a single target.

2) Next, convert $S_i(x,y)$ from a two-dimensional image to a one-dimensional array $S_i(j)$. For each loop iteration, traverse the depth value corresponding to each pixel. If a pixel's depth value is greater than the mean or equal to zero, replace that pixel with the mean value.

3) Finally, convert $S_i(j)$ back into a two-dimensional image. Repeat the above process in a while loop to obtain the set of single-target depth maps $SD=\{D_1(x,y),D_2(x,y),...,D_i(x,y)\}$, which represents the extracted depth maps for each single target from the stacked depth maps.

### *C. Attention Residual Grasping Network ARG-Net*

We utilize the segmentation network SS-Net to highlight candidate target in unconstrained stacked scenes and then employ the grasp dataset to train a grasp detection network for estimating the grasp configuration $G_i$ from the single-channel depth image $D_i(x, y)$, defined by:

$$G_i=(p_i,\theta_i,w_i,q_i) \quad (2)$$

where $p_i$=($x_i$, $y_i$) is the grasp position coordinates in pixel space, $\theta_i$ is the grasp rotation angle in the image plane, $w_i$ is the predicted grasp width, determined by the degree of label opening and closing in the datasets, and $q_i \in [0,1]$ is the grasp quality score for the $i$-th predicted grasp configuration.

Fig.3 shows the specific structure of the Attention Residual Grasping Network (ARG-Net). The depth of a single target is extracted from the stacked depth image using the target separation method (Algorithm 1), and the input target depth is normalized to the range [-1, 1]. We employ 9×9 odd convolution kernels to enhance the network's global feature extraction capability, while smaller 3×3 convolution kernels are used for additional feature extraction to reduce overall computational cost. Through down-sampling convolution, the feature map size is reduced to 1/8 of the original image dimensions.

>Subsequently, the feature maps are fed into five cascaded feature extraction modules. A residual module with a bottleneck layer, termed the Bottleneck Residual Module (BRM), is used to encode, convolve, and decode the feature map. In this paper, the Convolutional Block Attention Module (CBAM) is constructed by channel attention and spatial attention [29]. In the channel attention module, the maximum pooling and average pooling features of the original feature $F$ are processed by a multi-layer perceptron (MLP) to produce the intermediate feature $Fc$:

$$Fc = sigmoid\left(MLP\left(AvgPool(F)\right) + MLP\left(MaxPool(F)\right)\right) \otimes F \tag{3}$$

Where $\otimes$ denotes element-wise multiplication.

In the spatial attention block, the maximum pooling and average pooling features of $F_c$ are convolved with 7 × 7 kernels to produce the feature $F_s$:

$$Fs = sigmoid\left(Conv\left(AvgPool(Fc); MaxPool(Fc)\right)\right) \otimes Fc \tag{4}$$

Finally, the grasp configuration $G_i$ is generated in parallel through the deconvolution channel.

### *D. The loss function*

The loss function of the SS-Net model consists of the region proposal loss $L_{rpn}$ and the target mask prediction loss $L_{mask}$, defined as:

$$Loss = \alpha L_{rpn} + (1-\alpha) L_{mask} \tag{5}$$

Where $\alpha$ is the weight parameter. The candidate anchor boxes are predicted using a cross-entropy loss function, and bounding box offsets are treated as a regression task, where smooth $L1$ loss is applied to compute the error in bounding box offsets:

$$\begin{cases} L_{rpn} = \dfrac{1}{n}\sum_{i=1}^{n} smooth\ L_1\left(t_i, t_i^*\right) - \beta \sum_{i=1}^{n} p_i \log(p_i^*) \\ smooth\ L_1\left(t_i, t_i^*\right) = \begin{cases} 0.5\left(t_i - t_i^*\right)^2, & \text{if } |t - t^*| < 1 \\ |t - t^*| - 0.5, & \text{otherwise} \end{cases} \end{cases} \tag{6}$$

here, $p_i$ represents the ground-truth anchor box, $p_i$* represents the predicted anchor box, $\beta$ is a weight parameter, $t_i$ represents the ground-truth offsets for each anchor box, and $t_i$* represents the predicted offsets by the network for each anchor box.

The target mask prediction contains only two values, 0 or 1. Therefore, the loss function for mask prediction uses binary cross-entropy:

$$L_{mask} = \frac{1}{n}\sum_{i=1}^{n}\left(-y_i \log(y_i^*) - (1-y_i)\log(1-y_i^*)\right) \tag{7}$$

where, $y_i$ represents the true pixel values from the ground-truth labels, and $y_i^*$ represents the predicted pixel values from the network.

To ensure a fair comparison with the baseline, the grasp detection network ARG-Net was trained on the Jacquard dataset, and it loss function is defined as:

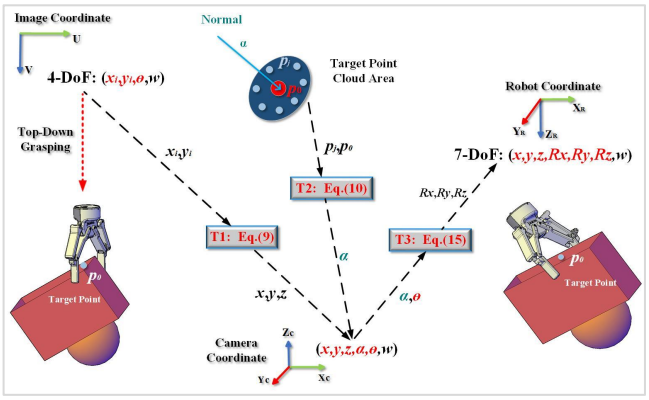



**Fig.4.** The 7-DoF grasp representation for stacked objects.

$$\begin{cases} Loss = \dfrac{1}{n}\sum_{i=1}^{n} smooth\ L_1(G_i - G_r) \\ smooth\ L_1\left(G_i - G_r\right) = \begin{cases} 0.5\left(G_i - G_r\right)^2, & \text{if } |G_i - G_r| < 1 \\ |G_i - G_r| - 0.5, & \text{otherwise} \end{cases} \end{cases} \tag{8}$$

where $G_i$ is the predicted grasping pose, and $G_r$ is the ground truth.

## IV. 7-DOF GRASP POSE PREDICTION

The conventional top-down grasping strategy, such as the 4-DoF approach, often demonstrates a lower success rate in real-world stacked scenarios due to the limited grasp posture flexibility of the robot[18],[32]-[33]. As shown in Fig. 4, when an object is tilted in a chaotic scene, the top-down grasp method may lead to an unstable grasp due to improper alignment with the object's center of gravity. Additionally, existing methods that directly generate a fixed grasp width cannot adequately scale for different sizes of candidate objects..

To overcome these issues, we propose a 7-DoF grasp method based on ARG-Net and pose spatial representation, allowing the robot gripper to approach inclined targets perpendicularly to their surfaces and predict an adaptive grasp width. This adjustment enhances the stability and suitability of the grasp. As shown in Fig.4, there are three steps to transform the grasp configuration $G_i$ in Eq.(5) into a 7-DoF spatial representation.

Firstly, the grasp position in image coordinate $p_i$=($x_i$, $y_i$) as predicted by ARG-Net, is converted into camera coordinate $p_0 = (x, y, z)^{\mathrm{T}}$ using the camera's intrinsic parameter matrix $T_1$.

$$p_{0=}\begin{pmatrix} x \\ y \\ z \end{pmatrix} = T_1\left(x_i, y_i\right) = \begin{pmatrix} \dfrac{x_i - c_x}{f_x} \\ \dfrac{y_i - c_y}{f_y} \\ 1 \end{pmatrix} dz \tag{9}$$

where $c_x$, $c_y$, $f_x$, and $f_y$ represent the intrinsic camera parameters, and $dz$ denotes the depth value at the pixel point.

Next, the normal vector of the target point $p_0$ is calculated from the local point cloud $p_j$, with the transformation denoted as $T_2$:

$$\begin{pmatrix} pitch \\ yaw \\ roll \end{pmatrix} = T_2\left(p_j, p_0\right) = I + N^2\left(1-\cos\varphi\right) + N\sin\varphi \quad (10)$$

where:

$$N = \begin{pmatrix} 0 & -n_z & n_y \\ n_z & 0 & -n_x \\ -n_y & n_x & 0 \end{pmatrix} \quad (11)$$

$$\varphi = angle\left(axis\_z, \alpha\right) \quad (12)$$

$$\alpha = \begin{cases} v_{\min(\lambda_C)}, s\cdot\alpha > 0 \\ -v_{\min(\lambda_C)}, s\cdot\alpha < 0 \end{cases} \quad (13)$$

$$C = \frac{1}{n}\sum_{i}^{n}\left(p_i - p_0\right)\left(p_i - p_0\right)^{\mathrm{T}} \quad (14)$$

where, $v$ represents the eigenvector corresponding to the smallest eigenvalue $\min(\lambda_C)$ of the covariance matrix $C$, $S$ denotes the camera direction. $I = R^{3\times3}$ is the identity matrix, $\varphi$ is the angle between the $z$-axis and the normal vector $\alpha$ , and $N$ is an antisymmetric matrix representing the orthogonal vector to the $z$-axis and the normal vector $n = \left(n_x, n_y, n_z\right)$.

Finally, using the Rodrigues transformation $T_3$, the rotation matrix is converted into three rotational axis variables $Rx$, $Ry$, and $Rz$.

$$\begin{pmatrix} Rx \\ Ry \\ Rz \end{pmatrix} = Rodriges\left(pitch, yaw, \theta\right) \quad (15)$$

Thus, the 7-DoF grasp pose is measured in robot coordinate as:

$$G_r = \left(x, y, z, R_x, R_y, R_z, w\right) \quad (16)$$

where, $w$ is the grasp width predicted by the ARG-Net grasp module.

## V. Comparison And Evaluation

This section presents the validation results of the proposed SS-ARGNet schema, with separate evaluations of the SS-Net and ARG-Net modules. Additionally, the same public dataset and evaluation strategy were used to ensure a fair comparison with different state-of-the-art algorithms.

### *A. Dataset reconstruction*

Currently, there are few datasets specifically tailored for grasp tasks in stacked scene. Datasets such as Cornell, Jacquard and Graspnet-1Billion [9] collects information from non-stacked scenes in real environment. However, existing point cloud segmentation techniques struggle to utilize these datasets effectively in scenarios where objects are stacked. Therefore, overcoming this challenge requires constructing appropriate labels to train a robust network capable of segmenting stacked targets. As shown in Fig.5, the dataset was reconstructed based on the instance data of Graspnet-1Billion. Firstly, we sampled images from different scenes and numbered them sequentially. Next, mask labels were used to separate and binarize the scene objects. Finally, each object was labeled with a unique number and corresponding mask region, and the images were renumbered in a format suitable for network training. This process establishes segmentation labels for individual targets within multi-target images, allowing for multiple predictions on the same image during SS-Net training.

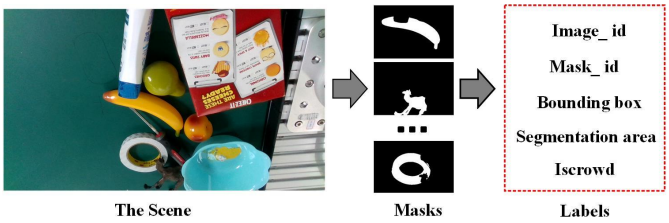



**Fig.5.** The reconstructed dataset based on Graspnet-1Billion for stacked scene segmentation.

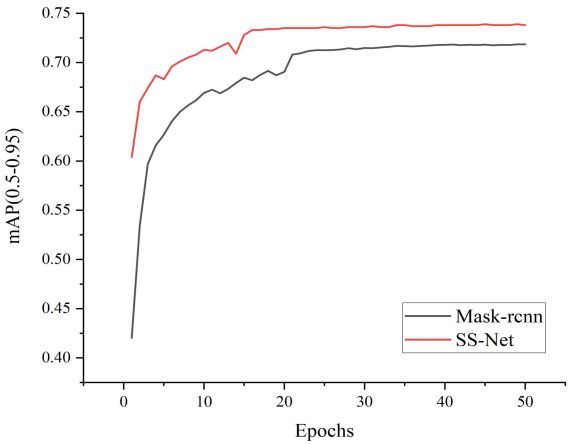



**Fig.6.** Verification accuracy of our SS-Net and Mask-rcnn using the same training strategy.

### *B. Segmentation module SS-Net evaluation*

To evaluating the performance of the instance segmentation module SS-Net, the Intersection over Union (IoU) metric was used to quantify overlap between the predicted region and the ground truth label. We divided our re-annotated Graspnet-1Billion datasets into a 4:1 ratio for model training and validation, respectively. To balance computational efficiency and model convergence speed, we employed the Adam optimizer for model training over 50 epochs, with a batch size of 4. The momentum parameter was set to 0.9, and the initial learning rate was set to 0.001.

The same training strategy was applied to the state-of-the-art method Mask-rcnn [30] and our SS-Net. The training results are shown in Fig.6. For Mask-rcnn, the initial batch accuracy is around 0.42, and the iteration loop stabilizes at approximately 25 batches, with a peak validation accuracy of 0.718 within 50 batches. In contrast, the initial batch accuracy for SS-Net is approximately 0.6, with validation accuracy improving rapidly within the first 20 batches, and a faster convergence rate than Mask-rcnn. The highest

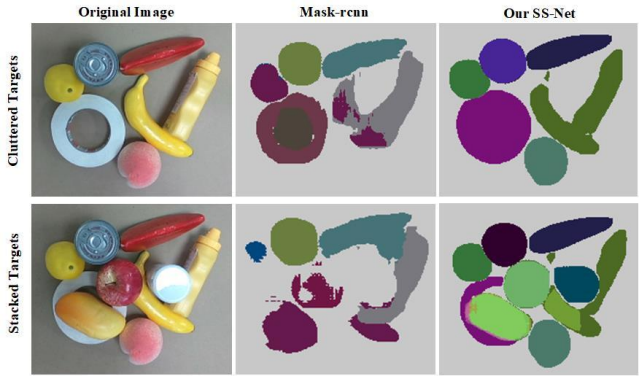



**Fig.7.** The comparison of our SS-Net with state-of-the-art method Mask-rcnn for mask prediction in different scenes.

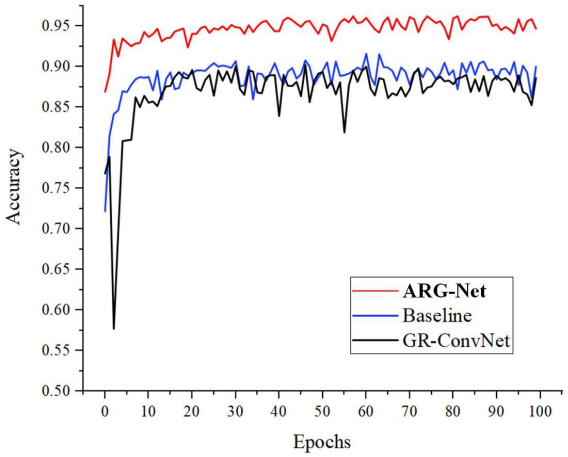



**Fig.8.** The verification accuracy between our ARG-Net, Baseline and GR-ConvNet.

validation accuracy for SS-Net within 50 batches is 0.743. Thus, compared to Mask-rcnn, our method improves steady-state accuracy and convergence rate by 3% and 25%, respectively.

The visualized results of our SS-Net and Mask-rcnn are shown in Fig.7. For cluttered targets, the masks predicted by our SS-Net are more refined, capturing almost the full edge of each target. In contrast, for stacked object segmentation tasks, the results predicted by Mask-rcnn exhibit missing targets and significant shape inaccuracies. However, our SS-Net nearly preserves the original shapes of all targets, even when they are stacked. The results clearly demonstrate the effectiveness of the proposed instance segmentation module, SS-Net, for accurate mask prediction.

### *C. Grasp detection module ARG-Net evaluation*

The proposed grasp detection module, ARG-Net, was trained on the Jacquard dataset using an input depth image size of 240 × 240 and a training batch size of 16. A total of 100 batches were trained using the Adam optimizer.

Ablation tests were conducted to validate the effectiveness of the CBAM attention mechanism and the bottleneck residual module. ARG-Net with the bottleneck residual module but without CBAM was considered the Baseline, while the state-of-the-art reference model selected for comparison was GR-ConvNet [13]. The evaluation results are shown in Fig.8.

TABLE I
COMPARISON IN NETWORK PARAMETERS

| Attribute | GR-ConvNet[13] | Baseline | Our ARG-Net |
|---|---|---|---|
| Input size (pix) | | 240×240 | |
| Parameters (MB) | 7.63 | 6.33 | 7.52 |
| Speed (ms) | 19.6 | 17.1 | 18.7 |
| Accuracy (IoU %) | 90.2 | 91.6 | **96.2** |

TABLE II
RESULT OF DIFFERENT METHODS ON JACQUARD DATASET

| Approaches | Input | Accuracy (IoU) | Inference Time |
|---|---|---|---|
| Depierre et al.[4] | RGB-D | 81.9 % | - |
| Morrison et al. [7] | D | 84.0 % | 19ms |
| Zhang et al.[6] | RGB | 90.4 % | 40ms |
| Chalvatzaki et al. [8] | D | 91.1 % | - |
| Zhou et al.[11] | RGB | 91.8 % | 118ms |
| Cheng et al.[2] | RGB | 92.3 % | 9ms |
| Song et al.[12] | RGB | 93.2 % | - |
| Tian et al.[24] | RGB-D | 94.0 % | 20ms |
| Kumra et al.[25] | RGB-D | 94.6 % | 26ms |
| Wang et al.[36] | RGB-D | 94.6 % | 42ms |
| Tong et al.[1] | RGB-D | 94.6 % | 33ms |
| Yu et al.[37] | RGB-D | 95.9 % | 35ms |
| Ours | D | **96.2 %** | 18.7ms |

It is evident that the introduction of the bottleneck residual module significantly improves the training stability, especially within the first 10 training batches. Compared with GR-ConvNet [13], the Baseline model shows improved training accuracy, with the highest gain reaching 1.4%, demonstrating the effectiveness of the bottleneck residual module. When ARG-Net incorporates the CBAM attention mechanism, it achieves a significant improvement in training accuracy, up to 4.6% compared to the Baseline. In addition, the training iteration curve is smoother, further proving the effectiveness of the CBAM attention mechanism.

The specific parameters for model training are shown in Table I. The Baseline model with bottleneck residuals has a size of only 6.33MB, which is 1.3MB less than GR-ConvNet and demonstrates the fastest inference speed. This shows that introducing bottleneck residuals can reduce both the parameters and computational complexity of the model. The addition of the CBAM attention mechanism in ARG-Net increases some network parameters, resulting in a slight impact on inference speed. However, compared with GR-ConvNet, ARG-Net still shows an improvement in inference speed and significantly enhances verification accuracy.

To verify the generalization performance of ARG-Net, comparative experiments were conducted on the large-scale Jacquard dataset. The experimental results are presented in Table II. ARG-Net achieves a grasp accuracy of 96.2%, which is higher than the best-performing method [37], and also exhibits a detection speed that is 16.3ms faster. Moreover, the grasp accuracy of ARG-Net surpasses that of most existing methods. While SKGNet [37] and BA-Grasp [1] achieve

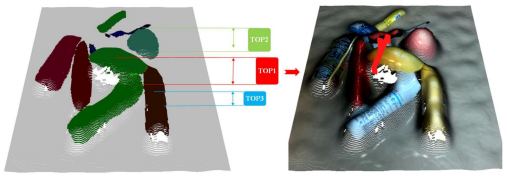



**Fig.9.** Collision detection based on the targets' height values.

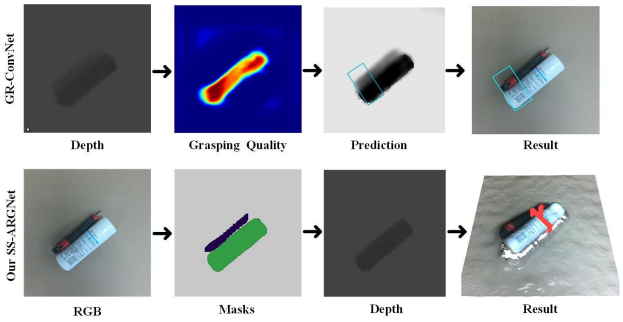



**Fig.10.** Illustration of grasp detection for closely adjacent targets.

optimal performance on the Jacquard datasets, their grasp accuracy sharply declines in scenarios with closely adjacent and stacked objects. In contrast, ARG-Net's grasp accuracy exceeds that of SKGNet and BA-Grasp by over 30%, as these methods are designed for non-stacked objects and lack adaptability to complex stacked scenarios.

### *D. SS-ARGNet with collision-free strategy evaluation*

The SS-ARGNet schema, which cascades SS-Net and ARG-Net was further tested, incorporating two strategies to improve grasp success rate. Firstly, to address frequent collisions between objects and the robot end-effector, as shown in Fig.9, where adjacent objects often differ in height, we sort the mask areas in depth, prioritizing the object with the highest value in each iteration. By detecting and ordering the height values of targets, the robot end-effector minimizes collisions with nearby objects during the grasping process. Secondly, as shown in Fig.10, in scenarios where objects are closely adjacent, existing methods that predict grasp quality (such as GR-ConvNet) are more likely to mistakenly treat two targets as a single entity, leading to the grasping of multiple objects simultaneously. In contrast, SS-Net segments the stacked scene, and objects are ordered by depth maps. ARG-Net then predicts an adaptive grasp width for each individual target using the 7-DoF grasp method. This approach indirectly encourages the robot to grasp only parts belonging to a single object, thereby avoiding collisions and ensuring successful grasping.

To fairly compare the performance of our SS-ARGNet with state-of-the-art methods, such as GR-ConvNet [13], PointnetGPD [22], and GraspNet [9], we visualized the grasp detection results in various complex scenes. As shown in Fig.11, for a *multi-object scene*, both GR-ConvNet and PointnetGPD often result in grasping the edge of objects, whereas our SS-ARGNet, similar to GraspNet, outputs a grasp pose closer to the object's center of gravity. In *lightly stacked scenes*, objects located at the bottom are often assigned maximum confidence by GR-ConvNet, PointnetGPD and GraspNet, resulting in incorrect grasps on the bottom object and potentially disrupting the scene's original structure. Our

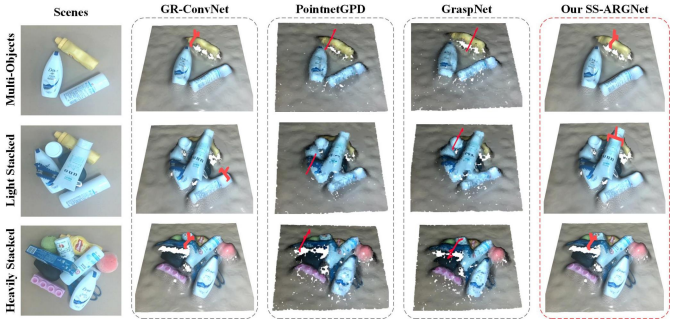



**Fig.11.** Grasp detection results of GR-ConvNet [13], PointnetGPD [22], GraspNet [9] and our SS-ARGNet for adjacent, lightly stacked, and heavily stacked scenarios, respectively.


**Fig.12.** (a) The robot grasp platform, and (b) the different objects used to grasping test.

SS-ARGNet method, however,employs height sorting of local object and predicts the maximum confidence within the segmented area containing the target object, effectively avoiding incorrect grasps. Similarly, in *heavily stacked scenes*, GR-ConvNet assigns a certain maximum grasp confidence across the entire scene as if it were a single object, resulting in grasping multiple objects simultaneously, which can also occur with PointnetGPD and GraspNet. This approach often causes the end-effector to collide with nearby objects. In contrast, our SS-ARGNet method, which cascaded segmentation and grasp prediction, reliably outputs a grasp pose for each individual target.

## VI. Robot Practical Grasping Experiment

We conducted practical grasping experiments in a real robot platform to empirically demonstrate the effectiveness of the proposed cascaded schema, SS-ARGNet, for challenging tasks.

*Platform and Metrics:* As shown in Fig.12(a), the platform consists of a UR5 6-DoF manipulator and a Realsense D435i vision sensor, which has a depth range of 0.2$m$ to 10$m$. The candidate objects typically have altitudes ranging from 0.01$m$ to 0.3$m$. To ensure proper initialization of the manipulator, we maintained a distance of more than 0.5$m$ between the objects and the camera. The camera's RGB and depth images were captured and published at 30Hz. To enhance the depth image, we applied a median filter for smoothing, followed by a registration algorithm to align the depth map with the RGB image. We used a *Robotiq-2F* gripper compatible with the UR5 robot. The default grasping width was set to 1$cm$, with

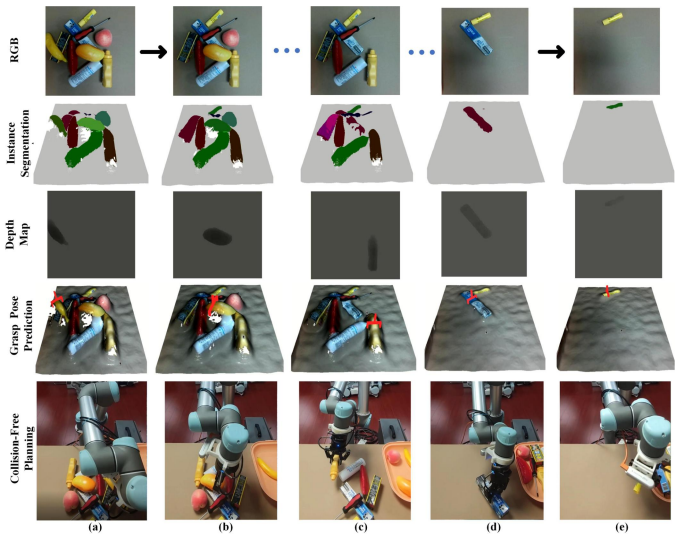



**Fig.13.** The results of the robot cleaning the scene in which the objects are similar to those in the training dataset.

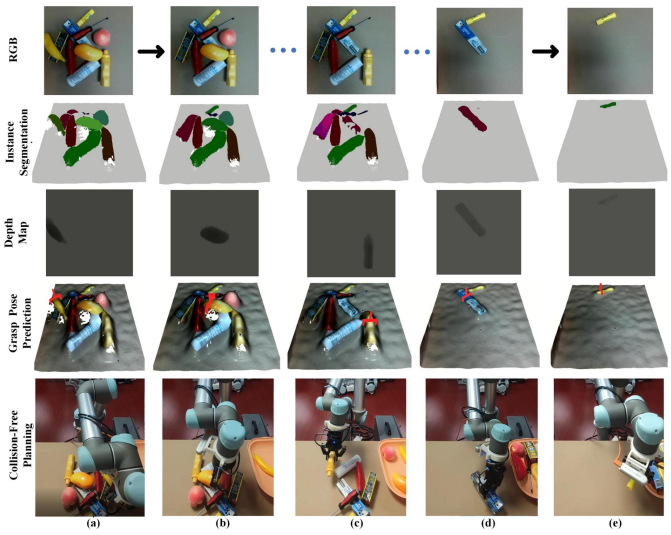



**Fig.14.** The results of the robot's cleaning manipulation in the scene with unknown objects.

predicted opening and closing widths ranging from 0 to 8.5*cm*. The network models were trained using an Intel Core i9 10900k processor and an Nvidia RTX3090 GPU, running on the PyTorch framework with the Ubuntu 18.04 operating system. As shown in Fig.12(b), there are 14 similar objects covering household items, fruits, and office supplies.These test objects simulate the diversity of household scenarios, including variations inshape, color, and texture. Additionally, there are 14 unknown objects that were not included in the network training, spanning household items, outdoor sports equipment, office supplies, and industrial components. Furthermore, 5 adversarial objects from Dex-Net [31] were created using 3D printing, posing challenges to both the segmentation and grasp detection models due to their adversarial geometry and textureless surfaces.

In a statistical setting, the number of objects in the scene and the robot's attempted grasping trials were recorded to calculate the grasp completion rate $C\_R$. The metrics is as follows:

$$C\_R = \frac{N\_O}{SG + FG} \tag{17}$$

Where $SG$ is the number of success grasps, $FG$ is the number of failed grasps, and $N\_O$ is the total number of objects.

*Case 1 test for similar objects:* As shown in Fig.13, the robot continuously performed grasping and placing until all objects were successfully cleared from the workspace. In the first two scenarios, as shown in Fig.13(a) and (b), complex stacking of objects is depicted. Our SS-Net module performed well in segmenting candidate objects and excluding non-target objects (see the second row), providing reliable input to the ARG-Net module for grasp prediction (see the third row). As expected, ARG-Net predicted a reasonable grasp pose for each object (see the fourth row). Combined with collision-free strategies, the position and posture of the 7-DoF gripper allowed it to finely grasp the objects without any collisions (see the fifth row).

In Fig.13(c), where the targets are closely adjacent, the width-based prediction enabled the robot to precisely grasp a specific object with the appropriate configuration, avoiding grasping multiple objects at once. Notebly, in Fig.13(d), the robot arm demonstrated its ability to grasp a fragile target that could easily deform if the gripper's closing width was incorrect. Compared to traditional 6-DoF prediction, the empty box target was finely grasped with minimal deformation, thanks to the additional width prediction in the 7-DoF configuration. In Fig.13(e), the robot showcased its capability to grasp small target, highlighting the versatility of the SS-ARGNet cascaded schema.

*Case 2 test for unknown objects:* As shown in Fig.14, we further tested stacked scenes with unknown objects to verify the generalization capability of the SS-ARGNet schema.

In Fig.14(a), the segmented depth map contains clutter and irregular shapes, representing various parts of objects. This clutter was fed into the ARG-Net grasp detection module, and the output grasp point was selected based on maximum confidence from the quality map, allowing our framework to measure a reliable grasp pose robustly. In the scenario shown in Fig.14(b), the mask output by the SS-Net segmentation module generally represented the shape of the object and excluded surrounding parts. As the number of stacked objects gradually decreases, the object contours were fully segmented during the grasping process, as seen in Fig.14(c), with smooth clutter-free edges for the predicted objects. In Fig.14(d), the robot demonstrated its capability to grasp unknown soft targets, while in Fig.14(e), it showcased its ability to grasp unknown small targets. This proves that the proposed SS-ARGNet schema can adapt to a stacked environment with unknown objects.

The gripper pose was visualized in the point cloud image, and the scene objects were clearly reconstructed (see the fourth row in Fig.14). The robot's grasping pose was consistent with the measured pose, achieved through the pose conversion matrix and Rodrigues formula (see the fifth row in Fig.14). This demonstrates successful prediction and alignment of the grasping pose between the end-effector and the gripper's position in the point cloud image.

TABLE III
GRASPING COMPLETION RATE IN DIFFERENT SCENARIOS OF THE PROPOSED CASCADED SCHEMA SS-ARGNET

| Objects | Scenes | Scene Number | Total *N_O* | Trials *SG+FG* | Failure *FG* | Completion Rate |
|---|---|---|---|---|---|---|
| Similar Objects | Adjacent | 9 | 90 | 97 | 7 | 92.8% |
| | Stacked | 9 | 90 | 103 | 13 | 87.4% |
| Unknown Objects | Adjacent | 9 | 90 | 101 | 11 | 89.1% |
| | Stacked | 9 | 90 | 106 | 16 | 84.9% |
| Adversarial Objects | Adjacent | 9 | 36 | 39 | 3 | 92.3% |

TABLE IV
PERFORMANCE COMPARISON BETWEEN DIFFERENT FRAMEWORKS

| Algorithm | Stacked | Unknown Objects | DoF | Completion Rate |
|---|---|---|---|---|
| OCID-Grasp [32] | - | √ | 4 | 71.6% |
| ROI-GD [33] | √ | - | 4 | 83.8% |
| GR-ConvNet [13] | √ | √ | 4 | 83.3% |
| 6-DoF GraspNet [34]+ Ins. Segmentation [35] | - | √ | 6 | 62.7% |
| Object Instance + CollisionNet [27] | - | √ | 6 | 80.3% |
| SS-ARGNet (Ours) | - | √ | 7 | 89.1% |
| | √ | - | 7 | 87.4% |
| | √ | √ | 7 | 84.9% |

*Performance analysis:* According to our statistics of the grasped objects, we tested scenes with 8, 10, and 12 objects, as well as scenes with 3, 4, and 5 adversarial objects. As shown in Table III, five groups of adjacent and stacked scenes with different object attributes were created, and each group was tested over nine rounds. Ideally, this totaled 90 grasps for similar and unknown objects, and 36 grasps for adversarial objects.

When similar objects were adjacent to each other, the robot achieved a grasp completion rate of 92.8%. In contrast, when similar objects were stacked, the completion rate was 87.4%. These results demonstrate the effectiveness of SS-ARGNet in successfully grasping similar objects in both adjacent and stacked scenarios. For scenes with unknown objects,the robot achieved a grasp completion rate of 89.1% when the objects were adjacent, and 84.9% when they were stacked. For adversarial objects, the grasp completion rate was 92.3%, comparable to existing methods in single-object grasping tasks. These results illustrate that our proposed SS-ARGNet schema is capable of handling novel objects of various shapes and arrangements effectively.

In Table IV, we present a comparison between state-of-the-art methods and our approach. Previous works have often struggled to handle the complexities of unknown and stacked objects, typically focusing on scenarios with multiple targets but not on the truly stacked scenes encountered in real world. In contrast, our approach considers stacked scene with unknown objects, resulting in significant improvements in grasp completion rate. Practical testing indicates the response time of our entire system, including segmentation and grasp prediction, is less than 0.9 seconds, making it suitable for real-time applications in various scenes. To the best of our knowledge, our algorithm outperforms the most advanced methods currently available for grasping unknown and tacked targets.


**Fig.15.** The ablation experiment for collision-free strategy, (a) represents the collision between the grasped target and the other target, (b) represents the collision between the gripper and the around target during the robot movement.

We further conducted an ablation test to verify the actual effectiveness of the collision-free strategy. As shown in Fig.15(a), the proposed SS-ARGNet schema without the collision-free strategy uses a traditional global maximum confidence measuring strategy. This approach prioritizes targets with higher grasp quality rather than objects at higher altitudes, increasing the risk of the gripper colliding with the scene during grasping and retreat movements. In Fig.15(b), the robotic arm plans its posture and adjusts the gripper angle, which can lead to unexpected collisions between the gripper and surrounding targets, ultimately resulting in failed grasp attempts. Through extensive practical grasping tests, we found that these failed grasping often occurred due to collision. Without the collision-free strategy, the grasp completion rate was 68.7%, representing an 18.7% decrease compared to the results achieved with the collision-free strategy.

## VII. CONCLUSIONS

In this work, we proposed a cascaded schema, SS-ARGNet, to address the challenges of grasping tightly adjacent and stacked objects by breaking down the instance segmentation and pose prediction tasks. In the segmentation module, the proposed SS-Net was tailored for category-agnostic objects, aiming to assign masks to each object in the scene to facilitate grasp detection for closely adjacent and stacked objects. In the grasp module, the proposed ARG-Net performed 7-Dof pose prediction, incorporating a bottleneck residual structure and a convolutional block attention module. Additionally, to improve practical grasping performance, a collision-free detection strategy is introduced to avoid collisions that often occur with stacked objects.

Through extensive testing on publicly available datasets, our method achieved a validation accuracy of 96.2%,

outperforming other state-of-the-art algorithms. In experiments, our cascaded schema achieved grasp completion rates of 87.4% and 84.9% in stacked scenes with similar and unknown objects, respectively. The system's response time was less than 0.9 seconds, meeting the real-time performance requirements for most grasping applications.

In future work, we plan to enhance our approach by incorporating force balance to prevent target slippage, thereby improving the stability and reliability of the grasping process. These improvements will contribute to advancing robotic grasping capabilities in complex and dynamic scenarios.

## REFERENCES


[1] L. Tong, K. Song, H. Tian, Y. Man, Y. Yan and Q. Meng, "A novel RGB-D cross-background robot grasp detection dataset and background-adaptive grasping network," IEEE Transactions on Instrumentation and Measurement, Online, DOI:10.1109/TIM.2024.3413164, 2024.

[2] H. Cheng, Y. Wang, and M. Q.-H. Meng, "A robot grasping system with single-stage anchor-free deep grasp detector," *IEEE Transactions on Instrumentation and Measurement,* vol. 71, pp. 1-12, Apr. 2022.

[3] S. Yu, D. Zhai, Y. Xia, "EGNet: Efficient Robotic Grasp Detection Network," IEEE Transactions on Industrial Electronics, vol. 70, no. 4, pp. 4058-4067, Apr. 2023.

[4] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2018, pp. 3511-3516.

[5] R. Newbury, M. Gu, L. Chumbley, et al., "Deep Learning Approaches to Grasp Synthesis: A Review," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3994-4015, Oct. 2023.

[6] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping scenes," in 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Nov. 2019, pp. 4768-4775.

[7] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," The International journal of robotics research, vol. 39, no. 2-3, pp. 183-201, Jun. 2020.

[8] G. Chalvatzaki, N. Gkanatsios, P. Maragos, and J. Peters, "Orientation attentive robotic grasp synthesis with augmented grasp map representation," *arXiv preprint arXiv:2006.05123*, 2020.

[9] H. S. Fang, C. Wang, M. Gou and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in Proc. IEEE/CVF Comput. Vision. Pattern. Recognit. (CVPR), Jun 2020, pp. 11444-11453.

[10] D. Morrison, P. Corke and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," Robotics: Science and Systems (RSS), arxiv.org/abs/1804.05172, May. 2018.

[11] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2018, pp. 7223-723.

[12] Y. Song, L. Gao, X. Li, and W. Shen, "A novel robotic grasp detection method based on region proposal networks," Robotics and Computer Integrated Manufacturing, vol. 65, p. 101963, Oct. 2020.

[13] S. Kumra, S. Joshi and F. Sahin, "Antipodal Robotic Grasping using Generative Residual Convolutional Neural Network," 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Oct. 2020, pp. 9626-9633.

[14] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in Robotics: Science and Systems (RSS), arxiv.org/abs/1703.09312, Mar. 2017.

[15] S. Wang, X. Jiang, J. Zhao, X. Wang, W. Zhou and Y. Liu, "Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images," in Proc. IEEE Int. Conf. Roboti. Biom. (ROBIO), Dec. 2019, pp. 474-480.

[16] C. Shi, C. Miao, X. Zhong, X. Zhong, H. Hu and Q. Liu, "Pixel-Reasoning-Based Robotics Fine Grasping for Novel Objects with Deep EDINet Structure," Sensors, vol. 22, no.11, pp.4283, May.2022.

[17] X. Dong, Y. Jiang, F. Zhao and J. Xia, "A Cascaded Multi-Stage Grasp Detection Method for Kinova Robot in Stacked Environments," Micromachines, vol. 14, no.1, pp. 117, Nov. 2022.

[18] H. Sun, Z. Zhang, H. Wang, Y. Wang, and Q. Cao, "A Novel Robotic Grasp Detection FrameworkUsing Low-Cost RGB-D Camera for Industrial Bin Picking," IEEE Transactions on Instrumentation and Measurement, vol. 73, pp. 2513212-2513212, 2024.

[19] M. Danielczuk, M. Matl, S. Gupta and A. Li, "Segmenting category agnostic 3d objects from real depth images using mask r-cnn trained on synthetic data," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2019, pp. 7283-7290.

[20] B. Wen, W. Lian, K. Bekris and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2022, pp. 6401-6408.

[21] L. Zeng, WJ. Lv, ZK. Dong and YJ. Liu, "PPR-Net++: Accurate 6-D Pose Estimation in Stacked Scenarios," IEEE Transactions on Automation Science and Engineering, vol.19, no. 4, pp. 3139-3151, Oct. 2022.

[22] M. Sundermeyer, A. Mousavian, R. Triebel and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in Proc. IEEE Int.Conf. Robot. Autom. (ICRA), Oct. 2021, pp. 13438-13444.

[23] A. Murali, A. Mousavian, C. Eppner, C. Paxton and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in Proc. IEEE Int.Conf. Robot. Autom. (ICRA), May 2020, pp. 6232-6238.

[24] H. Tian, K. Song, S. Li, S. Ma, and Y. Yan, "Lightweight pixel-wise generative robot grasping detection based on RGB-D dense fusion," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-12, Aug. 2022.

[25] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Feb. 2020, pp. 9626-9633.

[26] H. Liang, X. Ma, S. Li, M. Görner and S. Tang, "Pointnetgpd: Detecting grasp configurations from point sets," in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2019, pp. 3629-3635.

[27] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang and L. Li, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," in Proc. IEEE/RSJ Int. Conf. Intell. Robot. Sys. (IROS), Oct 2021, pp. 3571-3578.

[28] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, "Swin transformer: Hierarchical vision transformer using shifted windows," in Proc. IEEE/CVF Int. Conf. on Comp. Vis. (ICCV), Oct. 2021, pp. 10012-10022.

[29] S. Woo, J. Park, JY. Lee and IS. Kweon, "Cbam: Convolutional block attention module," in Proc. Euro. Conf. on comp. vis. (ECCV), Sep. 2018, pp. 3-19.

[30] YC. Fu, JF. Fan, SY. Xing, Z. Wang, FS. Jing, M. Tan, "Image Segmentation of Cabin Assembly Scene Based on Improved RGB-D Mask R-CNN," EEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1-12, Jan. 2022.

[31] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy and K. Goldberg, "Dex-Net 3.0: Computing Robust Vacuum Suction Grasp Targets in Point Clouds Using a New Analytic Model and Deep Learning," IEEE International Conference on Robotics and Automation (ICRA), May 2018, pp. 5620-5627.

[32] S. Ainetter and F. Fraundorfer, "End-to-end Trainable Deep Neural Network for Robotic Grasp Detection and Semantic Segmentation from RGB," IEEE International Conference on Robotics and Automation (ICRA), May 2021, pp. 13452-13458.

[33] H. Zhang, X. Lan, S Bai, X. Zhou, Z. Tian, N. Zheng, "ROI-based Robotic Grasp Detection for Object Overlapping Scenes," in Proc. IEEE/RSJ Int. Conf. Intell. Robot. Sys. (IROS),Nov. 2019, pp. 4768-4775.

[34] A. Mousavian, C. Eppner and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in Proc. IEEE/CVF Int. Conf. on Comp. Vis. (ICCV), Oct. 2019, pp. 2901-2910.

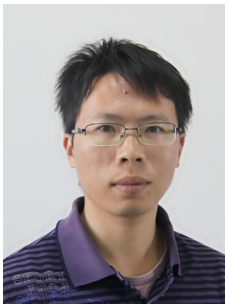
[35] C. Xie, Y. Xiang, A. Mousavian, D. Fox, "The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation," in Proc. Robot Learning (CoRL), May 2020, pp. 1369-1378.

[36] S. Wang, Z. Zhou, and Z. Kan, "When transformer meets robotic grasping: Exploits context for efficient grasp detection," IEEE robotics and automation letters, vol. 7, no. 3, pp. 8170-8177, Jul. 2022.

[37] S. Yu, D.-H. Zhai, and Y. Xia, "SKGNet: Robotic grasp detection with selective kernel convolution," IEEE Transactions on Automation Science and Engineering, vol. 20, no. 4, pp. 2241-2252, Oct. 2022.

**Xungao Zhong** (Member, IEEE) received the B.E. degree in electronic information engineering from Nanchang University, Nanchang, China, in 2007, the M.S. degree in electromechanical engineering from Guangdong University of Technology, Guangzhou, China, in 2011 and the Ph.D. degree in control theory and control engineering from Xiamen University, in 2014. He is currently an associate Professor with the School of Electrical Engineering and Automation, Xiamen University of Technology at Xiamen, China. His current research interests include machine learning, robotic grasping manipulation and visual servoing. He is selected as distinguished young scientific research talent of Fujian province, China.

**Jiaguo Luo** received the B.S. degree in automation from Nanjing University of Information Science and Technology in 2021. He is currently pursuing the M.S. degree in control engineering at Xiamen University of Technology. His research interests include robot motion control, convolutional neural network, and the application of robot grasping.



**Xunyu Zhong** received the M.E. degree in mechatronics engineering from Harbin Engineering University, Harbin, China, in 2007, and the Ph.D. degree in control theory and control engineering from Harbin Engineering University, in 2009. He is currently an associate Professor with the Department of Automation, Xiamen University, Xiamen, China. He is an academic visitor of the School of Computer Science and Electronic Engineering, University of Essex, U.K., for one year from Sept. 2017. His current research interests include robot motion planning, autonomous robots, unmanned system intelligent perception and autonomous navigation system.

**Huosheng Hu** (Senior Member, IEEE) received the M.Sc. degree in industrial automation from Central South University, China in 1982, and the Ph.D. degree in robotics from the University of Oxford, U.K. in 1993. Currently, he is a Professor with the School of Computer Science and Electronic Eng., University of Essex, U.K., leading the Robotics Group. He has authored over 500 research articles published in journals, books and conference proceedings. His research interests include autonomous robots, human–robot interaction, multi-robot collaboration, embedded systems, pervasive computing, sensor integration, intelligent control, and networked robots. Prof. Hu is Fellow of the Institute of Engineering and Technology, Fellow of the Institution of Measurement and Control, and a Chartered Engineer in the U.K. He currently serves as Editor-in-Chief for the International Journal of Automation and Computing, Editor-in-Chief of MDPI Robotics Journal, and an Executive Editor for the International Journal of Mechatronics and Automation.

**Qiang Liu** received his B.Eng. and M.Eng. degrees in automation and control engineering from the Harbin Institute of Technology, Harbin, China, in 2012 and 2014 and the Ph.D. degree in computer science from the University of Essex, Colchester, U.K. He is currently a lecturer in the Department of Engineering Mathematics and Technology at the University of Bristol, Bristol, U.K. Additionally, he holds the position of honorary research scientist in the Department of Psychiatry at the University of Oxford, Oxford, U.K. His research interests include machine learning algorithms and deep neural networks; statistical analysis; medical data analysis; image classification, segmentation and clustering; biomedical signal processing; smart sensors, wearable sensors, sensor integration and data fusion algorithms; visual SLAM and scene understanding; NLP algorithms.