scientific reports

OPEN

Robot multi-target high performance grasping detection based on random sub-path fusion

Bin Zhao[1,2,3,4,5]✉, Lianjun Chang[1,5], Chengdong Wu[3] & Zhenyu Liu[1]

To address the challenge of grasping multi-target objects with uncertain shape, attitude, scale, and stacking, this study proposes a high-performance planar pixel-level grasping network called random sub-path grasp fusion network (RSPFG-Net). The paper introduces the agile grasping representation (AGR) strategy for dexterous grasping of target objects and constructs a Multi-objects Grasping Dataset (NEU-MGD). Secondly, the article introduces the Multi-Scale random sub-path fusion (MSRSPF) module. This module effectively prevents overfitting and improves the robustness of the grasping network in unstructured scenes. The MSRSPF module is connected with the DeepLab v3 network to form the RSPFG-Net for pixel-level grasping and multi-target high-performance grasp detection. Finally, the experiments conducted with RSPFG-Net on publicly available Cornell, Jacquard, and NEU-MGD datasets resulted in an average grasping detection accuracy of 97.85%. In real-world scenarios, the robot achieved an average grasping success rate of 94.31%. These results demonstrate the excellent performance and robustness of RSPFG-Net when it comes to multi-target grasping problems.

Keywords Grasp detection, RSPFG-Net, Adaptive grasping model, Random subpath

Robot vision grasping technology based on depth vision is a current key direction of research in the robotics industry. It is widely used in application scenarios such as assembly, grasping, and palletizing[1,2]. However, due to the uncertainty of multi-target object categories, sizes, shapes, poses, and stacks, it poses challenges to the stability and accuracy of robot grasping. Obtaining the appropriate grasping attitude and position for reliable sensing information in multi-target grasping is a challenging task. Deep visual grasping is a widely researched topic in robotics, and the methods used can be summarized as analytical and empirical. The objective of multi-target grasping detection is to identify a stable position and attitude for the manipulator to grasp the target in different scenes based on visual information. There are two main types of grasping methods: analytical and empirical. Analytical methods use mathematical and physical models of geometry, kinematics, and dynamics to calculate stable grasping parameters[3,4]. Geometric model for grasping objects often struggle to perform well in real-world scenarios due to the challenge of accurately modeling the physical interactions between the robot arm and the target object. Empirical methods, on the other hand, do not rely on 3D models of objects. Instead, they train a grasping model using a known object and apply this model to detect the grasping pose of an unknown object. In recent years, deep learning methods have been developed to detect a planar grasping representation and map the object to a grasping pose in the world coordinate system. These methods typically outperform traditional empirical methods. In recent years, for the problem of grasp detection, research on deep learning methods using two-dimensional images as input has achieved fruitful theoretical and practical results. Research on methods related to deep visual grasping has made significant progress, which has been demonstrated in many kinds of literature. In their publication, Fang et al.[5] introduces the GraspNet - 1 billion benchmark test, which includes a variety of real-world cluttered scenes and detailed annotations. The dataset comprises 97,280 RGB-D images, each with over 1 billion grasping poses. A total of 190 clutter scenes were collected, with 100 serving as training sets and 90 as test sets. Furthermore, Wang et al.[6] introduced a new model called the Oriented Arrow Representation Model (OAR-model) to represent gripper configurations for both parallel jaws and three-finger grippers. This model enhances the applicability of the gripper to different types of grippers to some extent. However, Kumra et al.[7] proposes a new Generative Residual Convolutional Neural Network (GR-ConvNet)

[1]School of Information Science and Engineering, Shenyang University of Technology, Shenyang 110870, China. [2]College of Interdisciplinary Sciences, Liaoning University of Technology, Jinzhou 121001, China. [3]School of Information Science and Engineering, Northeastern University, Shenyang 110819, China. [4]SIASUN Robot & Automation Co., Ltd., Shenyang 110169, China. [5] Bin Zhao and Lianjun Chang contributed equally to this work. ✉email: zhaobin@stumail.neu.edu.cn

model that solves the problem of generating and executing an enantiomorphic robot grasping unknown objects from n-channel scene images. Additionally, Ribeiro et al.[8] proposes a multimodal hierarchical generative grasping CNN (MMH-GGCNN) with a small number of parameters based on the lightweight network GG-CNN. This approach aims to improve grasping performance by utilizing the multimodal and hierarchical nature of grasping components. However, these algorithms are ineffective for predicting grasping of multi-scale stacked targets, and accurately predicting the grasping attributes of targets remains a challenge. The size of the gripper jaw is set empirically in the rectangular representation, which can cause issues for neural network detection performance. Previous research papers mostly focused on predicting only the width of the mechanical jaws' spread, which may result in missing some actual gripping positions. Both the publicly available Cornell and Jacquard fetching datasets focus on single-target fetching tasks. The multi-objective Graspnet-1 billion and Acronym datasets achieve grasping robustness through point cloud processing, and obtain grasping gestures from 6D attitude projection scenes of objects. It takes a lot of computing resources and time to train the model, which cannot be deployed in embedded system environment. There is a lack of relevant research on how to quickly deploy and effectively deal with multi-object stacking scenarios in real scenarios. In view of this, this paper proposes a high-performance grasping detection method called RSPFG-Net, which generates pixel-level grasping information data using randomized sub-paths. The method consists of four parts: AGR-strategy, NEU-MGD dataset, MSRSPF module, and RSPFG-Net network. The summary of our research contributions is as follows.

(1) This paper presents an intelligent robotic system that introduces an AGR strategy based on the directional arrow model and adaptively represents the grasping attributes of the objects to resolve angle conflicts during training.
(2) A multi-target NEU-MGD was established based on this, which includes single-target, discrete multi-target, and stacked multi-target. The NEU-MGD dataset built in this paper has been open source, and the download addresses of related datasets are shared: https://github.com/SimonZhaoBin/NEU-MGD
(3) A method called multi-scale randomized sub-path fusion (MSRSPF) is proposed. It combines sub paths of different lengths to allow the network to select the most suitable set of sub paths, thus enhancing robustness in unstructured scenes.
(4) The RSPFG-Net is formed by connecting the MSRSPF module with the DeepLab v3 network in sequence. This helps to prevent missing grasping poses and reduces the required number of calculations. RSPFG-Net is used for pixel-level grasping of multi-target high-performance grasping detection. Its architecture can predict suitable grasping configurations for multi-target objects in visual scenes. The article uses the relevant open-source code as the reference: https://github.com/dougsm/ggcnn

Our approach

The aim of grasp detection is to predict an appropriate grasping pose in various scenarios by utilizing information about the target obtained from the intelligent camera. The manipulator can then accurately and stably grasp the target with closed fingers[9–15].

A. Multi target grasping system

Related studies have demonstrated that scholars have effectively enhanced robot vision-based task performance by combining RGB vision and thermal imaging in unstructured environments[16,17]. Relying solely on RGB vision, however, proves insufficient for the precise grasping of objects at varying heights or in stacked configurations, as RGB images lack sufficient depth information to distinguish the relative positions of the targets. Moreover, due to its inability to detect material differences between targets, thermal imaging technology significantly impacts grasping accuracy when differentiating objects of the same material. Given this, this paper uses the RGB-D camera to obtain image information. RGB-D cameras provide more accurate grasp height identification through depth information in multi-target stacking tasks, improving task execution precision. This paper analyzes the technical difficulties of multi-object visual grasping, and establishes the depth vision grasping system used in this paper, as shown in Fig. 1[18–21]. Grasping system comprises two main modules: the grasping target information prediction module and the robot grasping control module.

(1) The module for predicting target grasping information acquires RGB and depth images of the scene from the RGB-D camera. Based on the graphical information, it deduces the appropriate bit position of Kinova's grasping object in the camera's field of view, and adopts the AGR strategy of the directional arrow model for target grasping, where the robot's robotic arm approaches the target and closes its jaws.
(2) The grab control module of the robot utilizes the grab information generated by the RSPFG-Net. This information is then communicated to the robot using trajectory planning and the ROS interface of the controller to execute the desired actions. The network model recognizes the grasping information to control the robot and plan and adjust the position and attitude of the grasping target in real-time.

B. AGR strategy

Unlike typical multi-finger grippers, as shown in Fig. 2, each finger of a three-finger gripper can only face towards the center of the gripper. To represent parallel jaw grippers and three-finger jaw grippers with one gripping representation, the three-finger gripper is simplified into a parallel jaw gripper with two jaws of different sizes. The target object is grasped by keeping the two adjacent fingers moving synchronously[22,23].

The strategy for representing dexterous grasping is as follows:

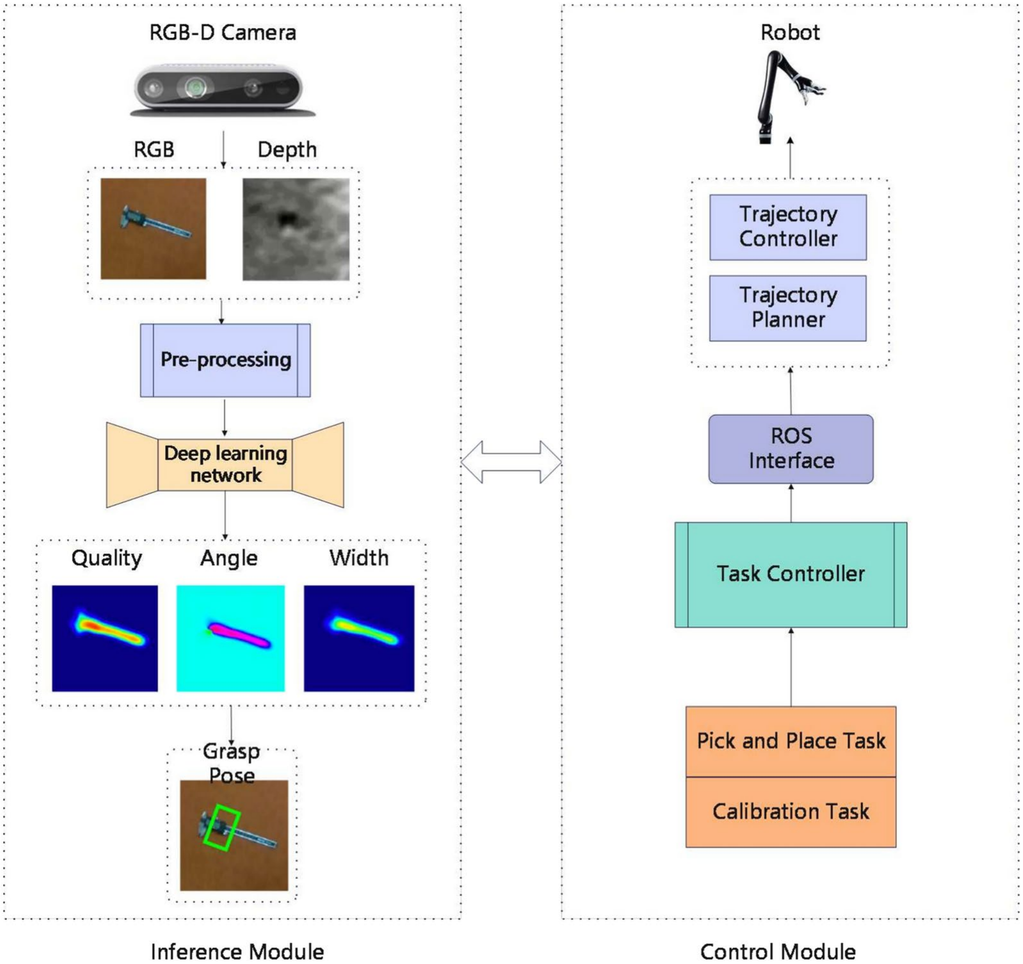

Fig. 1. Multi-target visual grasping system.

(1) The maximum unfolding width W_h and finger width W_f for two-finger and three-finger robotic arms are determined. For rectangular and cylindrical objects with $W_0 < W_h$, a symmetrical gripping method is used. The captured directional arrow model's AGR strategy is represented in the blue area of the figure.

$$S = (x, y, \Theta, W) \tag{1}$$

where (x, y) represents the grasping central point, W is the grasping width, Θ is the grasping angle. Compared with Dex-Net 2.0, the grasping width of the symmetric grasping model is variable, and objects with different widths can be learned.

(2) For spherical objects with diameter $D < \omega_h$ spherical objects using circular grasp mode. The circular grasping model is represented as:

$$C = (x, y, D) \tag{2}$$

where (x, y) also represents the grasping central point, and D represents the sphere's diameter and the grasping width. There is no angle set here. For a circle, the angle can be anywhere between 0 and 360. When predicting, an arbitrary angle value is also generated.

(3) For the hollow circular cylinder, as shown in the following figure, if the outer diameter $D_0 < W_h$, the circular grasping method is adopted; otherwise, when $(D_0 - D_1)/2 < W_h$ and $D_1 > W_f$, the multi-segment symmetric grasping method is adopted. It should be noted that the single finger is always located in

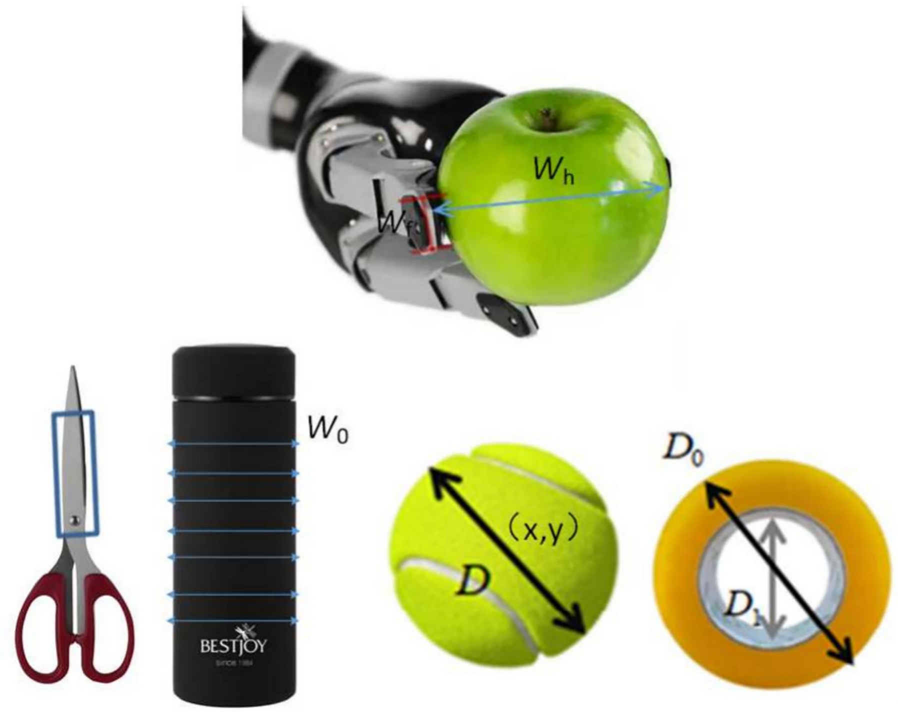

Fig. 2. Manipulators and grasping objects.

the inner circle if it is a three-finger manipulator. The robot arm can grip hollow cylindrical objects with a maximum width greater than that of other objects.

C. RSPFG-Net grasp network

In deep learning, the depth of the network is often related to the performance and learning ability of the model. The academic community now agrees that deeper networks can learn higher-level, more abstract feature representations, improving the model's ability to understand and express complex data, as long as overfitting is avoided. However, as the depth of the network increases, gradient disappearance and explosion may occur. These are common problems in deep learning networks, where the gradient can become very small or very large during backpropagation, making the network difficult to train.

Figure 3 illustrates the overall structure of RSPFG-Net. The directional arrow model is used to label the NEU-MGD multi-object data, which is then converted into pixel-level grasp detection using the AGR-strategy method. A random subpath grasping network, RSPFG-Net, is formed by sequentially connecting subpaths with the DeepLab v3 network. This allows the network to select the appropriate subpath set and quickly generate optimal grasping information data to guide the robot's grasp.

This paper presents a scheme for the RSPFG-Net randomized sub-path grasping network, which consists of three parts: (1) DeepLab v3 network with ASPP[24]; (2) MSRSPF module; (3) Grasping output module. The RSPFG-Net will be described in the following scheme as follows.

(1) Baseline: DeepLab v3 feature extraction details have been preserved, using ResNet-101 and ASPP as the shared encoder backbone. ASPP Module: Atrous Spatial Pyramid Pooling (ASPP) is employed in the Backbone section to utilize the backbone feature extraction network to obtain shallow and deep features. The ASPP consists of the following five components: a 1×1 convolution; a 3×3 convolution with a dilation rate of 6; a 3×3 convolution with a dilation rate of 3; another 3×3 convolution with a dilation rate of 3; and pooling applied to the input feature layer. The multi-scale feature extraction method of DeepLab v3 can effectively grasp the feature information of different scales through a parallel cavity convolution module, improving the depth and breadth of feature extraction.

(2) The MSRSPF module: The MSRSPF allows the network to select a suitable set of sub-paths by combining sub-paths of different lengths. The MSRSPF module contains multilevel residual connections, HASPP and random path selection three functions. 1. HASPP(High Atrous Spatial Pyramid Pooling) architecture consists of five main parts:1×1 convolution with an expansion rate of 6; 3×3 convolution with an expansion rate of 12; 3×3 convolution with an expansion rate of 18; 3×3 convolution with an expansion rate of 24; Pool the input feature layer. 2. The MSRSPF module introduces a path dropout strategy. Within each structural block, the local dropout strategy probabilistically discards each branch but ensures that at least one path from the input to the output is retained. The global dropout strategy guarantees at least one path from the input to the output across the entire network. In contrast, if each structural block were to apply the local dropout strategy, it would result in a global failure of paths across the entire network. This situation is pre-

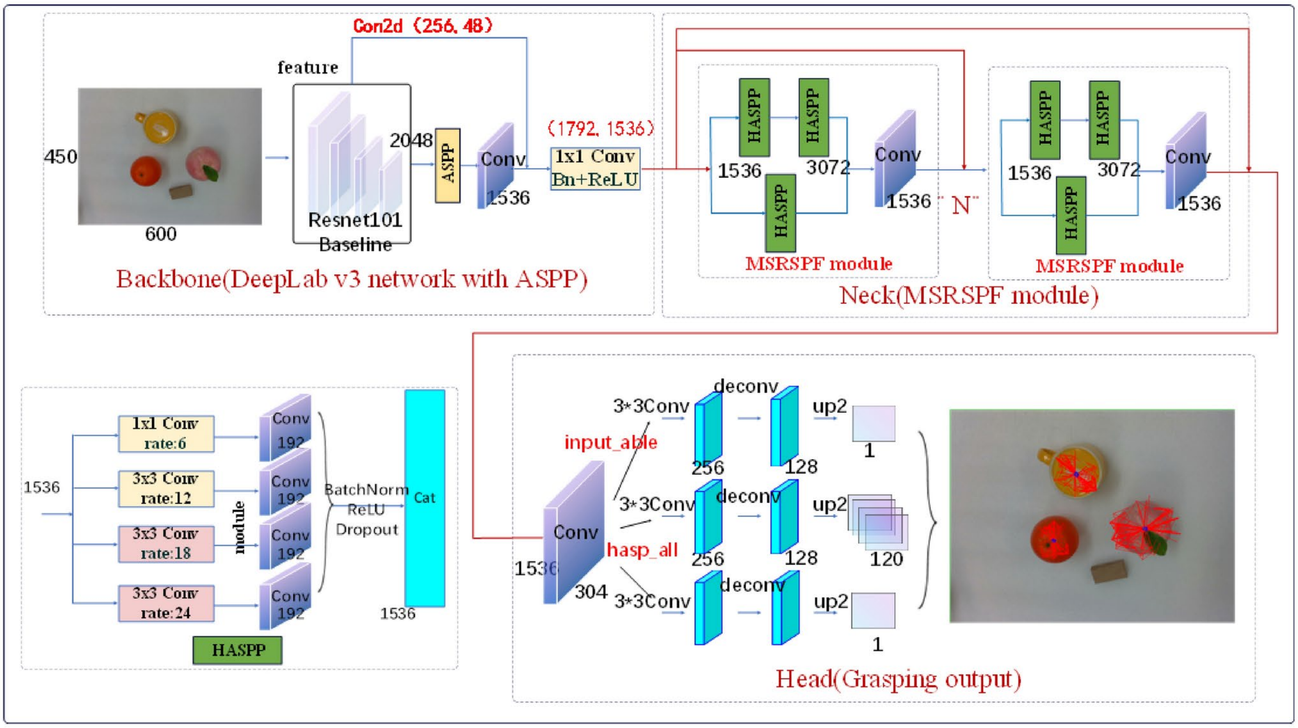

Fig. 3. Structure of the RSPFG-Net.

vented in this network design. 3. Multilevel residual connections alleviate the issues of gradient vanishing and gradient explosion during the training process of deep networks by directly adding the input to the output through skip connections. Residual connections accelerate the convergence of training by reducing the difficulty of parameter updates during network training. The network not only receives information from the current layer but also retains the input information from the previous layer, preventing the loss of information as it propagates through the deeper layers of the network.

(3) Grasping output module: The grasp detection problem is divided into three sub-problems: with the grab point $R = (R_x, R_y)$, grab angle Θ, and grab width W. The output channels of these heads are $(1, k, 1)$ (K=120 in this study). The header for the region outputs the confidence level for each pixel's location within the grasp region R. The Angle head outputs the class k of the grab angle corresponding to each point. From this, we calculate the grab angle by $\theta = 2\pi k/K$. Predicting grasping angles is a task that involves multiple labels within a single category. To avoid competition between categories, we normalize the output of the angle header using a sigmoid function. The width head outputs the grab width corresponding to each point.

Data sets and evaluation

A. Multi-objective NEU-MGD dataset production

This paper introduces the NEU-MGD Northeastern University multi-target grasping dataset, which includes single targets, discrete multi-targets, and stacked multi-targets. The NEU-MGD dataset contains 4000 datasets of various objects, including over 100 types of objects such as fruits, tools, cartons, beverage bottles, and other common objects in daily life. The AGR-strategy directional arrow model is used for labeling. The dataset includes single targets, multiple targets, and multiple stacked targets. The dataset includes pictures with sizes of 640*480 and 1 8 objects per picture. Figure 4 shows some of the labeled samples obtained from the dataset. Grasping models are primarily classified into pinching grasp models and circular grasp models. The circular grasp model allows for grasping at any angle within a 360° range, providing greater flexibility. This model can adaptively represent the grasping attributes of objects, thereby enhancing its applicability to various types of grippers. It meets the criteria for a multi-objective large-scale dataset with sufficient variety and quantity. The dataset also avoids overfitting problems that could cause the network to be unable to learn effective features[8–13].

This paper employs two methods for dividing the dataset into test and training sets: Image-wise split and object-wise split. The former randomly assigns the dataset into training and test sets, while the latter divides the dataset based on object instances. The test set in the object-wise split method consists of objects that have not appeared in the training set. In this paper, Image-wise split method is chosen.

B. Pre-training and evaluation indicators

This paper addresses the multi-target object grasping problem of the Kinova robotic arm. Therefore, we evaluate the RSPFG-Net model using the AGR-strategy model volume and optimize the RSPFG-Net network model

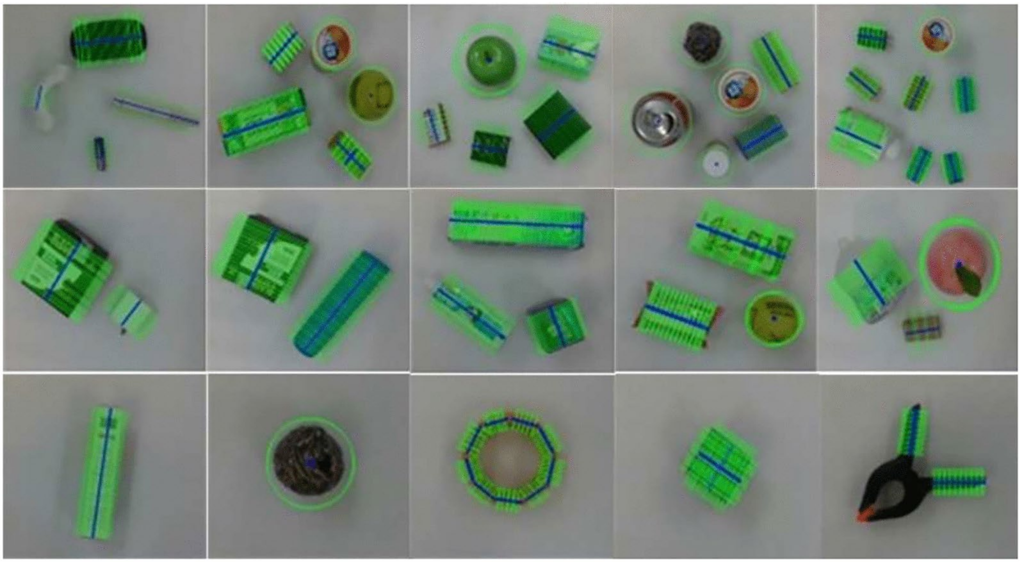

Fig. 4. Example of NEU-MGD grab dataset.

using the loss. The candidate grasping position, grasping width, and grasping angle are considered valid only when the task loss evaluation meets the criteria.

(1) *Grab region* (R_x, R_y) The task at hand involves binary classification for predicting the grasping region. To ensure accurate results, we begin by normalizing the predictions using the sigmoid function. The loss is then calculated using the Binary Cross-Entropy (BCE) function.

$$L_{reg} = -\frac{1}{N}\sum_{n=0}^{N}\left[y_q^n \cdot \log\left(p_q^n\right) + \left(1 - y_q^n\right) \cdot \log\left(1 - p_g^n\right)\right] \tag{3}$$

where N represents the size of the output feature map, p_q^n represents the predicted probability at position n, and y_q^n represents the corresponding label.

(2) *Grab angle* Θ A After normalizing the output of the angle head using a sigmoid function, the loss for the grasping angle is calculated using the Binary Cross-Entropy (BCE) function.

$$L_{ang} = -\frac{1}{N \times L}\sum_{n=0}^{N}\sum_{l=0}^{L}\left[y_l^n \cdot \log\left(p_l^n\right) + \left(1 - y_l^n\right) \cdot \log\left(1 - p_l^n\right)\right] \tag{4}$$

where p_l^n represents the probability that the predicted grasp angle within the range of position $\left[\frac{l}{L} \times 2\pi, \frac{l+1}{L} \times 2\pi\right]$, and y_l^n is the corresponding label.

(3) *Grabbing width W* Predicting grab width is a regression task. The loss of the grab width branch is computed using the BCE function.

$$L_{wid} = -\frac{1}{N}\sum_{n=0}^{N}\left[y_w^n \cdot \log\left(p_w^n\right) + \left(1 - y_w^n\right) \cdot \log\left(1 - p_w^n\right)\right] \tag{5}$$

The width of the predicted crawl for the first n positions is represented by p_w^n, while the corresponding label is y_w^n.

(4) *Multitasking LOSS* To achieve balance between the losses in each branch, the final multitasking loss is defined as:

$$LOSS = \gamma_1 \times L_{reg} + \gamma_2 \times L_{ang} + \gamma_3 \times L_{wid} \tag{6}$$

The loss has weighting coefficients, γ_1,γ_2 and γ_3. In this study, we set the coefficients to γ_1=0.7, γ_2=0.2, and γ_3=0.1.

(5) *Accuracy* The number of objects that can be successfully grabbed out of the total number of all grab poses predicted by the algorithm during the grab. Grasp prediction accuracy measures the accuracy of the algorithm in predicting the grasp pose.

$$LOSS = \gamma_1 \times L_{reg} + \gamma_2 \times L_{ang} + \gamma_3 \times L_{wid} \tag{7}$$

Experiment

The experimental environment is as follows:

Operating system: Ubuntu MATE16.04
CPU: Intel(R) Xeon(R) CPU E5-2620 v4@2.10GHz
GPU: Titan X
Python version: 3.7.13
Torch version: 1.10.1+cu111
Torch-vision version: 0.11.2++cu111

Under the conditions of 500 epoch and 16 batch_size, training the RSPFG-Net network on the GPU requires 7213 MB of memory and takes approximately 6 hours. The algorithm has already been ported to the SIASUN robot controller, which runs on the domestic OpenEuler operating system. Techniques such as model compression, pruning, and quantization are typically employed to accommodate the hardware limitations of embedded systems. These techniques can significantly reduce the model size and computational complexity while maintaining high accuracy during inference.

The paper evaluates the proposed method's performance on the publicly available Cornell and Jacquard datasets, as well as the NEU-MGD dataset of multi-target objects proposed in this study. The main performance metric is the robot grasping success rate, as it is difficult to quantitatively measure the robot grasping detection accuracy. Related experiment videos: https://www.bilibili.com/video/BV1vT421y7Pq/D

A. Ablation experiment

Overfitting occurs when a model performs well on training data but poorly on unseen test data. Deeper networks are more susceptible to overfitting. This section includes an ablation study on the proposed network to assess the impact of the proposed MSRSPF module score on performance. To test the MSRSPF module's ability to accelerate model convergence, prevent gradient vanishing, and avoid overfitting, we plot the ACC and LOSS images over time for both the training and validation sets.

Figure 5 displays the ACC and LOSS images over time for RSPFG-Net without the MSRSPF module at approximately 300 Epochs. The plotted learning curves for the model on the training and validation sets show that the curve of Loss and ACC shows an unstable jump with the increase of Epoch, resulting in severe overfitting of RSPFG-Net. Even when monitoring the model's performance on the validation set and stopping training when the performance no longer improves, the performance of the trained network remains suboptimal.

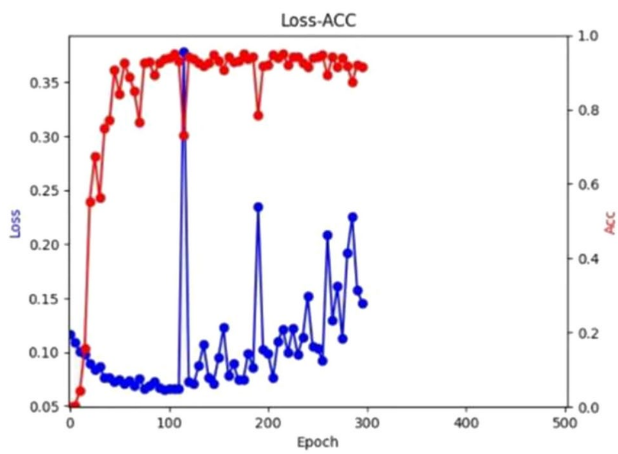

Fig. 5. LOSS and ACC in RSPFG-Net without MSRSPF.

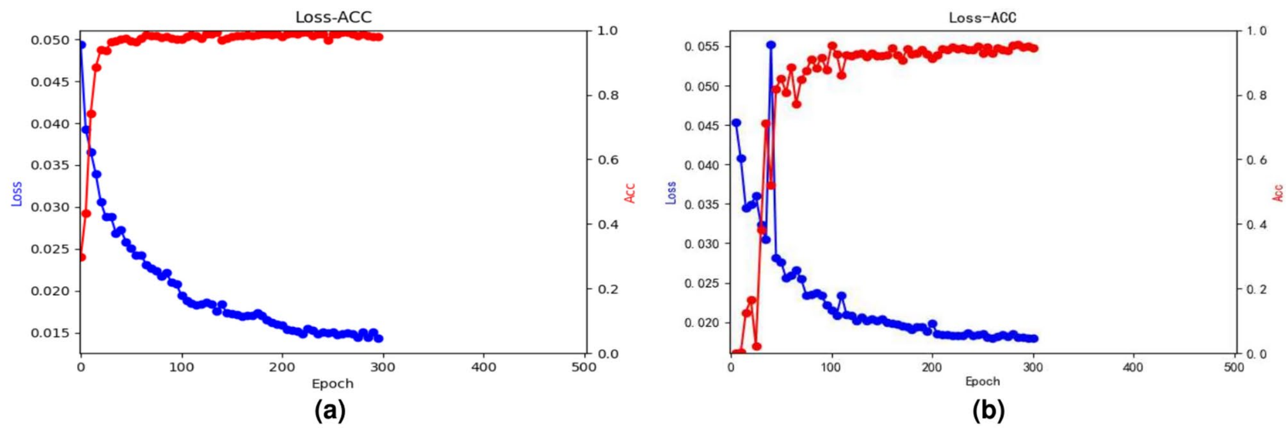

Fig. 6. Comparison experiments prevent overfitting.

Network name	ACC (%)	LOSS	Fps
RSPFG-Net	96.39	0.0175	45
SqueezeNet	74.32	0.025	14
Faster-RCNN	85.6	0.035	11
CascadeR-CNN	91.5	0.024	22
GG-CNN	73	0.0272	19
AlexNet	88	0.026	13
FCGN	97.7	0.0226	9
ResNet-50x2	89.2	0.0275	10

Table 1. Experimental results on the NEU-MGD multi-objects data set.

Gripper objects (M)	success rate
Known	97.52%
Unknown	94.31%

Table 2. Robotic grasp success rate of our method (%).

To obtain a more comprehensive picture of the model's generalization performance, Compare the preventing overfitting experiment of RSPFG-Net with MSRSPF and RSPFG-Net with dropout. Figure 6a shows that the RSPFG-Net with MSRSPF model converges more quickly at around 50 epochs, improving the accuracy and convergence speed of the network. Figure 6b shows that the RSPFG-Net with dropout model converges and stabilizes in terms of LOSS and ACC around the 130 epoch. The MSRSPF module exhibits higher accuracy and convergence efficiency than the dropout module. The designed MSRSPF module effectively solves the overfitting problem caused by a deep network. Therefore, the introduction of the MSRSPF module is practical.

B. Multi-object NEU data set grasp detection

As shown in Table 1, in over 3600 experiments involving robot grasping, the RSPFG-Net has demonstrated good performance in terms of ACC, LOSS, and Fps. The grasping method of RSPFG-Net has an inference speed of 45ms, achieving a balance between grasping accuracy and running speed.

As shown in Table 2, for the re-arrangement and combination of the target objects in the established NEU-MGD data set, 1000 grasping experiments are carried out. The RSPFG Net network architecture achieves 97.85 % accuracy for grasping known objects and 93.3% for grasping unknown objects in both single and multi-object scenarios. These results demonstrate the effectiveness of our method for grasping untrained objects.

The RSPFG-Net network is evaluated for single object grasping detection on the publicly available Cornell, Jacquard dataset under the same experimental conditions. As shown in Fig. 7, the RSPFG-Net model can effectively recognize and detect objects such as scissors, glasses, high heels, and tape while providing multiple grasping points for each object. This capability supports collaborative robots in executing grasping tasks in practical applications, and the grasping results can accurately represent the grasping characteristics of objects, thereby enhancing the model's applicability across different types of grippers.

Multi-target grasp detection of the RSPFG-Net network is carried out on the NEU-MGD data set, and comparative evaluation is carried out under the same experimental conditions. The design and execution of the

Fig. 7. Cornell and Jacquard datasets grasp detection.

Fig. 8. The effect of discrete NEU-MGD data set detection.

discrete multi-object visual grasping experiment and the stacked multi-object visual grasping experiment aim to evaluate the performance of the RSPFG-Net model across different grasping scenarios. These experiments selected various target objects with differing shapes, sizes, and materials, including scissors, building blocks, glasses, tools, calculators, and tape, and so on. As shown in Fig. 8, the discrete multi-object visual grasping experiment primarily focuses on how a collaborative robot can effectively identify and grasp multiple target objects when they are relatively far apart within the visual sensor's field of view. Each object is placed at a specific location, and the goal is to assess whether the RSPFG-Net model can accurately detect multiple targets in such a dispersed configuration and generate appropriate grasping points for each object. The model provides multiple grasping points, enabling the collaborative robot to perform the grasping task based on these points efficiently.

The results prove that the RSPFG-Net network method can effectively predict and grasp different types of objects, and we use less scene information to obtain the maximum accuracy.

As shown in Fig. 9, the stacked multi-object visual grasping experiment further validates the RSPFG-Net algorithm's performance in more complex scenarios, where multiple target objects are stacked together. This type of scene imposes higher demands on grasping algorithms, as the relative positions of the target objects are more compact, and some objects may be occluded or overlapped by others. The experiment demonstrates that, when facing stacked objects, the RSPFG-Net is capable of identifying suitable grasp points and successfully grasping the target objects, providing appropriate grasp points for each object. Despite the overlap or occlusion between objects, the model can still effectively predict the grasp points, allowing the collaborative robot to perform accurate grasps based on this information. In cases where occlusion between objects is severe, the RSPFG-Net successfully avoids mis grasping or grasp failure through precise grasp point optimization strategies.

C. Grasping detection in real scenarios

To assess grasping detection performance in a real-world setting, we conducted an experiment using conditions similar to those found in an actual project. The RealSense D435i depth camera grasps multiple objects from real scenes, while RSPFG-Net receives RGB images as input. The detection and recognition process outputs the grasping results of the target object with the highest confidence. As illustrated in Fig. 10, take hold of the target object and place it randomly within the workspace. Specifically, the human operator samples the randomized pose of the object by shaking the object in the box and placing it upside down in the workspace.

As shown in Fig. 11, The grasping success rate in multi-object scenes is slightly lower than in single-object scenarios. This phenomenon can be attributed to the complexity of object stacking in multi-object environments, which leads to increased depth missing values and errors. The white target objects are more likely to be misclassified as background, making distinguishing them from objects with similar background colors difficult. Furthermore, the closer proximity between multi-target objects increases the likelihood of collisions during the robotic arm's grasping process, which may cause the objects to scatter and result in grasping failure. Experiment 11 shows the three main types of RSPFG Net grabbing detection failures: (1) The gripper is blocked by other objects when approaching the object, causing some relative sliding and deviation from the original position. (2) White target objects are often misidentified as backgrounds due to errors in grasping information recognition. (3) In chaotic scenes, the three-finger gripper is more likely to be blocked by other objects, leading to grasping failure.

In contrast to point cloud-based approaches, our method circumvents the need to collect numerous point clouds for network training and eliminates the need to learn the disparity between simulation and reality.

Conclusion

To address the challenge of grasping objects with uncertain shape, attitude, scale, and stacking, this paper presents a NEU-MGD multi-target dataset of common objects in daily life. Additionally, we propose a new AGR-strategy that utilizes a directed arrow model to characterize the objects to be grasped for strategy decision-making. Directional

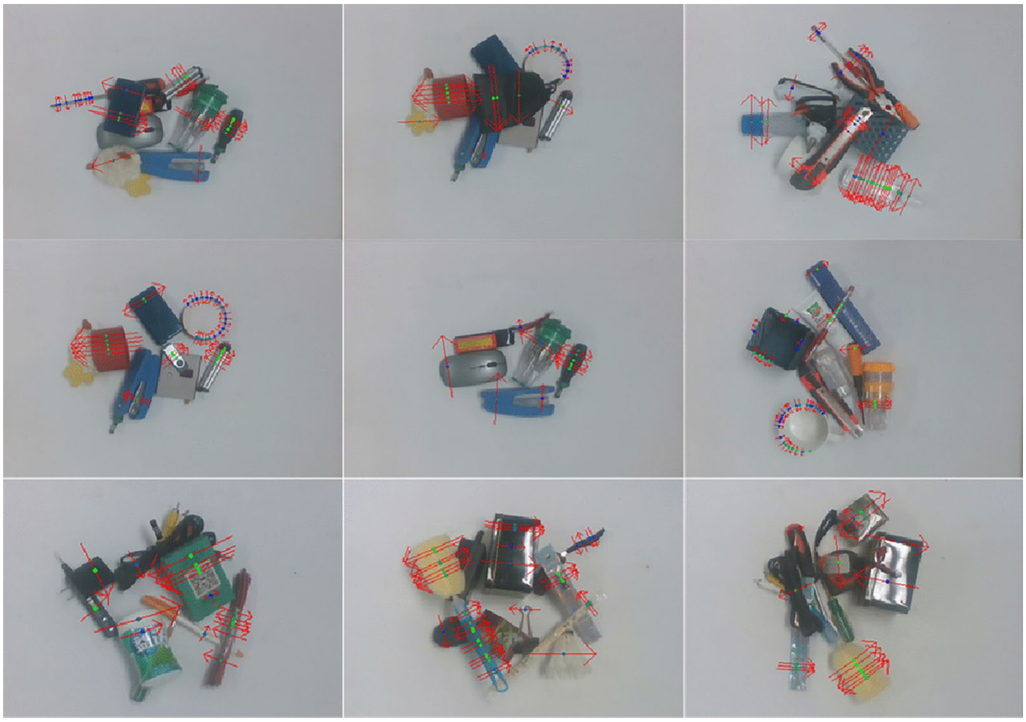

Fig. 9. Stacked multi-object data grasp detection.

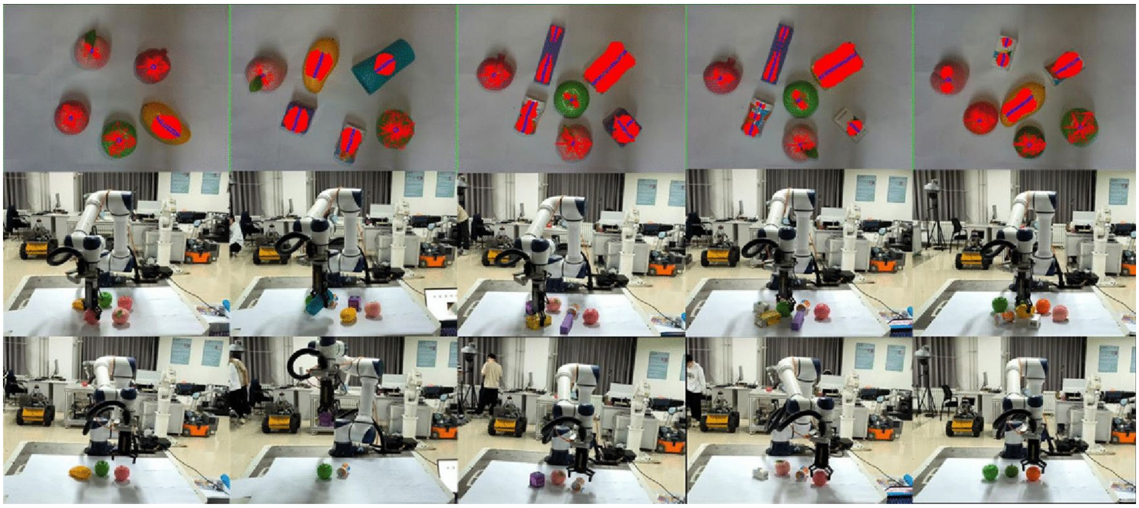

Fig. 10. Robot grasping experiment in real scenes.

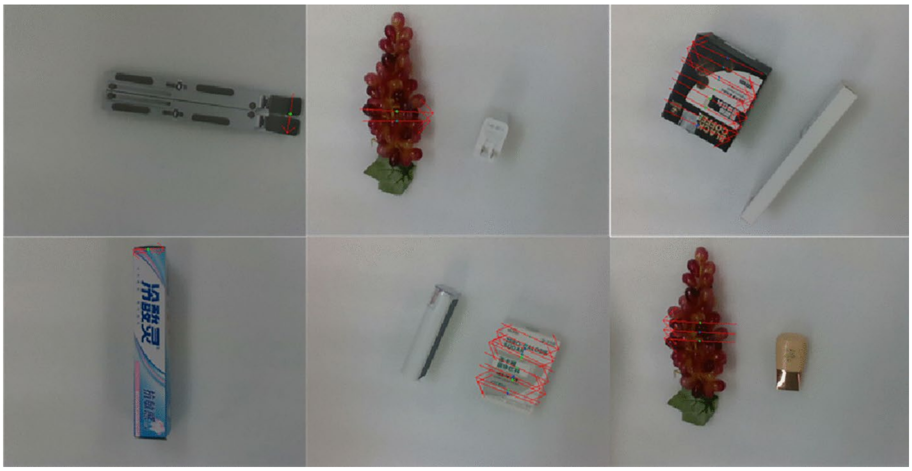

Fig. 11. Examples of grasping detection failures.

arrows model object grasping attributes, resolving possible angle conflicts in training and avoiding the complex pixel-level labeling process. Additionally, a MSRSPF method with DeepLab v3 network sequential connection of subpaths is proposed to form the RSPFG-Net network. RSPFG-Net predicts pixel-level grasping information data of RGB images to achieve high-performance detection, overcoming limitations of current deep learning grasping techniques for grasping candidate objects through discrete sampling and excessive computation time. Experiments were conducted on publicly available datasets, including Cornell and Jacquard, as well as the NEU-MGD dataset of multi-target objects proposed in this study. This text demonstrates the strong performance and robustness of our method in addressing the multi-target grasping problem. Compared with the point cloud-based approach, it avoids collecting a large number of point clouds and 6D gestures to train the network. The proposed method for grasping detection is currently limited to implementation in a two-dimensional plane. Future research will extend this approach to three-dimensional stacked multi-pose and multi-object grasping detection. It has important research significance for improving the performance of the object-grasping method.

Data availability

The datasets used and analysed during the current study are available from the corresponding author on reasonable request. corresponding. BinZhao email: zhaobin@stumail.neu.edu.cn https://github.com/SimonZhaoBin/NEU-MGD

Received: 14 November 2024; Accepted: 7 March 2025
Published online: 13 March 2025

References

1. Zhai, D.-H., Yu, S. & Xia, Y. Fanet: fast and accurate robotic grasp detection based on keypoints. *IEEE Transactions on Automation Science and Engineering* (2023).
2. Han, J., Chai, J. & Hayashibe, M. Synergy emergence in deep reinforcement learning for full-dimensional arm manipulation. *IEEE Trans. Med. Robotics Bionics* **3**, 498–509 (2021).
3. Hong, Q.-Q., Yang, L. & Zeng, B. Ranet: A grasp generative residual attention network for robotic grasping detection. *International Journal of Control, Automation and Systems* **20**, 3996–4004 (2022).
4. Fang, H.-S., Gou, M., Wang, C. & Lu, C. Robust grasping across diverse sensor qualities: The graspnet-1billion dataset. *The International Journal of Robotics Research* **42**, 1094–1103 (2023).
5. Haninger, K., Radke, M., Vick, A. & Krüger, J. Towards high-payload admittance control for manual guidance with environmental contact. *IEEE Robotics and Automation Letters* **7**, 4275–4282 (2022).
6. Wang, D., Liu, C., Chang, F., Li, N. & Li, G. High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network. *IEEE transactions on industrial electronics* **69**, 11611–11621 (2021).
7. Kumra, S., Joshi, S. & Sahin, F. Antipodal robotic grasping using generative residual convolutional neural network. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 9626–9633 (2020).
8. Ribeiro, E. G., de Queiroz Mendes, R. & Grassi, V. Jr. Real-time deep learning approach to visual servo control and grasp detection for autonomous robotic manipulation. *Robotics and Autonomous Systems* **139**, 103757 (2021).
9. Zhang, H. et al. Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter. *IEEE Robotics and Automation Letters* **7**, 2929–2936 (2022).
10. Chen, S., Tang, W., Xie, P., Yang, W. & Wang, G. Efficient heatmap-guided 6-dof grasp detection in cluttered scenes. *IEEE Robotics and Automation Letters* (2023).
11. Cao, B. et al. Real-time, highly accurate robotic grasp detection utilizing transfer learning for robots manipulating fragile fruits with widely variable sizes and shapes. *Computers and Electronics in Agriculture* **200**, 107254 (2022).
12. Le, T.-T., Le, T.-S., Chen, Y.-R., Vidal, J. & Lin, C.-Y. 6d pose estimation with combined deep learning and 3d vision techniques for a fast and accurate object grasping. *Robotics and Autonomous Systems* **141**, 103775 (2021).
13. Zhao, B. et al. Research on small sample multi-target grasping technology based on transfer learning. *Sensors* **23**, 5826 (2023).
14. Liu, H. et al. Mgbm-yolo: a faster light-weight object detection model for robotic grasping of bolster spring based on image-based visual servoing. *Journal of Intelligent & Robotic Systems* **104**, 77 (2022).
15. Laili, Y., Chen, Z., Ren, L., Wang, X. & Deen, M. J. Custom grasping: A region-based robotic grasping detection method in industrial cyber-physical systems. *IEEE Transactions on Automation Science and Engineering* **20**, 88–100 (2022).
16. Ghosh, S., Paral, P., Chatterjee, A. & Munshi, S. Rough entropy-based fused granular features in 2-d locality preserving projections for high-dimensional vision sensor data. *IEEE Sensors Journal* **23**, 18374–18383 (2023).
17. Ghosh, S., Mondal, A. S. & Chatterjee, A. Student's t-uniform mixture-based robust sparse coding model for sign language recognition from thermal images. *Measurement* **246**, 116619 (2025).
18. Ford, C. J. *et al.* Tactile-driven gentle grasping for human-robot collaborative tasks. *2023 IEEE International Conference on Robotics and Automation (ICRA)* 10394–10400 (2023).
19. Depierre, A., Dellandréa, E. & Chen, L. Jacquard: A large scale dataset for robotic grasp detection. *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 3511–3516 (2018).
20. Dong, M., Bai, Y., Wei, S. & Yu, X. Robotic grasp detection based on transformer. *International Conference on Intelligent Robotics and Applications* 437–448 (2022).
21. Yu, S., Zhai, D.-H., Guan, Y. & Xia, Y. Category-level 6-d object pose estimation with shape deformation for robotic grasp detection. *IEEE Transactions on Neural Networks and Learning Systems* 1857–1871 (2023).
22. Gilles, M. et al. Metagraspnetv2: All-in-one dataset enabling fast and reliable robotic bin picking via object relationship reasoning and dexterous grasping. *IEEE Trans. Autom. Sci. Eng.* **21**(3), 2302–2320 (2023).
23. Lee, S.-K., Myung, H. & Kim, J.-H. Mmh-ggcnn: Multi-modal hierarchical generative grasping convolutional neural network. *International Conference on Robot Intelligence Technology and Applications* 422–430 (2021).
24. Czajkowska, J., Badura, P., Korzekwa, S. & Płatkowska-Szczerek, A. Automated segmentation of epidermis in high-frequency ultrasound of pathological skin using a cascade of deeplab v3+ networks and fuzzy connectedness. *Computerized Medical Imaging and Graphics* **95**, 102023 (2022).

Acknowledgements

This research was funded by the National Natural Science Foundation of China under Grants (U20A20197), the Provincial Key Research and Development for Liaoning under Grant(2020JH2/10100040). Development and application of autonomous working robots in large scenes(02210073421003). Research and development of automatic inspection flight control technology for satellite denial environment UAV(02210073424000)

Author contributions

B.Z. and L.C. conceived the experiment(s), L.C. and Z.L conducted the experiment(s), B.Z. and C.W. analysed the results. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to B.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.