SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

March 2025, Vol. 68, Iss. 3, 132203:1–132203:16
https://doi.org/10.1007/s11432-024-4157-7

MetaCoorNet: an improved generated residual network for grasping pose estimation

Hejia GAO[1,2,4], Chuanfeng HE[1,2,4], Junjie ZHAO[1,2,4] & Changyin SUN[1,2,3,4*]

[1]*School of Artificial Intelligence, Anhui University, Hefei 230601, China*
[2]*Engineering Research Center of Autonomous Unmanned System Technology, Ministry of Education, Hefei 230601, China*
[3]*School of Automation, Southeast University, Nanjing 214135, China*
[4]*Anhui Provincial Key Laboratory of Security Artificial Intelligence, Anhui University, Hefei 230601, China*

Received 9 February 2024/Revised 14 May 2024/Accepted 10 September 2024/Published online 20 January 2025

Abstract Robotic grasping presents significant challenges due to variations in object properties, environmental complexities, and the demand for real-time operation. This study proposes the MetaCoorNet (MCN), which is a novel deep learning architecture specifically designed to address these challenges in robotic grasping pose estimation. By combining spatial and channel operators, the MetaCoor block is utilized to extract features efficiently. This architecture enhances feature selectivity by embedding location information into channel attention using a positional embedding technique within the coordinate attention mechanism. Consequently, the proposed MCN can focus on pertinent grasp-related regions. Furthermore, convolutional fusion blocks seamlessly integrate spatial and channel features, resulting in enhanced feature resolution and representation capabilities. This innovative design enables the proposed MCN to achieve state-of-the-art performance on the Cornell and Jacquard datasets, attaining accuracies of 98% and 91.2%, respectively. The effectiveness and robustness of MCN are further validated through real-world experiments conducted using a seven-degree-of-freedom Kinova manipulator.

Keywords generative ResNet, meta light block, coordinate attention, feature resolution, robot grasping

Citation Gao H J, He C F, Zhao J J, et al. MetaCoorNet: an improved generated residual network for grasping pose estimation. Sci China Inf Sci, 2025, 68(3): 132203, https://doi.org/10.1007/s11432-024-4157-7

1 Introduction

Robotic grasping, a fundamental challenge in robotics, involves the intricate interaction between robots and their environment, enabling a range of applications from industrial automation to service tasks [1]. While the potential benefits are vast, spanning across domains like manufacturing [2], domestic assistance, and component assembly [3], achieving robust and reliable grasping remains a complex endeavor. The inherent variability of objects in terms of shape, size, and material properties, coupled with environmental uncertainties such as occlusions and lighting variations, presents significant challenges for the robotic grasping research [4]. Furthermore, sensor noise and the intricacies of gripper mechanics introduce additional layers of complexity. Therefore, developing grasp pose estimation methods that are robust to these uncertainties and efficient in execution is crucial for advancing robotic grasping capabilities [5].

Grasping pose estimation has emerged as a keystone technique [6]. As the compass guiding robotic interaction, pose estimation [7] delivers the data required for a robot to ascertain the most favorable points of contact angles and positions, where the interplay of the robotic gripper and object results in the most stable and effective grasp. Conceiving this process as a regression challenge [8], where visual inputs, e.g., images or point clouds [9], morph into calculated grasping positions, has been made possible through advancements in deep learning methods [10]. These techniques can unravel the complex mappings [11] within the grasp-related data; however, computational costs and limited generalizability hinder current methodologies [12]. For example, contemporary methods either consume considerable resources with intensive deep networks or exhibit limited application ranges due to various factors, including the characteristics of the target objects or limited data availability. Our research presents a pathway to achieve unparalleled precision and adaptability, i.e., the MetaCoorNet, which represents a profound innovation

* Corresponding author (email: cysun@ahu.edu.cn)

in the field of deep learning for robotic grasping that strives for exceptional accuracy and efficiency across an array of scenarios and tasks, from seizing uncharted objects [13] to mastering the chaos of cluttered spaces [14] or articulating multi-fingered grips.

Thus, this study proposes an efficient neural network for the grasp pose estimation task. The proposed the MetaCoorNet (MCN) method calculates the grasping pose quickly and efficiently, and it builds upon the generative residual convolutional neural network (GRCNN) [15] and introduces several key innovations to enhance performance and robustness [16]. Specifically, the proposed MCN incorporates the meta light block (MLB) [17], which is a convolutional block that combines spatial and channel operators to enhance the feature extraction and computation efficiency of the network. In addition, coordinate attention (CA) [18], i.e., an attention mechanism that applies coordinate-wise attention to both the spatial and channel dimensions, is implemented in the proposed MCN. This allows the network to focus on the important regions and features for grasping. Furthermore, the convolution fusion block (CFB) module is implemented to enhance the feature extraction and fusion process. Using these modules, the proposed MCN can generate high-quality and diverse grasp poses from n-channel input images, e.g., RGB-D images. Compared with the existing grasp pose estimation networks, the proposed MCN achieves better grasp quality with less training time and faster reasoning speed.

The proposed MCN was evaluated experimentally on two public datasets, i.e., the Cornell and Jacquard datasets, and the performance of the MCN was compared with various baseline models that include analytical and empirical methods. To verify the rationality of the proposed method, we conducted extensive experiments and analyzed the impact of the MetaCoor block (MCB) and residual block on the results. These experiments focused on grasping household items, and they considered several relevant limitations and challenges [19], including single-object grasping tasks, multiobject grasping tasks [20] in complex scenes, and transparent objects [21]. The experimental results demonstrate that the proposed MCN performs well in terms of several performance indicators. For example, the proposed MCN exhibited competitive or excellent performance in terms of grasp accuracy and grasp speed. In addition, results demonstrate that the proposed MCN exhibits good generalizability and proficiency when dealing with various fetching challenges and complexities. This empirical evidence strengthens the feasibility and applicability of the MCN network. Similarly, the proposed MCN can handle other challenges and limitations, e.g., sensitivity to noise and occlusion, scalability to different types and shapes of claws, and effective generalization of transparent objects [22] and environments.

The primary contributions of this study are summarized as follows.

• Designing the MCB. The MCB module combines spatial and channel operators to enhance the feature extraction process and computation efficiency of the network, thereby resulting in a lighter and faster network that maintains sufficient performance.

• Designing coordinate attention with residual blocks. The proposed network incorporates CA, which embeds location information into channel attention, thereby enhancing feature selectivity. This mechanism is combined strategically with the residual blocks to further improve the information flow and prevent vanishing gradients, which ultimately results in improved accuracy and stability during training.

• Proposing the MCN architecture. The architecture of the proposed MCN effectively combines the MCBs, CA, residual blocks, repetition space operator (RepSO), thinning channel operator (RefCO), and CFB modules to realize robust grasp pose estimation. This architecture provides a lightweight but powerful solution that can predict grasp poses for diverse objects in cluttered scenes accurately, achieving state-of-the-art performance while maintaining efficiency suitable for practical real-time robotic applications.

The remainder of this paper is organized as follows. Section 2 reviews previous studies on robot grasping and deep learning-based grasping attitude estimation methods. Section 3 introduces the main problem formula and grasping pose representation used in the study, and Section 4 introduces the proposed network architecture, including its innovative component modules, in detail. Section 5 reports the results of experiments and ablation studies conducted on two benchmark datasets, and Section 6 discusses a thorough simulation and corresponding experiment (a flowchart of the experiment is shown in Figure 1). Finally, the paper is concluded in Section 7, including suggestions for potential future work.

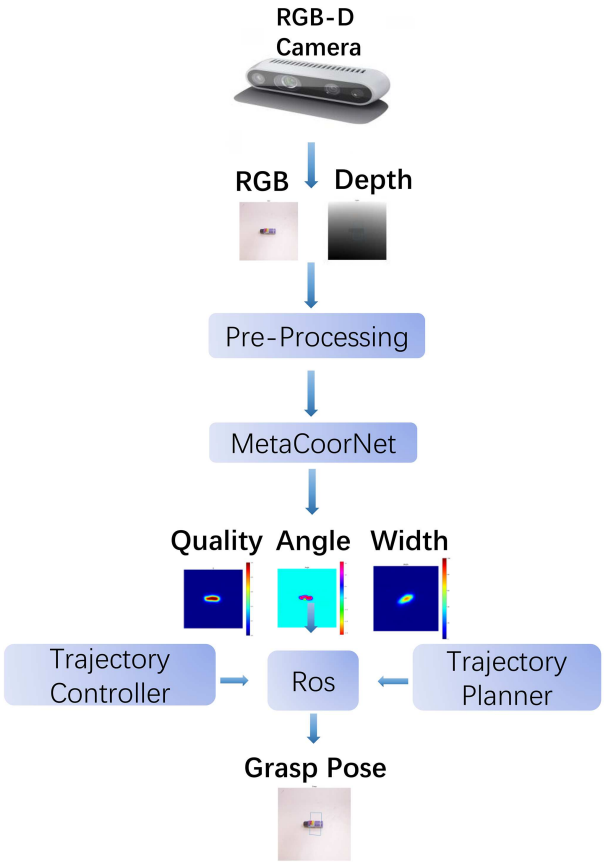

Figure 1 (Color online) Overall flow of grasping process.

2 Related work

Deep neural network-based robot grasping methods. Robot grasping is a challenging and active research field in the robot technology domain [23]. The goal of grasping is to enable robots to operate objects in various scenes [24], e.g., industrial, home, and medical environments [25]. To achieve stable and robust grasping, a previous study [26] analyzed and modeled the physical and geometric interactions between the grasp and the target object. In addition, Zhou et al. [27] studied a robot clutter target detection and acquisition method based on a cascaded depth convolution neural network. This method can be applied to the target capture scene in cluttered scenes and exhibits strong stability and robustness. In terms of using deep neural networks for grasp pose estimation, Yu et al. [28] proposed SE-ResUNet, which achieved high accuracy (98.2% and 95.7% on the Cornell and Jacquard datasets, respectively) and real-time performance (30 fps). This approach integrates residual blocks with channel attention to realize efficient grasp pose and quality prediction. Wu et al. [29] proposed an unanchored grasp detector based on a fully convolutional network, which was employed to detect multiple effective fetches. Compared with the former method, this method requires fewer parameters. However, robot grasping remains an open and challenging problem that requires further research and development. The current limitations and future developments focus on improving the pseudo-realistic transformation of the grasping model to enhance the generalizability and robustness of grasping algorithms by combining multimodal sensory feedback and uncertainty estimation, in addition to combining the grasping tasks with other higher-level tasks, e.g., planning, reasoning, and communication. Robot grasping technology is of great significance because it allows robots to perform a wide range of high-utility tasks that are beneficial to humans and society.

Generative adversarial network-based methods. Methods based on generative adversarial networks generate the grasping pose and control strategy of the manipulator using a small amount of data, e.g., image and depth data, using the confrontation training of the generator and discriminator. The goal of this method is to achieve innovative and diverse grasping that can adapt to various scenarios

and objects. The advantage of methods based on generative adversarial networks is that they can utilize a small amount of data to enhance the data and improve the diversity of grasping, which can reduce the dependence on large-scale and expensive data collection and annotation processes. However, such methods must balance the training of generators and discriminators, and they must also avoid pattern collapse [30] and pattern loss, which are common challenges in generative adversarial networks. For example, a previous study [31] proposed a modular robot system to solve the problem of generating and executing pedal robot grasping of unknown objects from n-channel images of the target scene. They proposed a generated residual convolution neural network (GR ConvNet) model that can generate robust foot grasping from the n-channel input in real-time (20 ms). The GR ConvNet model comprises a residual encoder-decoder network and a grasp quality network, which can learn the features of the input images and generate high-quality antipodal grasps. They evaluated the proposed model architecture on standard datasets and a different set of family objects. On the Cornell and Jacquard grasping datasets, their method achieved state-of-the-art accuracies of 97.7% and 91.2%, respectively. Ref. [32] proposed a deep convolution generated adversarial kinematics network (DCGAKN) to establish the inverse kinematics [33] of self-assembled manipulators. Here, they designed a robot system that utilizes a depth sensor and the YOLOv4 algorithm for object detection in a preliminary step prior to grasping. Then, they proposed the DCGAKN with an inverse kinematics model generator and discriminator with antagonistic evolutionary training, which was employed to control the self-assembled manipulator to solve the limited solution space and make the manipulator more adaptive in the dynamic environment. This DCGAKN model can learn the inverse kinematics mapping from the end-effector pose to the joint angles without requiring prior knowledge or analytical solutions. Their findings demonstrated the effectiveness of the DCGAKN model for various self-assembled manipulators and tasks.

Supervised learning-based grasping method. Conventional methods based on supervised learning [34] train the grasping model of the manipulator using the classification, regression, and convolution neural network techniques [35] according to tagged data, e.g., grasping point, grasping angle, and grasping success rate data, to achieve efficient and accurate grasping. The advantage of supervised learning methods is that they can utilize existing data, reduce manual intervention, and improve grasping performance, which can enhance the reliability and robustness of the manipulator. The disadvantage of such methods is that they require large amounts of labeled data, as well as appropriate network structures and parameters, which can increase the complexity and cost of the data preparation and model training processes. For example, a previous study [36] derived a learning algorithm from the application perspective to reduce the amount of requirement training data significantly. The main improvements in terms of time and data efficiency are summarized as follows. First, the geometric consistency between the lossless depth image [37] and the task space is utilized, which can avoid the distortion and noise caused by the image projection and compression. This study used a relatively small full convolutional neural network to predict the grasp [30] and grasper [38] parameters, which reduced the computational and memory costs and improved the training and inference performance. Second, under the incentive of approximately 3% of the success rate of small random grasping, the grasping space was explored systematically, which can increase the diversity and coverage of the data samples. In approximately 60 h, the final system was trained on 23000 grip attempts, increasing the current solution by an order of magnitude, which can demonstrate the efficiency and effectiveness of the learning algorithm. For a typical dustbin sorting scenario [39], the capture success rate was measured to be 96.6%, which demonstrates the high accuracy and precision of the grasping model. Further experiments demonstrated that the system can generalize and transfer knowledge to new objects and environments, which proves the adaptability and scalability of the system. In a previous study [40], to reduce the time costs associated with data acquisition and annotating and improve the grasping success rate, Peng et al. developed a self-supervised learning mechanism to control the grasping tasks performed by the manipulator. First, the manipulators automatically collected point clouds of objects from multiple angles to improve the efficiency of data collection, which can reduce human labor and intervention costs. The complete point cloud of the object was obtained using the hand-eye vision of the manipulator [41] and the truncated symbol distance function algorithm [42], which can reconstruct the 3D shape and pose of the object from partial and noisy observations. Then, a series of six-degree-of-freedom grasping poses was generated using the object's point cloud data, and the force closure decision algorithm [43] was employed to add grasping quality labels for each grasping posture to realize the automatic data marking, which can eliminate the need for human annotations and ensure the consistency and reliability of the data labels. Finally, the point cloud in the closed area of the gripper corresponding to each grasp pose was obtained and used to train the grasping quality classification

model [44] of the manipulator, which can learn the features and patterns of successful and unsuccessful grasps. The experimental results demonstrated that the proposed self-supervised learning method can improve the success rate of manipulator grasping, which indicates the feasibility and applicability of the method.

Existing grasping pose estimation methods can generally be divided into the above categories. Although the above methods can solve most grasping pose estimation problems, they frequently struggle when handling complex real-world scenes. For example, although supervised learning-based methods have shown promising results, they require large amounts of labeled data and can lack generalizability. In addition, deep learning methods are powerful tools; however, the current models face challenges in terms of incorporating physical constraints and handling different scraping scenarios. Differing from these previous methods, the method proposed in this paper comprises several key modules that integrate sparse concepts, e.g., the CA, MLB, RepSO, RefCO, and CFB modules. The CA module integrates location information into the channel attention [45] to improve the selectivity and expressiveness of the features. This allows the network to focus on relevant features for grasping, e.g., object edges and corners, which are crucial for determining stable grasp poses. This improves the selectivity and expressiveness of the features, leading to better grasp pose estimation. In addition, the MLB module combines spatial and channel operators to improve the feature resolution and representation efficiency. By capturing and combining low-level and high-level features, the MLB modules can form a rich and compact representation that reduces redundancy and increases the information content of the features. The RepSO module applies multiple spatial convolutions to capture complex patterns. For example, the RepSO module can learn to extract fine-grained and diverse features that are useful for grabbing, e.g., edges, corners, and textures, which can increase the variety and specificity of the features. Furthermore, the RefCO module utilizes an attention mechanism to refine the features. For example, the RefCO module can learn to suppress noise, enhance the signals in the features, and improve feature quality and robustness, which can eliminate interference and improve the accuracy and stability of the features. In addition, the multilevel convolution structure is employed in the CFB module to improve the network's representation ability. For example, the CFB module can combine the characteristics of different stages and sizes to generate comprehensive and consistent representations, which ensures the coherence and completeness of the representations.

3 Problem formulation

Robotic grasping is a fundamental task in robotics that involves finding a suitable grasp pose for an object in a given scene. A grasp pose is a configuration of the robot's end-effector, e.g., a gripper or a suction cup, that can successfully pick up and manipulate the object. Robotic grasping is challenging because it requires dealing with various shapes, sizes, textures, and poses of objects, as well as uncertainties in perception and control, e.g., sensor noise, occlusion, and slippage.

This study tackles the robotic grasping problem using a vision-based approach that takes an n-channel image of the scene as the input. The image can contain depth, color, segmentation, or other information that can help the robot locate and identify the objects, as well as their geometric and semantic features. The output is the predicted grasp pose for the target object in the image, which can be used to control the robot's arm and gripper. The predicted grasp pose comprises the position, orientation, and width of the end-effector, as well as the grasp quality score, which indicates the likelihood of performing a successful grasp.

The grasp pose in the robot's frame of reference is expressed as follows using a four-tuple:

$$S_r = (T, U_r, V_r, W), \tag{1}$$

where $T = (x, y, z)$ is the position of the end-effector's center, U_r is its orientation around the z-axis, V_r is the width of the gripper, and W is the grasp quality score. The quality score is a scalar that reflects how likely the grasp is to succeed based on several criteria, e.g., stability, robustness, or clearance.

The predicted grasp pose in the image's frame of reference is expressed as follows:

$$S_i = (x, y, U_i, V_i, W), \tag{2}$$

where (x, y) is the center of the grasp, V_i is the width of the gripper in the image plane, U_i is the orientation of the grasp in the camera's frame of reference, and W is the same scalar as in (1).

Table 1 Whole process of MCN network propagation.

Network propagation stage	Dimension change	Propagation effect
Input layer	n-channel image	Receive preprocessed image information
Feature extraction layer	128 filters	Extract image features through MCB and CA
Feature fusion layer	32 filters	Fuse spatial and channel features through RepSO, RefCO, and CFB
Output layer	4 single-channel probability maps	Predict grasp angle, width, and quality

The predicted grasp pose S_i in the image coordinate system is transformed into the robot coordinate system S_r by applying a series of transformations, shown as

$$S_r = X_{cr} X_{ic}(S_i), \tag{3}$$

where X_{ic} is the transformation matrix that converts a grasp pose S_i in the image space to the camera's 3D space and X_{cr} is the transformation matrix that converts the camera space to the robot space. These transformation matrices [46] can be obtained by calibrating the camera and the robot using standard methods.

To find a suitable grasp pose for a given object in the scene, a vision-based approach is proposed to indicate the likelihood of grasp success. This method takes an n-channel image of the scene as input, and it outputs the predicted grasp pose and the corresponding grasp quality score. Here, a quadruple is used to represent the grasp pose, including the grasp center position, orientation, width, and quality score, and we transform the grasp pose from the image coordinate system to the robot coordinate system.

4 Methodology

This section describes the proposed MCN network for grasp pose estimation in detail. The proposed network is designed to generate grasping poses efficiently and robustly from n-channel input images, which can contain various information, e.g., RGB, depth, and segmentation information. The network comprises four main components, i.e., the input layer, the feature extraction layer, the feature fusion layer, and the output layer, as shown in Figure 2. The dimensional changes of the complete communication process and the roles of the communication process are shown in Table 1.

Input layer. The input layer receives the preprocessed n-channel image and applies a convolution layer with 32 filters to extract the initial features. Here, the size of the input image is 224×224 pixels, and the size of the output feature map is also 224×224 pixels.

Feature extraction layer. The feature extraction layer attempts to capture the spatial and channel dependencies of the features and embed the positional information into the channel attention. Here, two MetaCoor blocks and three residual blocks with CA layers are employed to achieve this goal. Each MCB comprises a spatial convolution layer and a CA layer, which can enhance the feature extraction and computational efficiency of the network. Each residual block comprises two convolution layers and a CA layer, which can preserve the input feature map and avoid the gradient vanishing problem. The feature extraction layer utilizes two MetaCoor blocks and three residual blocks with CA layers, and each block maintains the spatial dimensions of the feature maps while enriching the feature representation, thereby resulting in a final feature map size of 56×56 with 128 filters.

Feature fusion layer. The goal of the feature fusion layer is to integrate the spatial and channel features into a rich and compact representation and to improve the feature resolution and diversity. Here, three RepSOs, three RefCOs, and a subsequent CFB are employed to achieve this objective. Each RepSO applies multiple convolution layers with the same kernel size, stride, and padding to the feature map, which enhances the spatial information and improves the network's efficiency. Each RefCO applies an attention mechanism to refine the channel information, which can improve the discrimination ability of the network and make it capture more information features in the image processing. The CFB utilizes a series of convolution layers and batch normalization to mix the spatial and channel information, which enhances the network's representation ability. Note that the feature maps retain a size of 56×56 with 128 channels throughout both the feature extraction and feature fusion layers, which enables the network to learn complex spatial relationships while extracting progressively more intricate feature representations.

Output layer. The output layer of the MCN is responsible for predicting the crucial grasp parameters from the extracted feature map, including the grasp angle, width, and quality. To achieve this, the network utilizes three transpose convolution layers with batch normalization and the ReLU activation

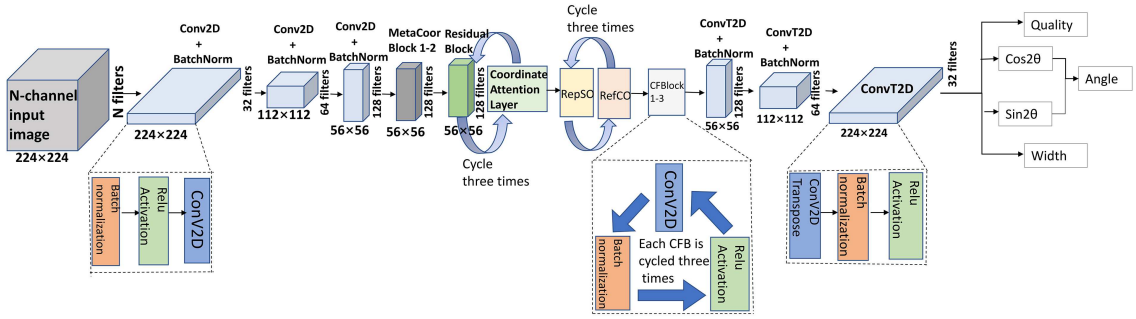

Figure 2 (Color online) Architecture of MCN. The MCN enriches the network's ability to capture detailed location information by embedding channel attention in the feature extraction layer. With the MCB, the MCN seamlessly integrates the CFB into the feature fusion layer, and the CFB focuses on merging spatial and channel features efficiently by balancing the space and channel considerations and reducing parameters and computational costs while maintaining the integrity of the feature graph. Through feature fusion, the MCN predicts the attitude of the target through the output layer accurately.

function, which is followed by four convolution layers with the sigmoid activation function. The transpose convolution layers upsample the feature map to the original image size of 224 × 224 pixels, effectively restoring the spatial resolution. Then, the four convolution layers generate single-channel probability maps, each representing the likelihood of each pixel belonging to a specific grasp attribute, i.e., the grasp quality, cosine of the grasp angle, sine of the grasp angle, and grasp width. This configuration allows the network to be trained in an end-to-end manner by combining the cross-entropy loss for grasp quality and the width with the mean squared error loss for the grasp angle (represented by its cosine and sine values).

4.1 MCB

The MCB module receives the feature map from the previous block as input and attempts to improve the feature extraction and computational efficiency by integrating spatial and channel operators. Here, the spatial convolution layer applies a 3×3 kernel to the input feature map, and it generates a spatial feature map with the same number of channels. Compared with only positional indexes, the feature map can capture richer information by allowing the network to learn more complex spatial relationships within the feature space.

Embedding. The CA layer embeds the positional information into the channel attention by adding a learnable coordinate embedding vector to each channel of the spatial feature map. This coordinate embedding vector has the same size as the spatial feature map, and it encodes the relative position of each pixel in the channel. As a result, channel attention can capture the spatial dependency of the features and focus on regions that are important for the grasping task.

Factorization. The CA layer factorizes the channel attention into two 1D feature encoding processes along the horizontal and vertical directions, respectively. The horizontal feature encoding process applies a $1\times k$ kernel to each row of the channel attention map, where k is the width of the map, and the vertical feature encoding process applies a $k\times 1$ kernel to each column of the channel attention map, where k is the height of the map. Then, the outputs of the two feature encoding processes are multiplied element-wise to generate a spatially selective attention map, which has the same size as the channel attention map. The spatially selective attention map can enhance the saliency and diversity of the features by assigning different weights to different regions of the channel attention map. Then, the CA layer multiplies the spatially selective attention map with the spatial feature map to obtain the output feature map of the MCB module. The output feature map has the same size and number of channels as the input feature map but with enhanced feature representations. Note that the MCB module is simple and efficient, and it can be flexibly plugged into classic mobile networks, e.g., MobileNetV2, MobileNeXt, and EfficientNet, to improve their performance on image classification tasks.

The structure of the MCB module is shown in Figure 3. In this study, the effectiveness and interpretability of the MCB were analyzed through ablation studies and visualizations. The ablation experiments were performed by training different numbers of MCBs and residual blocks on the Cornell dataset.

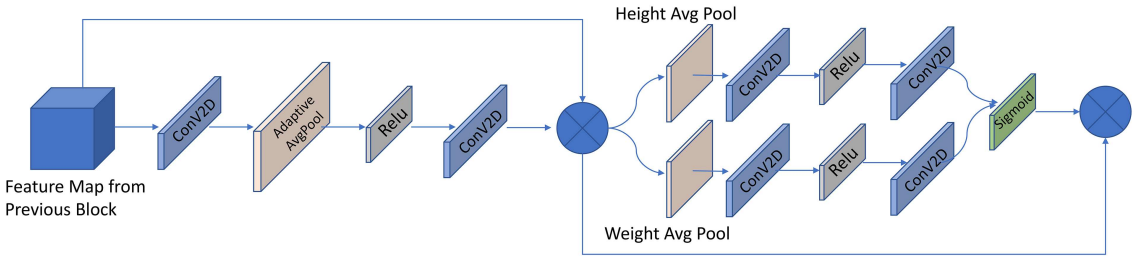

Figure 3 (Color online) Structure of the MCB.

4.2 Rest of the modules

In addition to MCB, the proposed network includes the RepSO, RefCO, and CFB modules, which are designed to enhance the spatial and channel features of the network and fuse them in a systematic manner. In addition, these modules are lightweight and efficient, which makes them suitable for real-time applications.

RepSO. The RepSO module applies multiple convolution layers with the same kernel size, stride, and padding to the feature maps. It is inspired by repeated convolution blocks, which are common building blocks in many convolutional neural networks (CNN). The RepSO module can enhance the spatial information of the feature maps and improve the network's efficiency, and it can capture complex patterns and extract fine-grained and diverse features, e.g., edges, corners, and textures, which are useful for the robot grasping task.

RefCO. The RefCO module attempts to enhance the channel relationship in feature maps in a refined manner. It is inspired by the attention mechanism and uses adaptive pooling and convolution operations to refine the channel information. By focusing on relevant channels, the goal of RefCO is to improve the discriminability of the network and make it capture more informative features during image processing. The RefCO module can also suppress noise, enhance signals, and improve feature quality and robustness.

CFB. The CFB module systematically integrates the spatial and channel features into a feature map. Here, a series of convolution layers and batch normalization are employed to mix the spatial and channel information. The goal of the CFB module is to enhance the network's representation ability through effective feature fusion. In addition, it can be used to balance computational efficiency and feature richness, which is beneficial to the overall expressiveness of the neural network. The CFB module also combines the features of different stages and scales to generate comprehensive and consistent representations, which can integrate local and global information while ensuring the coherence and completeness of the representations.

By combining these modules, the proposed network extracts and fuses spatial and channel information effectively, which results in accurate and efficient grasp pose estimation. First, the proposed MCN combines spatial convolution with CA through the MCB module to enrich feature representation using position information. This is particularly beneficial for the robot grasping task because it allows the network to focus on relevant object regions and identify grasp points for different object orientations and poses accurately. Second, enhanced feature fusion is employed in the proposed network. Here, by integrating the RepSO, RefCO, and CFB modules, the MCN can fuse the spatial and channel features effectively, thereby improving the feature resolution rate and representation. Note that a deeper understanding of object geometry is crucial for generating diverse and high-quality grasping poses for complex-shaped objects. These improvements directly address the challenges associated with grasp pose estimation, thereby leading to significant performance gains over the baseline GRCNN. In addition, the core concepts implemented in the proposed MCN, e.g., the integration of location information and the utilization of efficient feature fusion techniques, are not limited to grasping tasks. These concepts are equally beneficial in other vision tasks, e.g., object detection, semantic segmentation, and pose estimation, where understanding spatial relationships and feature dependencies plays a vital role.

5 Performance evaluation

5.1 Cornell dataset

The Cornell dataset [47] is a widely used benchmark for robotic grasp detection tasks that contains 885 RGB-D images of 240 different objects, with a total of 8019 annotated grasps. The grasps are represented by rectangles with four parameters, i.e., the center position, angle, height, and width. The dataset is divided into five splits, each containing 80% training images and 20% testing images. The Cornell dataset is challenging due to the diversity of objects, backgrounds, lighting conditions, and occlusions.

5.2 Jacquard dataset

The Jacquard dataset [48] is a large synthetic dataset for robotic grasp detection tasks. It is built on a subset of ShapeNet, which is a large dataset of CAD models, and it contains 54485 unique scenes from 11619 distinct objects, with a total of 4967454 grasps annotations. For each scene, a rendered RGB image, a segmentation mask, two depth images, and the grasps annotations are provided. The grasps are represented by rectangles with four parameters, i.e., the center position, angle, height, and width.

In our experiment, these two datasets were split randomly into training and test sets at a ratio of 9 : 1. Here, the batch size was set to 8 during training, and the Adam optimizer was used with a learning rate of 0.001. The training process comprises 50 epochs, and each epoch includes 1000 batches. The model was tested on a GeForce GTX 1660s graphics processing unit.

5.3 Grasp detection metric

To ensure a fair evaluation of the results, we employed the rectangle metric, which was introduced by Jiang et al. [47], to assess the performance of the proposed system. The rectangle metric defines the validity of a grasp based on two specific criteria.

Intersection over union (IoU) score. A grasp is considered valid if the IoU score between the predicted grasp rectangle and the ground truth grasp rectangle exceeds 25%. This measure gauges the extent of overlap between the predicted and actual grasp regions.

Orientation offset. The grasp orientation of the predicted rectangle should deviate by less than 30° from the orientation of the ground truth rectangle. This criterion accounts for the accuracy of the predicted grasp orientation.

Note that the rectangle metric requires the use of a grasp rectangle representation; however, our model generates an image-based grasp representation, denoted as S_i, using (2). Thus, to facilitate an effective comparison, each pixel value in the output image must be mapped to its corresponding rectangle representation. This conversion is crucial because it aligns the image-based grasp predictions with the requirements of the rectangle metric. To perform this conversion, we followed the approach proposed by Redmon and Angelova [49], which involves two main steps.

First, nonmaximum suppression (NMS) is applied to the output image S_i to obtain a set of candidate grasp rectangles, each with a confidence score. The NMS technique can eliminate redundant and overlapping rectangles while retaining only the most confident rectangles. Second, we select the best grasp rectangle among the candidates by applying a scoring function that considers both the confidence score and the IoU score relative to the ground truth rectangle. The scoring function is defined as follows:

$$\text{score} = \alpha \cdot \text{confidence} + (1 - \alpha) \cdot \text{IoU}, \tag{4}$$

where α is a weighting factor that balances the importance of the confidence and IoU score. We set α to 0.5 in our experiments.

6 Experiments and results

6.1 Setup

To verify the effectiveness of the proposed MCN in real-world scenarios, we conducted experiments using the Kinova robotic arm, which has a lightweight, compact design and seven degrees of freedom; thus, this device provides the dexterity and flexibility required for grasping tasks. To enable visual perception, we

Figure 4 (Color online) Kinova manipulator experimental platform.

equipped the robot with an Intel RealSense D435 stereo camera (Intel Corporation, Santa Clara, CA, USA) mounted on its hand, as shown in Figure 4. This camera captures high-resolution depth and color images of the scene, which facilitates object detection and pose estimation tasks. For grasping, we used a three-finger gripper with soft rubber fingertips. This gripper design provides sufficient versatility to handle objects of various shapes and sizes while providing an appropriate amount of friction to realize safe gripping.

6.2 Robotic grasping procedure

In this experiment, we used the RGB-D data captured by an external camera as the input to the trained MCN model. The proposed MCN can predict the best grasping attitude of any object in the target scene. A homography matrix is used to convert it from the image coordinate plane (S_i) to the robot coordinate system (S_r). Then, the inverse kinematics solution is used to calculate the joint angle of the Kinova arm and transfer it to the control system of Kinova to perform the grasping action. To evaluate the performance of the proposed system, we conducted experiments on different household items, e.g., boxes and transparent bottles with different shapes, sizes, and textures. We also introduced occlusion and clutter into the scene to increase the difficulty and authenticity of the task. To measure the success rate of the system, we defined a successful grab as lifting an object above a certain height and holding it for at least 3 s. In addition, the failure cases of the system were analyzed, and several limitations and challenges were identified, e.g., dealing with dynamic environments and extending them to invisible objects.

In this experiment, the network initially used the RGB-D data captured by the external camera as input to predict the optimal grasping attitude, opening width, and grasping quality of any object in the scene. Then, the grasping attitude was transformed from the image coordinate plane (S_i) to the robot coordinate system (S_r) using the homography matrix, where the homography matrix was estimated by the four corresponding points between the camera and the robot coordinate system. We used the

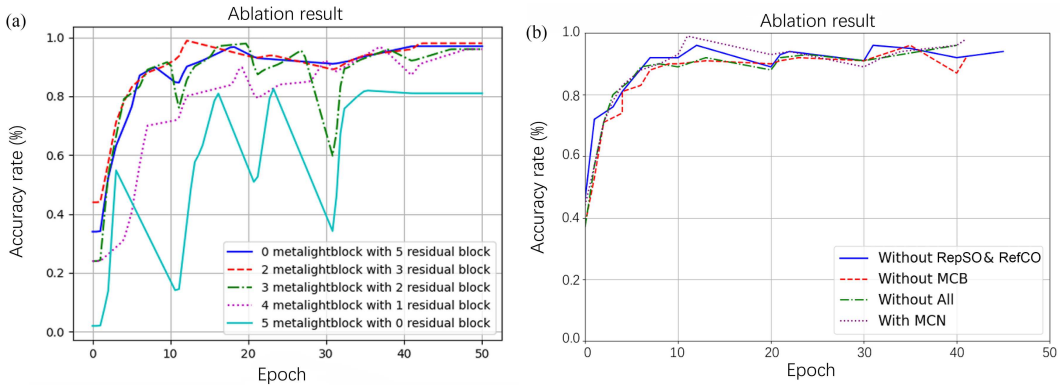

Figure 5 (Color online) Overall results of ablation tests. (a) Ablation results of different amounts of metalight blocks combined with different amounts of residual blcok; (b) ablation results of different component combinations of the feature fusion layer.

final Kinova arm, which is a lightweight and dexterous robotic arm that can perform various tasks. The Kinova device receives the desired joint angle calculated from the grasping attitude through the inverse kinematics solution, and it communicates with the control system through the robot operating system (ROS), which is a popular robot application middleware that provides a variety of communication, visualization, simulation, and debugging tools. We evaluated the proposed MCN in real-world clutter removal experiments, and the results prove that the proposed method can achieve high success rates and fast execution times without explicit collision checking or object segmentation.

To evaluate the performance of the proposed system, experiments were conducted on different household items, e.g., boxes and transparent bottles with different shapes, sizes, and textures, to increase the difficulty and authenticity of the task. Here, we conducted experiments and analyzed failure cases. To build the experimental platform, we used the Kinemia arm, an Intel RealSense D435 camera as an RGB-D sensor, and Ubuntu 20.04 and ROS Noetic as the computing units. The camera was installed on the robot arm, the object to be grasped was placed on the worktable, and a checkerboard mode was used to calibrate the camera and arm to acquire the homography matrix. A complete experimental environment was formed using Gazebo as the simulator and Rviz as the system's visualization tool.

A series of ablation studies were conducted to analyze the contribution of each module in the proposed MCN and optimize its architecture.

6.3 Impact of MCB and residual blocks

The initial ablation study (Figure 5(a)) investigated the impact of the number of MCBs and residual blocks on the network's performance. The results clearly demonstrated the importance of both components. MCB: We found that increasing the number of MCBs from zero to two improved accuracy significantly, highlighting the effectiveness of integrating spatial and channel information with positional embedding. However, further increasing the number of MCBs led to only marginal gains, which suggests that an optimal balance can be achieved with two MCBs. Residual blocks: We also found that adding residual blocks consistently improved accuracy while preventing overfitting, which highlights their value in terms of preserving information and avoiding vanishing gradients. We found that the best results were obtained using three residual blocks, and diminishing returns were observed with an increasing number of residual blocks.

6.4 Contributions of RepSO, RefCO, and CA modules

To further investigate the contributions of the individual modules implemented in the proposed MCN, we performed an additional ablation study (Figure 5(b)). Here, we compared the performance of the proposed MCN with and without the RepSO, RefCO, and CA modules. As discussed in the following, the findings confirm the essential role of each module. Without RepSO and RefCO: Removing the RepSO and RefCO modules resulted in a significant reduction in accuracy and slower convergence. This underscores their importance in capturing complex spatial patterns, refining channel information, and enhancing feature representation. Without CA: The absence of the CA module resulted in lower accuracy and reduced stability during training, which demonstrates the value of CA in focusing on relevant regions

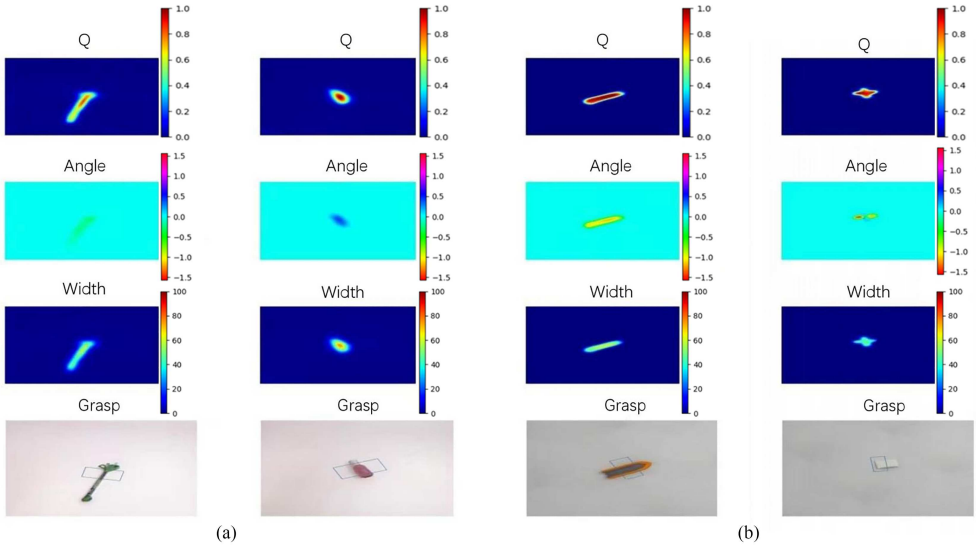

Figure 6 (Color online) Test results on Cornell and Jacquard datasets show that each column from top to bottom is grasping quality, grasping angle, grasping width, and grasping posture. (a) Occluded objects on Cornell datasets; (b) occluded objects on Jacquard datasets.

and improving feature selectivity. With MCN: The complete MCN architecture, incorporating all of the proposed modules, achieved the best performance in terms of accuracy and convergence speed. These findings emphasize the synergistic effect of combining these modules to realize optimal grasping pose estimation.

These ablation studies provide strong evidence for the effectiveness of each component implemented in the proposed MCN. The combination of MCBs, residual blocks, RepSO, RefCO, and CA enables the proposed MCN to achieve state-of-the-art performance in grasping pose estimation tasks.

In consideration of various factors, e.g., the size of the dataset and the type of training data, the Cornell and Jacquard data sets were used to evaluate the performance of the model. Here, the goal is to demonstrate the model's ability to generalize to various object categories. Note that the proposed MCN shows a strong ability to capture isolated objects at a single time, and it also exhibits a strong ability to predict objects that appear in cluttered scenes. The versatility of the model is emphasized by its adaptability to different object scenarios and its potential in real-world applications. We compared our model and several advanced methods on the two datasets, and the results demonstrate that our model achieved excellent performance in terms of accuracy and robustness. We also provide some qualitative results to illustrate the effectiveness of the model when handling challenging cases. Figure 6 shows the results obtained by our model on the Cornell and Jacquard datasets, where each column (from top to bottom) represents grasping quality, grasping angle, grasping width, and grasping posture. As can be seen, our model realized accurate and diversified capture of all kinds of objects, and the model could handle objects with different scales and directions, as well as different lighting conditions and backgrounds.

The qualitative results of various household items are shown in Figure 7, and the results prove the remarkable performance of the proposed model. The described grasping representation includes key parameters, e.g., the grasp quality measure Q, the desired grasping angle, and the required grasping width. To ensure that the model is sufficiently adaptable, we also conducted experiments in different scenarios, including single-object grabbing, multi-object manipulation, and handling transparent objects. The experimental results demonstrate that the model has good generalizability and accuracy. We found that the model can handle different and complex object categories, even in the presence of occlusion and clutter. We also observed that the model can handle objects of different scales and directions, as well as different lighting conditions and backgrounds.

To gain visual insight into the prediction of the model, we projected the rectangular capture representation onto the RGB image to enhance the interpretability of the capture prediction. The rectangular capture representation technique divides a geographic space into cell grids, where each cell represents the possible location and direction of the capture. RGB images are data inputs from a single perspective that may contain multiple objects in a busy scene. Grasping prediction is the output of the neural network, which can estimate the best grasping posture of each object instance. This visual representation helps us

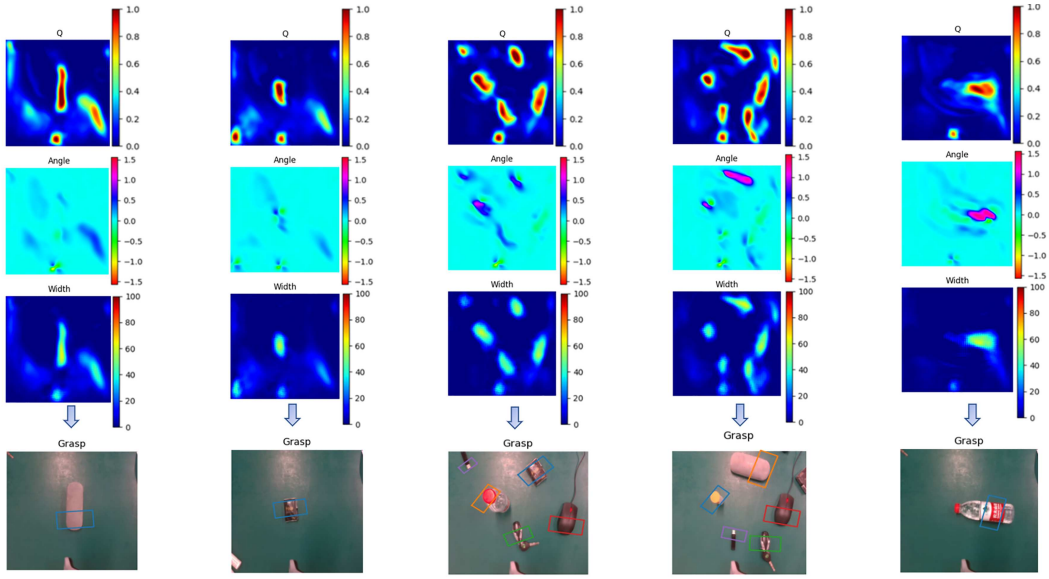

Figure 7 (Color online) Images showing good grasping results in single-object tasks, multi-object tasks, and transparent object grasping tasks.

Table 2 Comparison of accuracy on the Cornell and Jacquard datasets.

Author	Cornell accuracy (%)	Jacquard accuracy (%)	Algorithm
Cheng et al. [6]	93.3	89.6	Feature Pyramid Network
Kumra et al. [31]	97.7	94.6	GRCNN
Zhang et al. [4]	92.5	83.8	Region of Interest
Cao et al. [19]	97.8	95.6	Gaussian kernel
Wang et al. [45]	97.99	94.6	TF-Grasp
Our	98.88	91.2	MCN

to understand the model's decision-making process and promotes the transparency of the prediction. It also allows us to compare the performance of the model with other methods, e.g., communication-based and template-based methods. By projecting the rectangular capture representation onto the RGB image, we can observe how the model handles occlusion, blur, and uncertainty in the scene and how it selects the most appropriate grasp for each object.

In addition, to assess the model's performance comprehensively, we conducted comparative evaluations with other state-of-the-art methods. Evaluating the proposed MCN on different tasks with various types of objects established its effectiveness and capability. The comparison with existing methods is shown in Table 2, detailing the results of our experiments on the Cornell and Jacquard datasets. The findings strongly validate the proposed MCN's advantages in terms of learning feature extraction, emphasizing its superior ability and accuracy in object grasping. These results position the proposed MCN as a leading solution for diverse and challenging grasping scenarios. In addition, to measure the performance of the model comprehensively, we also performed a speed comparison test with other models.

Further substantiating the efficiency of the proposed MCN, we performed a comparative analysis of its inference speed against several established grasp detection models. As shown in Table 3 [6, 20, 27, 29, 31, 35, 45, 50], the proposed MCN exhibits exceptional performance, achieving an inference time of only 20 ms per image. This places it on par with the fastest of the compared models, and it surpasses the speed of many other prominent architectures. This remarkable efficiency, combined with its high accuracy, underscores the suitability of the proposed MCN for real-time robotic grasping tasks where swift and precise responses are critical.

To further prove the robustness and generalization of the proposed MCN, we conducted experiments in real scenes. Here, we collected a group of images from different environments, e.g., offices, kitchens, and warehouses, that included objects of different shapes, sizes, and textures. We applied the proposed MCN to these images and generated the grab pose performed by the robot arm. The success rate of MCN in these real-world grasping tasks was 93.6%, which is equivalent to or higher than that of traditional

Table 3 Speed comparison.

Author	Speed (ms)	Algorithm
Zhou et al. [27]	117	FCGN
Wang et al. [45]	–	Transformer
Karaoguz et al. [20]	200	CNN
Cheng et al. [6]	–	Feature Pyramid Network
Hu et al. [35]	40	GGS-CNN
Wu et al. [29]	26	FCN
Kumra et al. [31]	20	GRCNN
Qin et al. [50]	23	AE-GDN
Our	20	MCN

Table 4 Comparison of grasping success rates of different grasping networks in real-robot experiments.

Author	Success rate (%)	Algorithm
Breyer et al. [51]	80.0	VGN
Wang et al. [52]	93.6	AFFGA-Net
Yu et al. [53]	93.6	EGNET
Lenz et al. [54]	84.0	GraspCNN
Morrison et al. [55]	83.0	GG-CNN
Our	93.6	MCN

classical grasping networks.

To further validate the robustness and generalizability of the proposed MCN, we conducted experiments using the Kinova robotic arm equipped with an Intel RealSense D435 depth camera. Here, we collected image data of various household objects and introduced noise to simulate challenging grasping scenarios. Then, the proposed MCN and other grasping networks, including the VGN, AFFGA-Net, EG-Net, and GraspCNN networks, were used to predict grasp poses for the collected data. Based on these predictions, the robotic arm performed grasping actions, and we recorded the success rates obtained by each network. As shown in Table 4 [51–55], the results demonstrate that the proposed MCN achieved a competitive grasping success rate compared to the other state-of-the-art grasping networks, highlighting its effectiveness and practicality in real-world applications.

The experimental results demonstrate that the proposed MCN is effective and accurate for open datasets and that it is robust and generalizable to real-world data. The proposed MCN can handle a wide range of objects and environments to achieve fast and reliable grasp prediction, and it can produce a stable variety of high-quality grasps, covering different directions and positions of objects. The comparison of the results has shown that the proposed MCN can adapt to various grasping conditions and requirements, and it provides greater flexibility and robustness for the robot arm. The MCN demonstrates strong adaptability; thus, it can effectively learn the characteristics and representations of objects and grasps, achieve the most advanced performance in tasks involving grasping and operating objects, and generate the optimal grasping posture for the robot arm. Thus, we consider that the proposed MCN is a promising object grasping and operation model that can be applied to a wide range of applications, including industrial automation, service robots, auxiliary technology, and other fields.

7 Conclusion

This paper proposed a novel and efficient neural network for grasp pose estimation, called MCN, which can generate high-quality and diverse grasps for various objects in cluttered scenes. MCN is composed of four main parts: input layer, feature extraction layer, feature fusion layer, and output layer. In the feature extraction layer, we have introduced a new module, MCB, which combines spatial and channel operators to enhance the feature extraction and computation efficiency of the network. In the feature fusion layer, we have employed CA and CFB to improve the selectivity and expressiveness of the features, as well as the resolution and representation ability of the features. In the output layer, we have predicted the angle, quality, and width of the grasp pose. We have evaluated our MCN on two benchmark datasets, Cornell and Jacquard, and have achieved state-of-the-art performance in terms of accuracy and robustness. We have also conducted ablation studies and visualizations to analyze the effectiveness and interpretability

of our MCN. Moreover, we have tested our MCN on a real robotic arm, Kinova, and demonstrated its feasibility and applicability in various scenarios, such as single-object grasping, multi-object manipulation, and transparent object handling. Our experimental results show that our MCN can produce accurate and efficient grasps for different object categories, even in the presence of occlusion and clutter.

The proposed MCN provides a unique approach to grasp pose estimation by effectively incorporating both spatial and channel information. By embedding positional information into the channel attention mechanism through the innovative MCB, our network can prioritize crucial regions and features relevant to grasping, leading to enhanced accuracy and selectivity. This attention to spatial details allows MCN to excel in complex scenarios with cluttered backgrounds or occlusions. Furthermore, the utilization of lightweight modules such as MCB, CA, and CFB optimizes the network's architecture, reducing the number of parameters while maintaining or even improving performance compared to existing methods. This focus on efficiency makes MCN particularly suitable for real-time robotic applications where speed and accuracy are paramount.

Despite MCN's excellent performance in grasping pose estimation, there are still some limitations that need to be addressed. First, the model training mainly relies on synthetic data, which may not fully reflect real-world challenges such as sensor noise, lighting conditions, and object deformation. Second, MCN employs a fixed grasp representation, which may not be suitable for different types of end-effectors such as suckers or soft claws.

In the future, we will focus on the following directions to improve MCN.

• Incorporating real-world data. Combine real-world data with synthetic data to improve the robustness and generalization ability of the model in real environments.

• Developing adaptive grasp representation. Design grasp representation methods that can adapt to different end-effectors and object shapes.

• Integrating physical and geometric constraints. Incorporate robot kinematics, dynamics, and environmental constraints into grasp pose estimation for more reliable grasp planning and execution.

• Exploring multi-object grasping. Extend MCN to multi-object scenarios to achieve simultaneous grasping and manipulation of multiple objects.

We hope that our MCN will enable robots to interact with the real world and accomplish a variety of tasks that are beneficial to human beings and society.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant Nos. 62388101, 62303010), University Synergy Innovation Program of Anhui Province (Grant No. GXXT-2023-039), and Anhui Provincial Key Research Program of Universities (Grant No. 2022AH050087).

References

1 Babin V, Gosselin C. Mechanisms for robotic grasping and manipulation. Annu Rev Control Robot Auton Syst, 2021, 4: 573–593
2 Li D Y, Ma N, Gao Y. Future vehicles: learnable wheeled robots. Sci China Inf Sci, 2020, 63: 193201
3 Zhao H, Zhu C Y, Xu X, et al. Learning practically feasible policies for online 3D bin packing. Sci China Inf Sci, 2022, 65: 112105
4 Zhang H, Lan X, Bai S, et al. ROI-based robotic grasp detection for object overlapping scenes. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019. 4768–4775
5 Fu J J, Lv Y Z, Yu W W. Robust adaptive time-varying region tracking control of multi-robot systems. Sci China Inf Sci, 2023, 66: 159202
6 Cheng H, Wang Y, Meng M Q H. A vision-based robot grasping system. IEEE Sens J, 2022, 22: 9610–9620
7 Zhao C H, Fan B, Hu J W, et al. Homography-based camera pose estimation with known gravity direction for UAV navigation. Sci China Inf Sci, 2021, 64: 112204
8 Relaño C, Muñoz J, Monje C A. Gaussian process regression for forward and inverse kinematics of a soft robotic arm. Eng Appl Artif Intell, 2023, 126: 107174
9 Ren D Y, Wu Z Y, Li J W, et al. Point attention network for point cloud semantic segmentation. Sci China Inf Sci, 2022, 65: 192104
10 Bergamini L, Sposato M, Pellicciari M, et al. Deep learning-based method for vision-guided robotic grasping of unknown objects. Adv Eng Inf, 2020, 44: 101052
11 Ning B, Dong H R, Gao S G, et al. Distributed cooperative control of multiple high-speed trains under a moving block system by nonlinear mapping-based feedback. Sci China Inf Sci, 2018, 61: 120202
12 Feng J P, Wang X G, Liu W Y. Deep graph cut network for weakly-supervised semantic segmentation. Sci China Inf Sci, 2021, 64: 130105
13 Song Y N, Gao L, Li X Y, et al. A novel vision-based multi-task robotic grasp detection method for multi-object scenes. Sci China Inf Sci, 2022, 65: 222104
14 Xi L L, Peng Z H, Jiao L, et al. Smooth quadrotor trajectory generation for tracking a moving target in cluttered environments. Sci China Inf Sci, 2021, 64: 172209
15 Park S, Shin Y G. Generative residual block for image generation. Appl Intell, 2022, 52: 7808–7817
16 Chen H X, Huang X Y, Wu W, et al. Efficient and secure image authentication with robustness and versatility. Sci China Inf Sci, 2020, 63: 222301
17 Chen F, Li S, Han J, et al. Review of lightweight deep convolutional neural networks. In: Proceedings of Archives of Computational Methods in Engineering, 2023. 1–23
18 Xiao M, Yang B, Wang S, et al. Fine coordinate attention for surface defect detection. Eng Appl Artif Intell, 2023, 123: 106368

19 Cao H, Chen G, Li Z, et al. Efficient grasp detection network with Gaussian-based grasp representation for robotic manipulation. IEEE ASME Trans Mechatron, 2023, 28: 1384–1394
20 Karaoguz H, Jensfelt P. Object detection approach for robot grasp detection. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2019. 4953–4959
21 Tian Y, Fu H S, Wang H, et al. RGB oralscan video-based orthodontic treatment monitoring. Sci China Inf Sci, 2024, 67: 112107
22 Wang S, Zhang Y, Chen B L, et al. Invisibility cloaks from forward design to inverse design. Sci China Inf Sci, 2013, 56: 120408
23 Liu Y, Chen X B, Mei Y F, et al. Observer-based boundary control for an asymmetric output-constrained flexible robotic manipulator. Sci China Inf Sci, 2022, 65: 139203
24 Mei S H, Geng Y H, Hou J H, et al. Learning hyperspectral images from RGB images via a coarse-to-fine CNN. Sci China Inf Sci, 2022, 65: 152102
25 Haidegger T. Autonomy for surgical robots: concepts and paradigms. IEEE Trans Med Robot BIon, 2019, 1: 65–76
26 Liu D, Tao X, Yuan L, et al. Robotic objects detection and grasping in clutter based on cascaded deep convolutional neural network. IEEE Trans Instrum Meas, 2022, 71: 1–10
27 Zhou X, Lan X, Zhang H, et al. Fully convolutional grasp detection network with oriented anchor box. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018. 7223–7230
28 Yu S, Zhai D H, Xia Y, et al. SE-ResUNet: a novel robotic grasp detection method. IEEE Robot Autom Lett, 2022, 7: 5238–5245
29 Wu Y, Zhang F, Fu Y. Real-time robotic multigrasp detection using anchor-free fully convolutional grasp detector. IEEE Trans Ind Electron, 2021, 69: 13171–13181
30 Li W, Fan L, Wang Z, et al. Tackling mode collapse in multi-generator GANs with orthogonal vectors. Pattern Recognition, 2021, 110: 107646
31 Kumra S, Joshi S, Sahin F. Antipodal robotic grasping using generative residual convolutional neural network. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020. 9626–9633
32 Hsieh Y Z, Xu F X, Lin S S. Deep convolutional generative adversarial network for inverse kinematics of self-assembly robotic arm based on the depth sensor. IEEE Sens J, 2022, 23: 758–765
33 Wang X C, Zhao H, Ma K M, et al. Kinematics analysis of a novel all-attitude flight simulator. Sci China Inf Sci, 2010, 53: 236–247
34 Morgan A S, Bircher W G, Dollar A M. Towards generalized manipulation learning through grasp mechanics-based features and self-supervision. IEEE Trans Robot, 2021, 37: 1553–1569
35 Hu W, Wang C, Liu F, et al. A grasps-generation-and-selection convolutional neural network for a digital twin of intelligent robotic grasping. Robot Comput-Integrated Manuf, 2022, 77: 102371
36 Berscheid L, Rühr T, Kröger T. Improving data efficiency of self-supervised learning for robotic grasping. In: Proceedings of International Conference on Robotics and Automation (ICRA), 2019. 2125–2131
37 Li Y S, Chen W, Huang X, et al. MFVNet: a deep adaptive fusion network with multiple field-of-views for remote sensing image semantic segmentation. Sci China Inf Sci, 2023, 66: 140305
38 Yako C L, Yuan S, Salisbury J K. Designing underactuated graspers with dynamically variable geometry using potential energy map based analysis. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022. 4638–4645
39 Raptopoulos F, Koskinopoulou M, Maniadakis M. Robotic pick-and-toss facilitates urban waste sorting. In: Proceedings of IEEE 16th International Conference on Automation Science and Engineering (CASE), 2020. 1149–1154
40 Peng G, Ren Z, Wang H, et al. A self-supervised learning-based 6-DOF grasp planning method for manipulator. IEEE Trans Automat Sci Eng, 2021, 19: 3639–3648
41 Enebuse I, Foo M, Ibrahim B S K K, et al. A comparative review of hand-eye calibration techniques for vision guided robots. IEEE Access, 2021, 9: 113143–113155
42 Kim H, Hong H, Lee B. Improved iterative closest point algorithm using truncated signed distance function. In: Proceedings of the 18th International Conference on Control, Automation and Systems (ICCAS), 2018. 1620–1623
43 Tsintotas K A, Bampis L, Gasteratos A. DOSeqSLAM: dynamic on-line sequence based loop closure detection algorithm for SLAM. In: Proceedings of IEEE International Conference on Imaging Systems and Techniques (IST), 2018. 1–6
44 Pourkamali-Anaraki F, Nasrin T, Jensen R E, et al. Evaluation of classification models in limited data scenarios with application to additive manufacturing. Eng Appl Artif Intelligence, 2023, 126: 106983
45 Wang S, Zhou Z, Kan Z. When transformer meets robotic grasping: exploits context for efficient grasp detection. IEEE Robot Autom Lett, 2022, 7: 8170–8177
46 Martínez O, Campa R. Comparing methods using homogeneous transformation matrices for kinematics modeling of robot manipulators. In: Proceedings of Multibody Mechatronic Systems, 2021. 110–118
47 Jiang Y, Moseson S, Saxena A. Efficient grasping from RGBD images: learning using a new rectangle representation. In: Proceedings of IEEE International Conference on Robotics and Automation, 2011. 3304–3311
48 Depierre A, Dellandréa E, Chen L. Jacquard: a large scale dataset for robotic grasp detection. In: Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018. 3511–3516
49 Redmon J, Angelova A. Real-time grasp detection using convolutional neural networks. In: Proceedings of IEEE International Conference on Robotics and Automation (ICRA), 2015. 1316–1322
50 Qin X, Hu W, Xiao C, et al. Attention-based efficient robot grasp detection network. Front Inform Technol Electron Eng, 2023, 24: 1430–1444
51 Breyer M, Chung J J, Ott L, et al. Volumetric grasping network: real-time 6 DOF grasp detection in clutter. In: Proceedings of Conference on Robot Learning, 2021. 1602–1611
52 Wang D, Liu C, Chang F, et al. High-performance pixel-level grasp detection based on adaptive grasping and grasp-aware network. IEEE Trans Ind Electron, 2021, 69: 11611–11621
53 Yu S, Zhai D H, Xia Y. EGNet: efficient robotic grasp detection network. IEEE Trans Ind Electron, 2022, 70: 4058–4067
54 Lenz I, Lee H, Saxena A. Deep learning for detecting robotic grasps. Int J Robotics Res, 2015, 34: 705–724
55 Morrison D, Corke P, Leitner J. Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach. 2018. ArXiv:1804.05172