Article

Bilateral Cross-Modal Fusion Network for Robot Grasp Detection

Qiang Zhang[1],*, Xueying Sun[1],*, Mingmin Liu[2] and Zhiwei Fan[3,4,5]

1   School of Automation, Jiangsu University of Science and Technology, No. 666 Changhui Road, Zhenjiang 212100, Jiangsu, P.R. China
2   Central Research Institute, SIASUN Robot & Automation Co., LTD., NO.16 Jinhui Street, Hunnan District, Shenyang 110168, P.R. China
3   State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China;
4   Institutes for Robotics and Intelligent Manufacturing, Chinese Academy of Sciences, Shenyang 110169, China
5   University of Chinese Academy of Sciences, Beijing 100049, China
*   Correspondence: qzhang@ just.edu.cn; sunxueying@just.edu.cn

Abstract: In the field of vision-based robot grasping, effectively leveraging RGB and depth information to accurately determine the position and pose of a target is a critical issue. To address this challenge, we propose a tri-stream cross-modal fusion architecture for 2-DoF visual grasp detection. This architecture facilitates the interaction of RGB and depth bilateral information and is designed to efficiently aggregate multiscale information. Our novel modal interaction module (MIM) with spatial-wise cross-attention algorithm adaptively captures cross-modal feature information. Meanwhile, the channel interaction modules (CIM) further enhance the aggregation of different modal streams. In addition, we efficiently aggregate global multiscale information through a hierarchical structure with skipping connections. To evaluate the performance of our proposed method, we conduct validation experiments on standard public datasets and real robot grasping experiments. We achieve the image-wise detection accuracy of 99.4% and 96.7% on Cornell and Jacquard datasets respectively. The object-wise detection accuracy reaches 97.8% and 94.6% on the same datasets. Furthermore, physical experiments using the 6-DoF Elite robot demonstrate a success rate of 94.5%. These experiments highlight the superior accuracy of our proposed method.

Keywords: robot grasp detection; cross-modality fusion; channel interaction

1. Introduction

In the realm of robotics, the advancement of intelligence has significantly boosted the adoption of robots. As a result, visual detection of targets has become an increasingly crucial area of focus in robotics research. A robot's ability to grasp and transport objects, either independently or in response to user commands, can enhance its ability to assimilate into the environment and broaden the range of potential robotic applications. Presently, the utilization of RGB-D cameras is making remarkable strides in robot grasping, thanks to the evolution of vision sensor technology.

Our work focuses on RGB-D data-driven robot grasp detection. Many pioneers in the field have achieved remarkable results. In the past decade, convolutional neural networks (CNNs) [1–5] have become the most widely utilized solution for robot grasp detection due to their superiority in feature representation, resulting in outstanding detection accuracy and high efficiency. While CNNs excel at local feature representation, they tend to lose information with global relevance. Recently, transformer-based approaches have gained significant popularity for visual tasks and have demonstrated comparable or superior performance in classification, semantic segmentation, and object detection. Some researchers, such as S. Wang et al. [6], have demonstrated the applicability of transformers in robot grasp detection.

Despite the impressive strides made by deep learning in solving the problem of visually detecting and grasping targets, the robustness of grasp detection still requires further improvement. This is because, while either RGB or depth images can provide some information about the scene, they are only partial in nature and may not always be effective in obtaining reliable detection results across different scenarios. Therefore, it is essential to leverage the information provided by both modalities to enhance grasp detection.

To address this issue, researchers in the field have developed early [3,7–9] and late [2,10] multimodal fusion approaches for grasp detection. While these methods have yielded meaningful results, the correlation between multimodal data has only been partially exploited. Recent studies have focused on exploring the mechanisms of intermediate fusion [11,12]. Although these methods have improved the efficiency of RGB and depth modalities in robot grasp detection, making the most of the bilateral modal information still remains a challenge.

To solve the problem of multimodal fusion, we propose a tri-stream cross-modal fusion architecture to achieve bilateral information interaction. The key idea is to use the proposed MIM approach to capture the global association information between modalities. Subsequently, the aggregation of different modal streams is refined through adaptive CIM units. The main contributions of our work can be summarized as follows:

- We propose a tri-stream cross-modal fusion architecture facilitating the interaction of RGB and depth bilateral information and efficiently aggregating multiscale information.
- A novel spatial-wise cross-attention algorithm is developed to adaptively capture cross-modal feature information. The channel interaction modules further enhance the aggregation of different modal streams.
- The proposed method demonstrates state-of-the-art grasp detection accuracy on both the Cornell and Jacquard datasets, with image-wise detection accuracy reaching 99.4% and 96.7% on Cornell and Jacquard, respectively, and object-wise detection accuracy reaching 97.8% and 94.6% on the same datasets.
- The proposed method has also shown success in guiding gripping tasks in the real world, achieving a 94.5% success rate on household items.

The remaining parts of the article are structured as follows. Section 2 presents the deep regression model for detecting robot grasps. Section 3 describes the formulation of grasp detection. The proposed method is elaborated in detail in Section 4. The performance evaluation of the proposed method is presented in Section 5. Finally, Section 6 provides a summary and conclusion of the article.

2. Related works

2.1. Grasp model representation

The representation of the robot grasp model is a prerequisite for identifying the gripping position. In vision-based approaches, the object grasp can be divided into 2-DoF planar grasp and 6-DoF grasp based on various application scenarios. For instance, 2-DoF planar grasp implies that the target object is positioned on a flat working surface, and is confined from one direction. Thus, the grasping information is reduced from 6D to 3D, specifically 2D in-plane position and 1D rotation angle. On the other hand, 6-DoF grasping enables the gripper to hold objects from different angles in the 3D space.

In 2006, A. Saxena et al. [13] proposed a point-based model representation for 6-DoF grasp detection. This representation considers the target location to be a point in 3D space. The point is detected in the image, and the relative position of the point with respect to the robot end effector is estimated using either a binocular camera or motion recovery structure, enabling the robot to perform the grasp operation. In 2010, Q. V. Le et al. [14] proposed a multi-points linkage approach to express the grasp position. Subsequent studies, such as those outlined in references [14–17], have achieved significant progress in terms of detection accuracy, reliability, and efficiency.

For 2-DoF planar grasp detection, Y. Jiang [18] proposed a rectangular representation method for the robot grasp that bypasses the object detection and pose estimation process. In this method, each grasp is represented by a rectangle with its central coordinates, width, height, and rotation angle. This simplifies the model's complexity significantly. Since then, many researchers, such as [2–5,9,14,19,20], have focused on robot grasp detection using the rectangular model, and these studies have made efforts to improve the robustness of feature representation and the real-time performance of detection.

2.2. 2-DoF planar grasp detection approaches based on rectangular representation

When it comes to studies on robot grasp detection, many researchers have focused on improving the quality of deep neural networks in order to achieve better detection results. One such method was proposed by I. Lenz et al. [1] in 2014, which utilized a multilayer deep self-encoder for image feature extraction in combination with a support vector machine classifier. J. Wei et al. [21] proposed a similar approach in 2017 using Deep Extreme Learning Machine for automatic encoding. Trottier et al. [22] also proposed a detection method in the same year using a self-coding dictionary learning method with a support vector machine classifier, although it was found to be slow and not well-suited for robotic object grasping. Z. Wang et al. [23] proposed a unified model for object segmentation and grasp detection in 2016. The method combined a grasping detection network with a two-stage estimator to improves detection accuracy.

Due to the complexity of multi-stage detection methods, more researchers are focusing on the end-to-end approach. In 2015, J. Redmon et al. [24] proposed a robot grasp detection method based on multilayer convolutional neural networks, which allowed for end-to-end training and reduced manual involvement in the training process. This approach also significantly improved detection efficiency through direct regression.

In 2018, heatmap regression methods were first utilized by D. Morrison et al. [4] to indirectly obtain grasp detection results. In their follow-up study, D. Morrison et al. [5] introduced a generative convolutional neural network for robot grasp detection. In 2020, S. Kumra et al. [8] proposed an antipodal robotic grasp detection method using a residual convolutional neural network, achieving image-wise detection accuracy of 97.7% and 94.6% on Cornell and Jacquard datasets. Their work was further improved upon in 2022 by the same group [20]. H. Cao et al. [3] proposed an efficient convolutional neural network using Gaussian-based grasp representation in 2021, which achieved image-wise detection accuracy of 97.8% and 95.6% for Cornell and Jacquard datasets, respectively. Lastly, S. Ainetter and F. Fraundorfer [25] proposed an end-to-end method for robot grasp detection in 2021, using a semantic segmentation refinement engine to increase detection accuracy.

A recent development in the field of robot grasp detection is the transformer-based method proposed by S. Wang et al. [6]. In their study, they made a preliminary attempt to address the 2-DoF grasp detection problem using the transformer architecture, and achieved impressive detection accuracy and efficiency, proving to be a competitive method in the field.

Our work explores the effectiveness of hybrid models that integrate convolutional neural networks and transformer architectures to detect 2-DoF robot grasping. This approach offers new insights into the design of effective grasp detection systems.

2.3. Multiple modality fusion based grasp detection

With the wide adoption of RGB-D sensors, an increasing number of studies have turned their attention to the efficient fusion of multimodal data. Various approaches for multimodality fusion have been proposed, including early-fusion, late-fusion, and intermediate-fusion techniques.

In 2018, F. Chu et al. [7] introduced an early-fusion approach that integrated R, G, and depth channels to predict multi-grasps for multiple objects. Two years later, in 2020, S. Kumra et al. [8] presented a generative residual convolutional neural network for grasp detection, utilizing an early-fusion strategy with both RGB and depth images. Similarly,

in 2022, H. Cao et al. [3] proposed a Gaussian-based grasp representation method using a generative grasping detection model that incorporates both RGB and depth images as inputs. Also in 2022, S. Yu et al. [9] introduced another approach using a residual neural network and squeeze-and-excitation modules.

In 2017, Q. Zhang et al. [2] put forth a sturdy robot grasp detection method that integrates RGB and depth features in the prediction head based on the YOLO architecture [26]. In 2022, Y. Song et al. [10] also proposed a hierarchical late-fusion method for RGB-D data, utilizing two CNN branches in the form of U-Net [27]. The decoding process hierarchically merges the RGB and depth features.

H. Tian et al. [11] introduced an intermediate-fusion method for lightweight pixel-wise robot grasp detection, utilizing RGB and depth information. In 2023, H. Tian et al. [12] extended their work by introducing a rotation adaptive grasp detection approach, which also utilizes intermediate data fusion. They achieved a remarkable state-of-the-art accuracy of 99.3% and 94.6% on the Cornell and Jacquard datasets, respectively.

Research on multimodal fusion has yielded promising results. However, the effectiveness of cross-modal fusion is still limited by the quality of bilateral mutual information support. To address this issue, we propose a novel solution in this paper. Our approach offers an improved framework for cross-modal fusion that enhances the mutual information support between modalities and enables more effective integration of multimodal data. Our results demonstrate the viability and superiority of our proposed method in achieving better performance in robot grasp detection.

3. Problem formulation

The robot is capable of using different types of grippers, including two-finger, three-finger, or multi-finger grippers, to grasp objects. However, parallel two-finger grippers are commonly preferred due to their simple design and cost-effectiveness. For 2-DoF grasp applications, a grasp can be represented by a 5-dimensional tuple $g = \{x, y, \theta, w, h\}$ [1,8,24]. The tuple g describes a rectangle with the center coordinates (x, y), the gripper height size (h), the gripper opening distance (w), and the orientation of the grasp rectangle (θ) with respect to the horizontal axis. Typically, the gripper dimensions are known, which allows the grasp representation to be simplified to $g = (x, y, \theta, w)$.

Instead of the 5-dimensional representation, D. Morrison et al. [5] provide an improved version of a grasp described as follows:

$$\tilde{G} = (\tilde{Q}, \tilde{\emptyset}, \tilde{W}) \in \mathbb{R}^{3 \times H \times W} \tag{1}$$

In Equation (1), $\tilde{Q}_{i,j} \in [0, 1]$ denotes the detection quality of each pixel in the image, while $\tilde{\emptyset}_{i,j} \in [-\pi/2, \pi/2]$ represents the rotation angle of the gripper, and $\tilde{W}_{i,j} \in [0, W_{max}]$ specifies the required width of the gripper's opening. Our work involves the transformation of the grasp detection problem into a pixel-level prediction. Specifically, we propose a cross-modal fusion method to derive $\tilde{G}$ from an RGB-D image of the environment in which the grasping targets are located.

Equation (1) provides a comprehensive representation of the grasp image, but the rotation angle of the grip is challenging to determine due to its symmetrical values. To address this ambiguity, we encode the rotation angle using $\sin(2\tilde{\emptyset}_{i,j})$ and $\cos(2\tilde{\emptyset}_{i,j})$, which helps eliminate any discontinuities that may arise during the calculation. The angle of the grasp to be executed at each pixel can be obtained using equation (2):

$$\tilde{\emptyset}_{i,j} = \frac{1}{2} \times \arctan \left[\frac{\sin(2\tilde{\emptyset}_{i,j})}{\cos(2\tilde{\emptyset}_{i,j})} \right] \tag{2}$$

The optimal grasp within an image space is determined by identifying the pixel with the highest quality score in $\tilde{G}$. Additionally, the grasp can be straightforwardly mapped to physical space based on the internal and external parameters of the RGB-D camera.

4. Approach

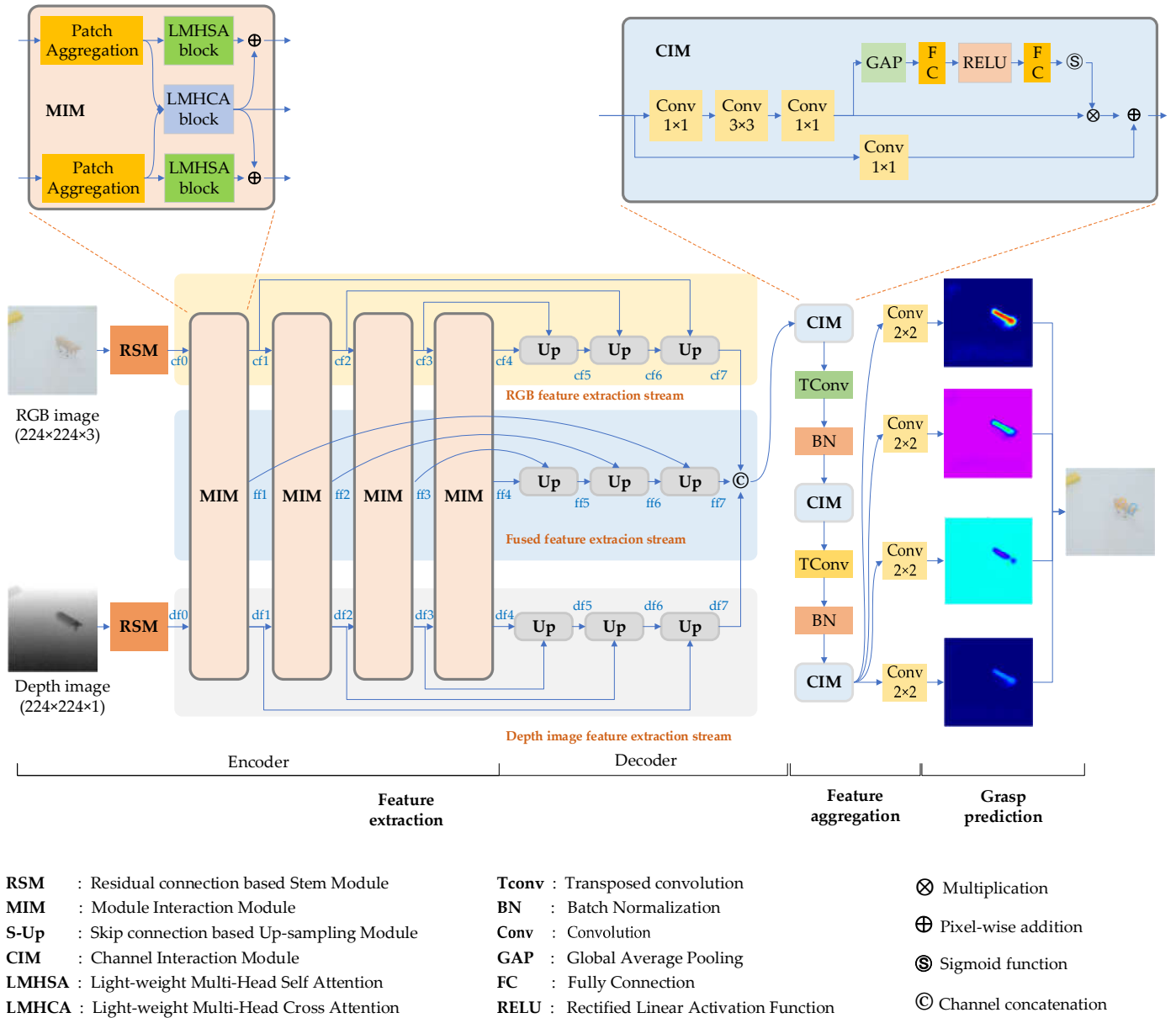*4.1. Overview of Bilateral Cross-Modal Fusion Network*

Figure 1. Bilateral cross-modal fusion network architecture.

Our robot grasp detection architecture is illustrated in Figure 1. It comprises three main components: feature extraction, feature aggregation, and grasp prediction. To ensure robust feature extraction, we employed two strategies. Firstly, we tackled the problem of modality interaction in feature fusion by assigning adaptive weights to RGB and depth image features during the fused feature extraction stage. Secondly, we adopted a channel interaction approach for feature aggregation.

The feature extraction method serves two purposes: extracting multi-scale features from the RGB and depth scene images, and constructing fused features from the two modalities. This process is accomplished through three streams: the RGB feature extraction stream, the depth feature extraction stream, and the fused feature extraction stream. Within each stream, the feature extraction process comprises two key stages: feature encoding and decoding.

The architecture of the RGB and depth feature extraction process, as shown in Figure 1, is based on U-Net-like structures. Similar to the approach in [28], each stream consists

of a residual connection based stem module (RSM) and a series of modal interaction modules (MIM) for encoding RGB and depth features. Down-sampling operations are used to obtain multi-scale features. However, low-level features may contain more details but also unnecessary information, while high-level features may have more semantic information but may not represent small target features well. To obtain more robust features, we employ feature decoders, which are composed of transposed convolution and up-sampling processes with skip connections.

As depicted in Figure 1, the initial step in the feature extraction process involves the utilization of the RSM on the RGB image, which is responsible for extracting fine-grained features and generating the cf0 feature. To produce features at multiple scales, the cf0 feature undergoes several stages of light-weight multi-head self-attention modules (LMHSA) to create multi-scale features, including cf1, cf2, cf3, and cf4. These encoding procedures also involve the fused features from using an addition operation. To ensure that the resulting feature map is consistent with the size of the input image, a series of up-sampling modules are employed at the decoding step. Additionally, concatenation operations are incorporated into these processes to fully exploit the low-level and high-level features, leading to the production of the cf5, cf6, and cf7 feature maps. The depth feature extraction stream is also capable of generating corresponding feature maps, ranging from df0 to df7.

Compared to the encoding processes for RGB and depth features, the fused feature maps (ff1, ff2, ff3, and ff4) are initially generated using light weight multi-head cross-attention (LMHCA) strategy. During feature fusion, adaptive weights are assigned to the RGB and depth information. Details about the LMHCA algorithm is described in 4.2.2. To ensure the resulting fused features are robust, the same decoders used in the RGB and depth feature decoding processes are employed. Consequently, ff7 is of the same size as cf7 and df7.

As previously stated, the feature extraction process results in three distinct feature maps: cf7, df7, and ff7. It is our hypothesis that the feature aggregation process should optimally utilize useful information from all three features while minimizing the impact of irrelevant information. Previous research, such as that conducted in [2,11], has made numerous attempts to explore this topic. However, both studies employ equal-weight feature aggregation for each channel. In an effort to enhance the efficacy of feature aggregation, we have implemented a channel interaction strategy. As illustrated in Figure 1, the RGB, depth, and fused feature maps are initially concatenated. CIM units are then utilized to assign adaptive weights to the feature channels. Subsequent transposed convolution based up-sampling processes further improve the resolution of the fused feature map.

The grasp prediction head is comprised of several convolution calculation modules and is able to predict grasp quality, $\cos(2\widetilde{\phi}_{i,j})$, $\sin(2\widetilde{\phi}_{i,j})$, and grasp opening width heatmaps, which are employed to construct grasp rectangles.

4.2. Feature extraction pipeline

4.2.1. Residual connection based stem module (RSM)

The residual network has proven to be highly effective in various applications such as image classification, object detection, and moving object tracking. To capitalize on its exceptional performance, we adopt a RSM module to generate a compact feature representation, thereby addressing the inferior feature representation capability of the linear projection.

Taking inspiration from [29], the stem module comprises two streams, as depicted in Figure 2. The first stream is composed of a sequence of convolution, Gaussian Error Linear Unit (GELU) [30], and Batch Normalization (BN) [31]processes. The convolution modules utilize 1×1 (stride=1), 3×3 (stride=2), and 1×1 (stride=1) kernels, respectively. This approach ensures that no information is neglected. The second stream includes a 2×2 (stride=1) average pooling stage, a 1×1 (stride=1) kernel convolution, GELU, and BN modules. This mechanism enables the expression of features while incurring minimal computational

costs. The outputs of these two streams are combined to generate the encoding features, which are then fed into the self-attention and cross-attention procedures.
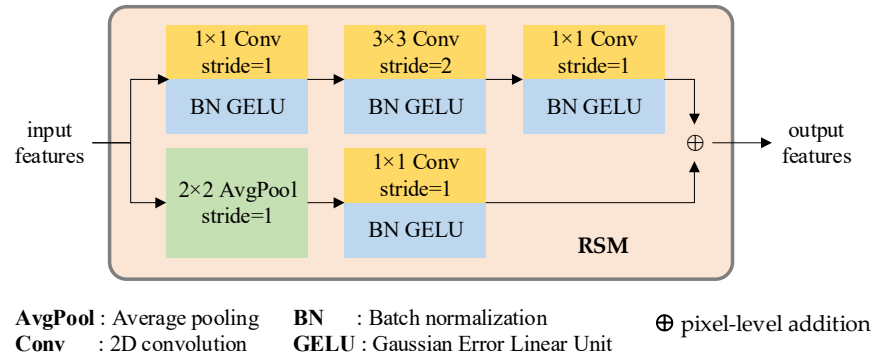

AvgPool : Average pooling   **BN**   : Batch normalization   ⊕ pixel-level addition
Conv   : 2D convolution   **GELU** : Gaussian Error Linear Unit

Figure 2. Input data feature extraction diagram using RSM.

4.2.2. Cross-modal feature encoding based on MIM

The proposed MIM module is use to execute feature extraction and bilateral RGB and depth cross-modal fusion strategies. As shown in Figure 3, Each MIM module consists of two patch embedding blocks, two LMHSA blocks, one LMHCA module and two summation units. The patch aggregation module, composed by a 2×2 convolution with the stride=2 and layer normalization block, is used to aggregate patches into a single image and produce hierarchical representation. The LMHSA module is used to extract features by spatial-wise self-attention. The LMHCA module is to compute fused features based on cross-attention. The summation operation helps to achieve mutual information support.
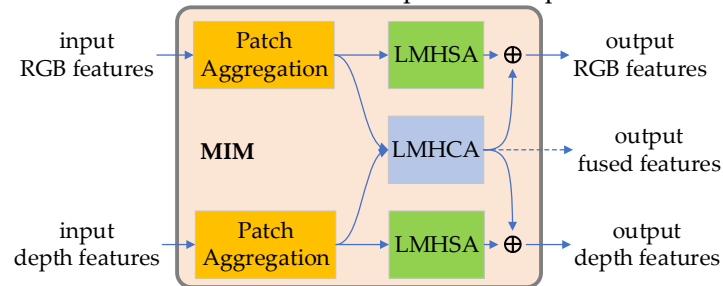

Figure 3. The structure of the MIM module

* Light weight multi-head self-attention (LMHSA) block

As shown in Figure 4, before performing self-attention calculations, we incorporated a local perceptual unit to address the issue of missing local associations and structural information when using the self-attention module [28]. This unit is depicted in light yellow in Figure 4. Next, we apply LMHSA computation to the feature transformation, represented in light green in Figure 4. To enhance the representation ability of tokens, we introduce the Inverted Residual Feed-Forward Network module, or IRFFN, which can perform dimensional expansion and non-linear transformation on each token.
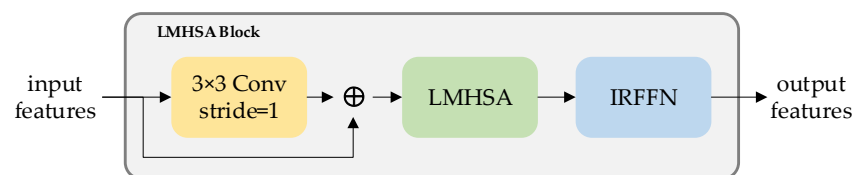

Figure 4. LMHSA block for RGB and Depth features extraction.

As described in [32], the original self-attention module utilizes linear mapping to derive the query matrix Q, key matrix K, and value matrix V. The dimensions of Q, K and V are given by $H \times W \times d_k$, $H \times W \times d_k$ and $H \times W \times d_v$, respectively, where $H \times W$

represents the number of image patches, and d_k and d_v are the dimensions of the tensor K and V. Subsequently, the self-attention module can be expressed as the following formula:

$$Attn(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3}$$

While the original self-attention algorithm can effectively handle various visual tasks, it is associated with high computational costs. As such, numerous researchers have dedicated efforts to addressing this issue. In our work, we adopt a similar approach to that in [32] which involves the use of $k \times k$ depth-wise convolutional operations with stride of s to reduce the dimensionality of the key and value matrices, thereby mitigating the computational burden. The computation associated with the depth-wise convolution can be expressed as follows:

$$\begin{cases} Q = Linear(X) \\ K = Linear(DWConv(X)) \\ V = Linear(DWConv(X)) \end{cases} \tag{4}$$

Equation (4) describes the lightweight self-attention mechanism used in our approach, where X represents the input feature, $DWConv(\cdot)$ denotes the depth-wise convolution operation, and $Linear(\cdot)$ is the linear operation. To further enhance performance, we incorporate a position bias term. Then the lightweight self-attention can be defined as:

$$LightAttn(Q,K,V,B) = Softmax(\frac{QK^T}{\sqrt{d_i}} + B)V \tag{5}$$

In the above formula, the dimensions of the query (K) and value (V) matrices are reduced to $1/s^2$ due to the application of stride s in the depth-wise convolution kernel. d_i is the dimension of Q. The structure of the lightweight multi-head self-attention module is depicted in Figure 5.
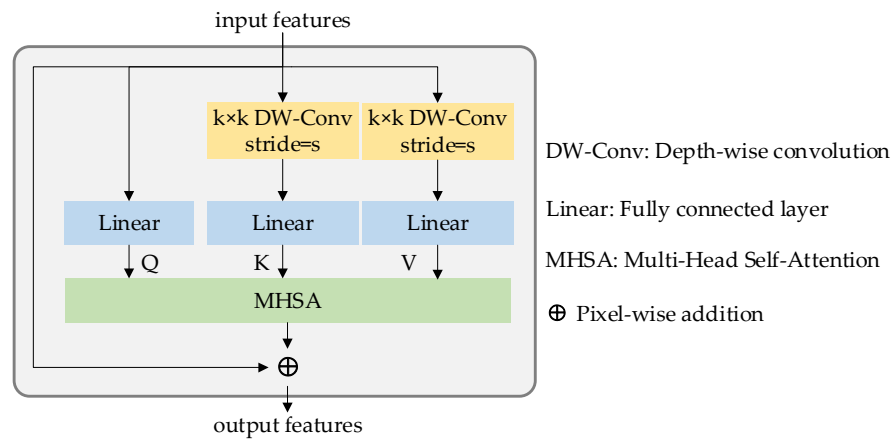

Figure 5. Lightweight multi-head self-attention block.

In the robot grasp detection architecture, the IRFFN layer serves to expand and reduce feature dimensions, allowing for non-linear transformation. However, unlike the structure proposed in [28], we utilize an improved IRFFN layer to boost the expressiveness of the features. Figure 6 displays the structure of our proposed IRFFN layer which consists of two branches.
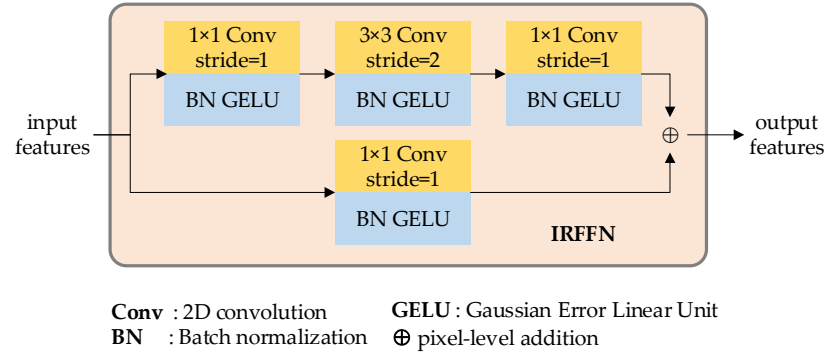
Conv : 2D convolution   **GELU** : Gaussian Error Linear Unit
BN   : Batch normalization   ⊕ pixel-level addition

Figure 6. IRFFN block diagram.

- Light weight multi-head cross-attention (LMHCA) module

The fusion algorithm we have developed tackles the challenge of how to optimize the utilization of information from RGB and depth modalities, taking into account their respective importance. To further enhance the robustness of the fused features, we have designed a cross-attention mechanism and a modal reweighting strategy. These techniques work in tandem to ensure that the most salient features from each modality are given the appropriate attention and weight in the final fusion result.

The proposed modal interaction strategy is illustrated in Figure 7. To address the issue of missing local associations and structural information during cross-attention computation, we incorporate local perception units into the design. Subsequently, multi-head cross-attention operations are employed to extract high-level semantic features from the RGB and depth features. Finally, the output features are fed into the IRFFN unit to enable dimensional expansion and non-linear transformation.
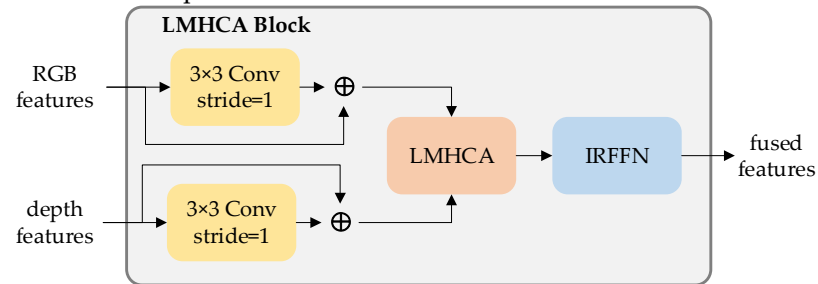

Figure 7. LMHCA block for fused feature extraction.

The token compounding method proposed by [33] has demonstrated outstanding performance in vision-and-language representation learning. Building upon this approach, we propose a lightweight multi-head cross-attention method to facilitate the fusion of RGB and depth features in robotic gripping applications. The lightweight multi-head cross-attention unit comprises two parts: a multi-head cross-attention component and a modality reweighting-based feature fusion component. Figure 8 presents a detailed schematic of the proposed procedure. Specifically, in the cross-attention module, we employ linear layer to acquire matrices Q_{rgb}, Q_d for RGB and depth streams respectively. Depth-wise convolution and linear mapping are used to obtain the matrices K_{rgb}, V_{rgb}, K_d, V_d accordingly. Then, the lightweight cross-attention computation can be expressed as equation (6).

$$\begin{cases} LightCrossAtten(RGB2depth) = Softmax(\frac{Q_{rgb}K_d{}^T}{\sqrt{d_i}} + B)V_d \\ LightCrossAtten(depth2RGB) = Softmax(\frac{Q_dK_{rgb}{}^T}{\sqrt{d_i}} + B)V_{rgb} \end{cases} \quad (6)$$

RGB and depth tokens are exported according the upper operation. These two tokens are concatenated and fed into the modality reweighting process to create fused features. Different from the strategy of [34], we developed a novel modal reweighting method and assign appropriate adaptive weights to RGB and depth tokens respectively to obtain

effective fused features. The detailed structure of the modality reweighting strategy can be seen in the bottom half of Figure 8. First, we concatenate the RGB and depth tokens, followed by using 1×1 convolution and SoftMax functions to learn adaptive weights for both token types. The fused features are obtained by reweighting the RGB and depth tokens and adding them at the pixel level.
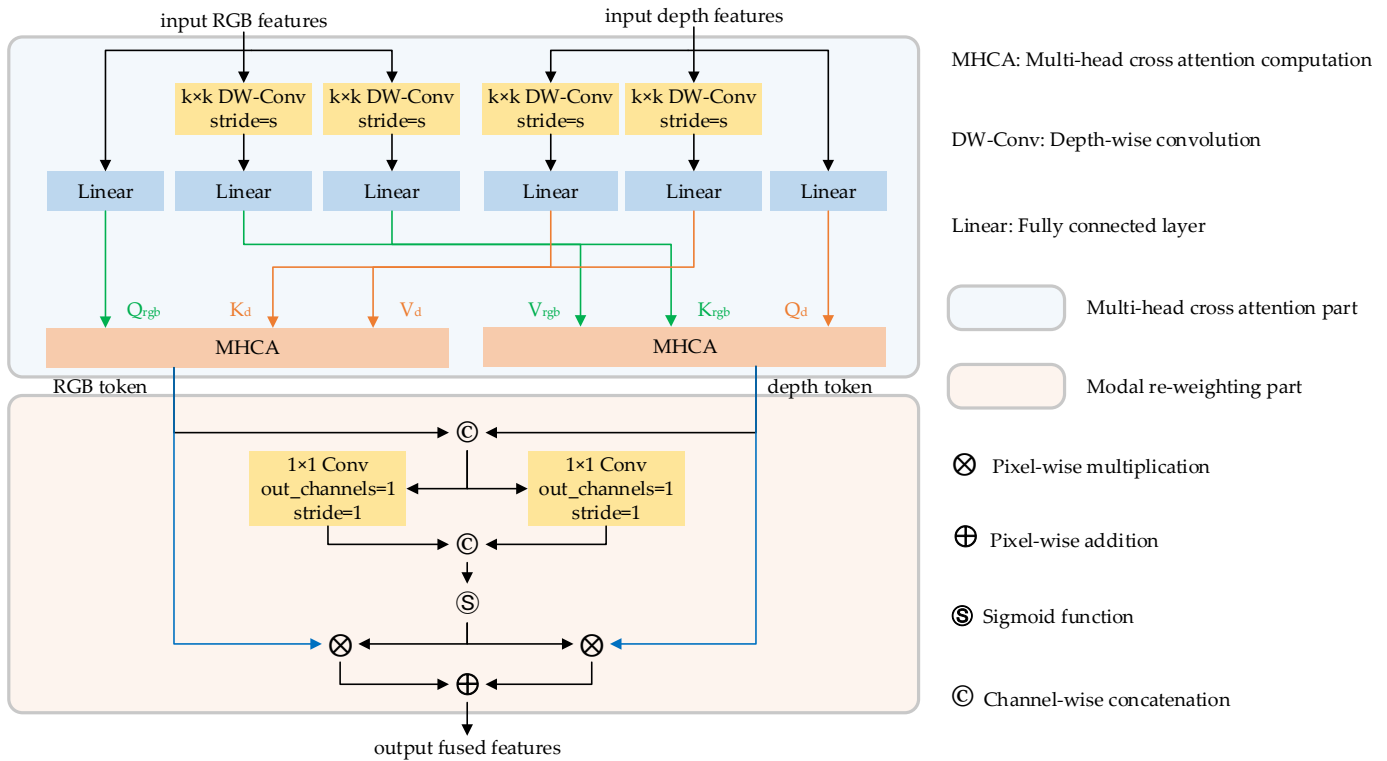

Figure 8. Multi-head cross-attention block diagram.

4.3. Feature aggregation based on channel interaction module (CIM)

In Figure 1, three feature maps cf7, df7 and ff7 can be obtained after the feature extraction procedures. These feature maps are subsequently aggregated for grasp prediction.
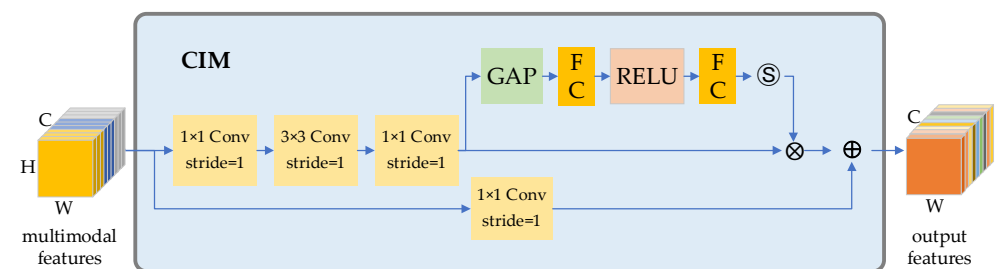

Figure 9. Channel interaction module.

Since each of these maps contains both valuable and non-valuable information, we apply the channel interaction method to reweight each channel of the connected features accordingly. To accomplish this, we use an SE-block [35] to improve the sensitivity of valuable channels and suppress useless ones. The structure of the CIM unit is shown in Figure 9.

To generate a feature map that is of the same size as the input image, we employ the same up-sampling techniques used in the feature decoding step in the final stages of the feature aggregation process.

4.4. Robot grasp prediction

As outlined in sections 3 and 4.1, the neural network we designed is expected to generate four heatmaps, namely $\tilde{Q}$, $\sin(2\tilde{\phi}_{i,j})$, $\cos(2\tilde{\phi}_{i,j})$ and $\tilde{W}$, to facilitate robot grasping. To accomplish this, four separate branches of 2-D convolutions are constructed.

4.5. Loss function

We train the neural network by minimizing the discrepancy between the predicted grasps ($\tilde{G}$) and the ground truth grasps (G). To accomplish this, we utilize the smooth L1 loss function [36] in our work. This loss function is defined as follows:

$$L(\tilde{G} - G) = \sum_i^N \sum_{k \in \{Q, \emptyset, W\}} smooth_{L1}\left(\widetilde{G_{i,j}^k} - G_{i,j}^k\right) \tag{7}$$

In Equation (7), N represents the number of pixels in the heatmap, and $smooth_{L1}(x)$ is defined as:

$$smooth_{L1}(x) = \begin{cases} 0.5 \times \frac{x^2}{\beta}, & if |x| < \beta \\ |x| - 0.5\beta, otherwise \end{cases} \tag{8}$$

Here, the hyperparameter β controls the extent of smoothness and separates the positive axis range into $L1$ loss and $L2$ loss parts. In our work, we set the parameter β to 1.

5. Evaluation

5.1. Experimental methodology

5.1.1. Experiment content

We have designed four experiments to comprehensively evaluate the proposed method. The first two experiments are comparison studies that aim to verify the performance of different approaches on the Cornell and Jacquard datasets, respectively. The third experiment is an ablation study that examines the effects of cross-attention and channel interaction strategies. In this experiment, we evaluate the effectiveness of the modality adaptive reweighting algorithm in the fused feature extraction stage and the effects of the channel interaction algorithm in the feature aggregation stage. Additionally, we have verified the effectiveness of the proposed algorithm through a fourth physical experiment.

5.1.2. Datasets

We utilized two datasets, the Cornell dataset [1] and the Jacquard dataset [37], in our experiments. The Cornell dataset is relatively small, comprising 240 distinct objects with 885 samples, while the Jacquard dataset is of medium size, consisting of 11,619 unique objects and 54,485 different scenes. Both datasets provide RGB images and 3D dense cloud data for each sample. Prior to training the neural network, we converted the 3D point cloud data into depth images and adjusted their resolution to 224×224. We allocated 90% of each dataset for training and the remaining 10% for testing. Given the small size of the Cornell dataset, we augmented the dataset by performing augmentation operations such as cropping and rotation.

5.1.3. Experiment environment

The training and validation process was conducted on the Ubuntu 20.04 operating system, utilizing an Intel Core i9-12900KF CPU clocked up to 5.20 GHz, 64GB DDR4 memory, and an NVIDIA GeForce GTX 3090-Ti graphics card.

We have set up a real-world robot grasp scenario, as depicted in Figure 10. For this experiment, we gathered 30 distinct objects on a desk. An Orbbec Femto-W RGBD camera is used as the image sensor, while a parallel gripper is installed at the end of the Elite EC-66 collaborative robot to act as the clamping mechanism. Prior to the experiment, a hand-eye calibration is performed to ensure proper operation of the system.

Figure 10. Physical experiment conditions. Experiment instruments include a Femto-W RGB-D camera, an EC-66 collaborative robot, a parallel gripper and some objects to be grasped.

5.1.4. Grasp Detection Metric

In all of our experiments, we employ the grasp intersection over union (IoU) metric, which is defined in equation (9). The use of this metric allows for a quantitative evaluation of the performance of our proposed method in terms of its ability to accurately predict grasps.

$$IoU = \frac{G_{det} \cap G_{GT}}{G_{det} \cup G_{GT}} \times 100\% \qquad (9)$$

The IoU metric is utilized in all experiments, as defined in equation (8). The numerator of the equation represents the area of overlap between the detected grasp rectangle and the ground truth, while the denominator represents their union. To be considered a valid detection, the results must exhibit the following properties:

- IoU should be above 25%
- The angle error between the detection result and the ground truth should be less than 30°.

5.1.5. Experiment configuration

Our proposed architecture has input dimensions of $224 \times 224 \times 3$ and $224 \times 224 \times 1$ for RGB and depth images, respectively. The detailed size of each feature map is listed in Table 1.

We utilized self-attention and cross-attention based feature encoders with 1, 2, 4, and 8 heads, and the corresponding number of block layers are 2, 2, 10, and 2, respectively.

Table 1. Size of each feature map.

Feature map	Size (H×W×C)	Feature map	Size (H×W×C)
cf0, df0	112×112×16	cf5, df5, ff5	14×14×184
cf1, df1, ff1	56×56×46	cf6, df6, ff6	28×28×92
cf2, df2, ff2	28×28×92	cf7, df7, ff7	56×56×46
cf3, df3, ff3	14×14×184	f8	112×112×46
cf4, df4, ff4	7×7×368	f9	224×224×32

5.2. Experiement results

5.2.1 Cornell dataset experiment results

To compare the grasp detection performance of recent methods with our proposed algorithm on the Cornell dataset, we conducted an experiment that evaluated image-wise and object-wise grasp detection separately. Our algorithm achieved state-of-the-art

accuracy of 99.4% and 97.8% in image-wise and object-wise grasp detection, respectively, as shown in Table 2. However, the average time expenditure is 17.7 ms, which is higher compared to algorithms in [10–12] due to the complexity of our algorithm.

Figure 11 depicts some typical examples of heatmap regression results for quality, angle, and width, as well as grasp detection results. As shown in the figure, the quality heatmaps demonstrate the robustness of our proposed method, which contributes to the superior performance of our grasp detection results.

Table 2. Grasp detection results of different algorithms on Cornell dataset.

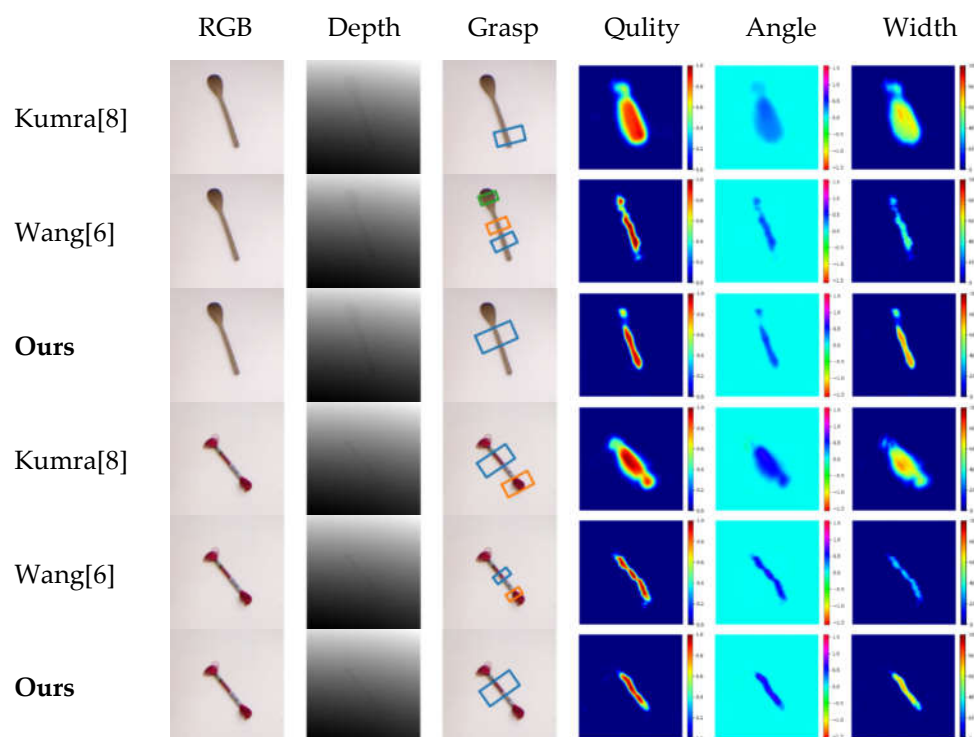| Method	Input	Accuracy (%)		Time (ms)
		Image-wise	Object-wise	
Lenz[1]	RGB-D	73.9	75.6	1350
Redmon[24]	RGB-D	88	87.1	76
Morrision[5]	D	73	69	19
Song[10]	RGB-D	92.5	90.3	17.2
Kumra[8]	RGB-D	97.7	96.6	20
Wang[6]	RGB-D	97.99	96.7	41.6
Yu[9]	RGB-D	98.2	97.1	25
Tian[11]	RGB-D	98.9	-	15
Tian[12]	RGB-D	99.3	91.1	**12**
Ours	RGB-D	**99.4**	**97.8**	**17.7**

Figure 11. Experiment results of the algorithms [8,6] and our method on Cornell dataset. The 1st and 2nd columns are RGB image and depth images. The 3rd column shows grasp detection results. The last three columns illustrate the quality, angle and width heatmaps.

5.2.2. Jacquard dataset experiment results

We also conducted a comparative analysis of our grasp detection algorithm with that of several other methods [5,6,8–12] using the Jacquard dataset. Table 3 presents the statistical results of our experiment with the Jacquard dataset. As evident from the table, our algorithm achieved the highest image-wise and object-wise detection accuracy of 96.7% and 94.6%, respectively, on the Jacquard dataset. Figure 11 shows several detection cases. Our algorithm offers superior quality heatmap prediction results.

Table 3. Grasp detection results of different algorithms on Jacquard dataset.

Method	Input	Accuracy (%)	
		Image-wise	Object-wise
Morrison[5]	D	84	-
Song[10]	RGB-D	93.2	-
Kumra[8]	RGB-D	92.6	87.7
Wang[6]	RGB-D	94.6	-
Yu[9]	RGB-D	95.7	-
Tian[11]	RGB-D	94	-
Tian[12]	RGB-D	94.6	92.8
Ours	RGB-D	**96.7**	**94.6**

The detection results presented in Table 3 and Figure 12 provide evidence that our proposed method, which leverages cross-attention and channel interaction for RGB-D feature fusion, can effectively utilize the information shared between the two modalities.
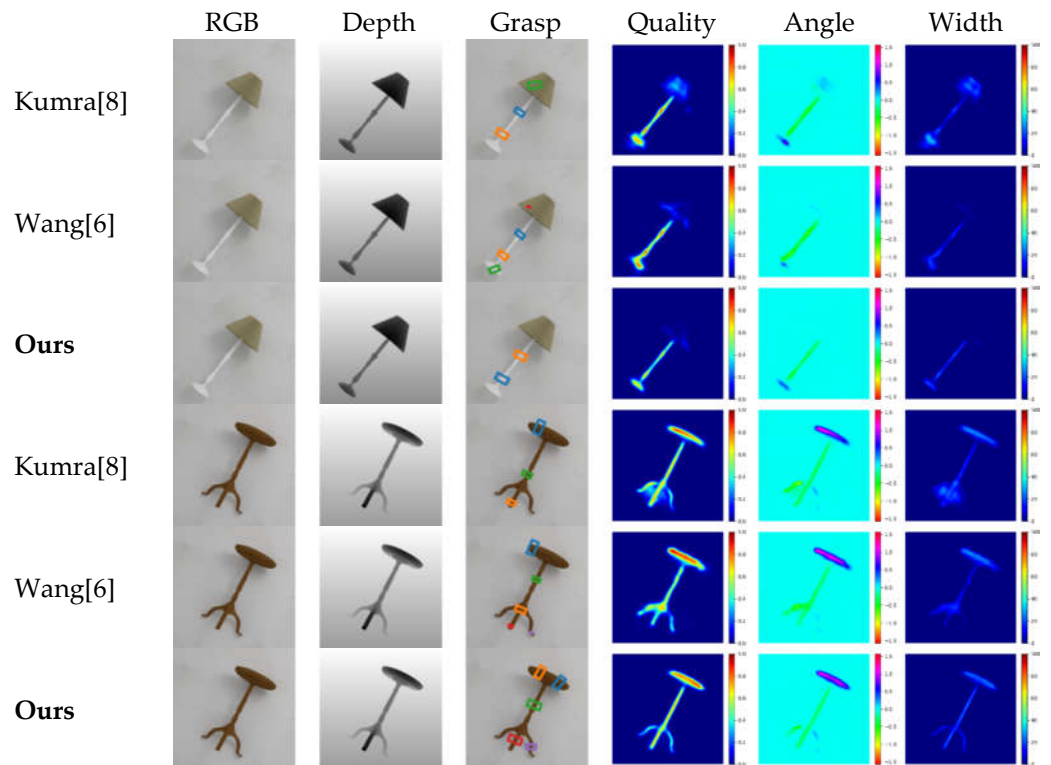

Figure 12. Experiment results of algorithms in [8,6] and our method on Jacquard dataset.

5.2.3 Ablation experiment

Since the cross-attention module is only involved in the fused feature encoding stage, we simplified our pipeline (shown in Figure 1) to produce the architecture depicted in

Figure 13.

To validate the effectiveness of the different modules in the proposed approach, we conducted several leave-one-out experiments on the Cornell and Jacquard datasets. Initially, we removed the MIM and CIM modules from the proposed architecture. The generated approaches are served as baseline approaches, shown in Figure 13(a) and Figure 13(b).
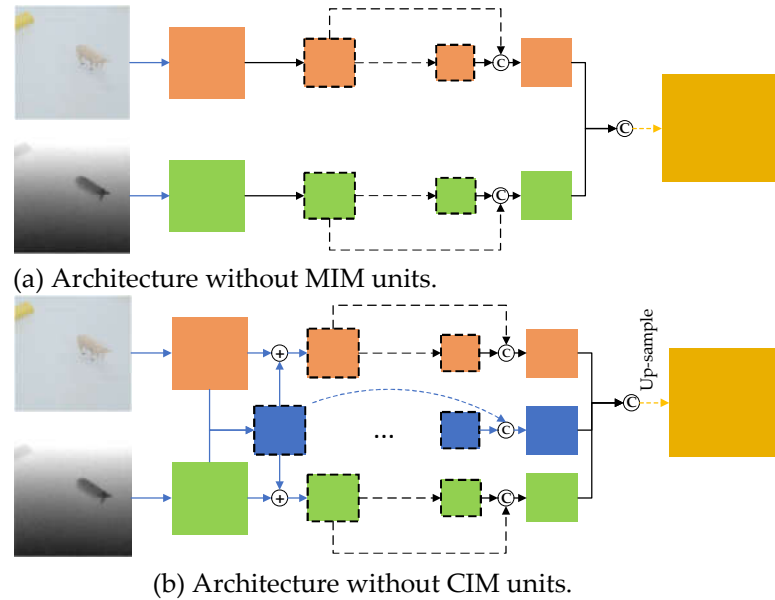

(a) Architecture without MIM units.

(b) Architecture without CIM units.

Figure 13. Architectures in ablation experiments. (a) has no MIM blocks. (b) has no CIM blocks.

In the ablation experiment, we evaluated the object-wise grasp detection accuracy on both the Cornell and Jacquard datasets. The corresponding statistical results are presented in Table 4.

Table 4. Object-wise grasp detection results of ablation experiment on Cornell and Jacquard datasets.

Methods	Accuracy of Cornell Dataset (%)	Accuracy of Jacquard Dataset (%)
Without MIM	89.7	84.6
Without CIM	96.4	92.6
With MIM and CIM	**97.8**	**94.6**

The results of the ablation experiment demonstrate that the bilateral modality interaction method based on cross-attention significantly enhances the accuracy of grasp detection. Additionally, the feature aggregation method based on channel interaction strategy has a fine-tuning effect on detection accuracy.

5.2.4. Physical experiment

The physical experiment was conducted on our in-house robotic platform, which comprises an Elite EC-66 robot with public open ROS interfaces, a parallel gripper, an Orbbec Femo-W RGB-D camera, and a computer server running Ubuntu. The experiment involved 30 different unknown objects in the scene, with the camera positioned relative to the desktop similar to that in the Cornell dataset. RGB-D image data was captured by the camera, and the server detected the position and pose of potential grasps. Following a coordinate transformation, the robot executed the grasping operation on the target object. The detailed grasp process is depicted in Figure 14.

During the experiment, we made a total of 200 attempts to grasp the target objects and successfully grasped the objects in 189 of those attempts, resulting in an average success rate of 94.5%. Figure 15 illustrates several examples of our grasp detection results.
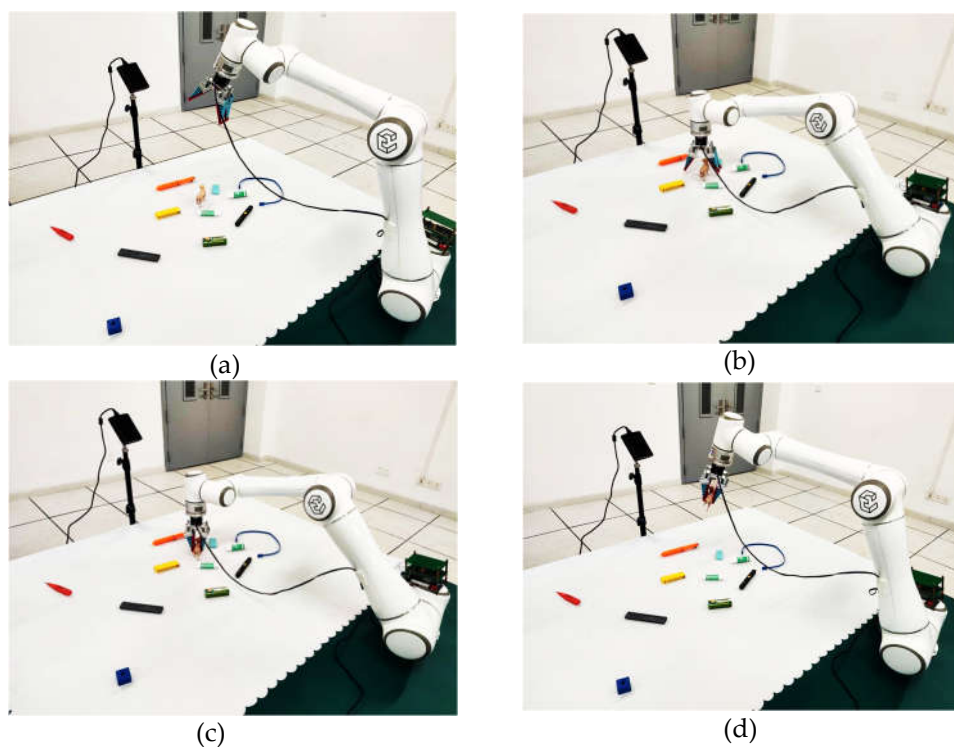

Figure 14. Four grasp stages in the physical experiment. (a) shows the initial position and posture of EC-66 robot. In this stage, grasp detection is performed. After detection, the parallel gripper moves to the effective grasping position, which can be seen in (b). The gripper then grasps the target, which is shown in (c). The robot completes the target grasping task in (d) finally.
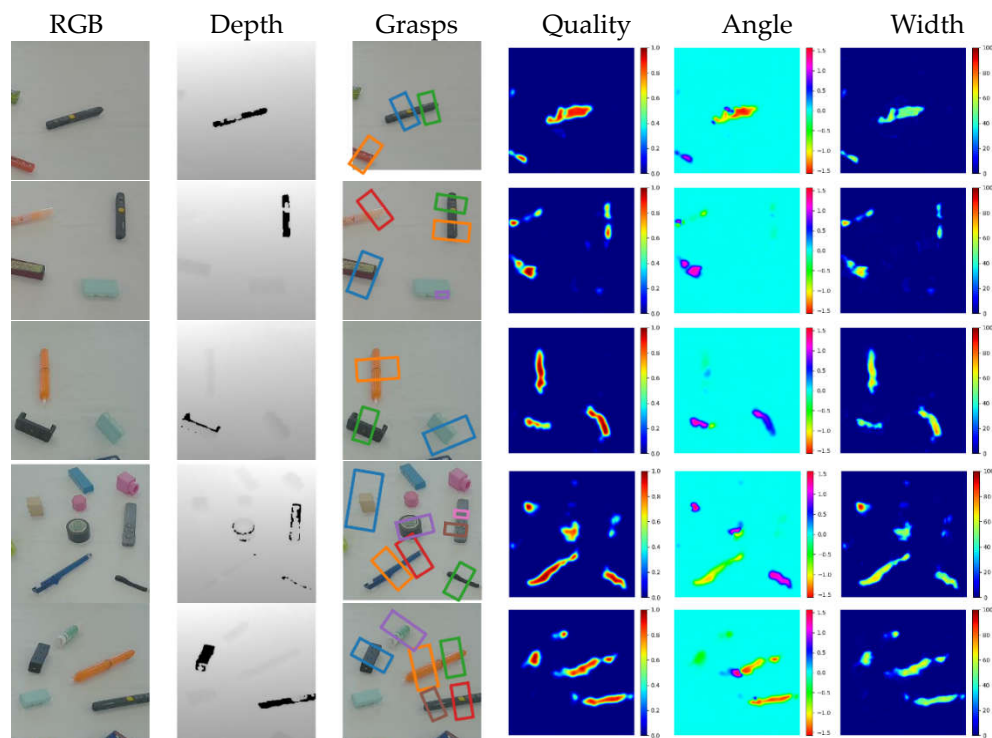

Figure 15. Detection results of the physical experiment.

6. Conclusions

This paper addresses the 2-DoF robot grasp detection problem by analyzing data fusion issues that affect grasp detection results. Our analysis shows that fully utilizing useful information from each modality and eliminating useless information is essential for achieving high accuracy in grasp detection. In response, we propose a cross-modality fusion method for 2-DoF robot grasp detection that uses a convolutional neural network and transformer structure. This method incorporates modal interaction and channel interaction strategies to adaptively retain essential information and reduce the impact of invalid information. To validate our approach, we conduct a series of comparison experiments with other methods and an ablation experiment. Our experimental results demonstrate that the proposed cross-modality fusion method achieves high accuracy for both image-wise and object-wise grasp detection and is effective in practical robot grasp detection scenarios. However, this method is somewhat time-consuming, which is a limitation that we will address in future work.

Author Contributions: Conceptualization, Q. Zhang; methodology, Q. Zhang and X. Sun; software, Q. Zhang and X. Sun; writing—original draft preparation, Q. Zhang; writing—review and editing, X. Sun and M. Liu; visualization, Z. Fan. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Cornell, Jacquard and multi-object multi-grasp RGB-D dataset experiment results are available on https://github.com/QZhSSLab/CrossModalityFusion4RobotGrasp. The code presented in this study is available on request from the corresponding author.

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

References

1. Lenz, I.; Lee, H.; Saxena, A. Deep Learning for Detecting Robotic Grasps. *Int. J. Robot. Res.* **2015**, *34*, 705–724, doi:10.15607/rss.2013.ix.012. [CrossRef]
2. Zhang, Q.; Qu, D.; Xu, F.; Zou, F. Robust Robot Grasp Detection in Multimodal Fusion. *MATEC Web Conf.* **2017**, *139*, 00060, doi:10.1051/matecconf/201713900060. [CrossRef]
3. Cao, H.; Chen, G.; Li, Z.; Feng, Q.; Lin, J.; Knoll, A. Efficient Grasp Detection Network With Gaussian-Based Grasp Representation for Robotic Manipulation. *IEEEASME Trans. Mechatron.* **2022**, 1–11, doi:10.1109/TMECH.2022.3224314. [CrossRef]
4. Morrison, D.; Corke, P.; Leitner, J. Closing the Loop for Robotic Grasping: A Real-Time, Generative Grasp Synthesis Approach. In Proceedings of the Robotics; MIT Press - Journals: United States of America, 2018. [CrossRef]
5. Morrison, D.; Corke, P.; Leitner, J. Learning Robust, Real-Time, Reactive Robotic Grasping. *Int. J. Robot. Res.* **2019**, *39*, 027836491985906, doi:10.1177/0278364919859066. [CrossRef]
6. 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, doi:10.1109/lra.2022.3187261. [CrossRef]
7. Chu, F.-J.; Xu, R.; Vela, P.A. Real-World Multiobject, Multigrasp Detection. *IEEE Robot. Autom. Lett.* **2018**, *3*, 3355–3362, doi:10.1109/LRA.2018.2852777. [CrossRef]
8. Kumra, S.; Joshi, S.; Sahin, F. Antipodal Robotic Grasping Using Generative Residual Convolutional Neural Network. In Proceedings of the 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); IEEE: Las Vegas, NV, USA, October 24 2020; pp. 9626–9633. [CrossRef]
9. Yu, S.; Zhai, D.-H.; Xia, Y.; Wu, H.; Liao, J. SE-ResUNet: A Novel Robotic Grasp Detection Method. *IEEE Robot. Autom. Lett.* **2022**, *7*, 5238–5245, doi:10.1109/LRA.2022.3145064. [CrossRef]
10. Song, Y.; Wen, J.; Liu, D.; Yu, C. Deep Robotic Grasping Prediction with Hierarchical RGB-D Fusion. *Int. J. Control Autom. Syst.* **2022**, *20*, 243–254, doi:10.1007/s12555-020-0197-z. [CrossRef]
11. Tian, H.; Song, K.; Li, S.; Ma, S.; Yan, Y. Lightweight Pixel-Wise Generative Robot Grasping Detection Based on RGB-D Dense Fusion. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12, doi:10.1109/TIM.2022.3196130. [CrossRef]
12. Tian, H.; Song, K.; Li, S.; Ma, S.; Yan, Y. Rotation Adaptive Grasping Estimation Network Oriented to Unknown Objects Based on Novel RGB-D Fusion Strategy. *Eng. Appl. Artif. Intell.* **2023**, *120*, 105842, doi:10.1016/j.engappai.2023.105842. [CrossRef]

13. Saxena, A.; Driemeyer, J.; Kearns, J.; Ng, A. Robotic Grasping of Novel Objects. In Proceedings of the Advances in Neural Information Processing Systems; MIT Press, 2006; Vol. 19. [CrossRef]

14. Le, Q.V.; Kamm, D.; Kara, A.F.; Ng, A.Y. Learning to Grasp Objects with Multiple Contact Points. In Proceedings of the 2010 IEEE International Conference on Robotics and Automation; May 2010; pp. 5062–5069. [CrossRef]

15. Liang, H.; Ma, X.; Li, S.; Gorner, M.; Tang, S.; Fang, B.; Sun, F.; Zhang, J. PointNetGPD: Detecting Grasp Configurations from Point Sets. In Proceedings of the 2019 International Conference on Robotics and Automation (ICRA); IEEE: Montreal, QC, Canada, May 2019; pp. 3629–3635. [CrossRef]

16. Gou, M.; Fang, H.-S.; Zhu, Z.; Xu, S.; Wang, C.; Lu, C. RGB Matters: Learning 7-DoF Grasp Poses on Monocular RGBD Images. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); IEEE: Xi'an, China, May 30 2021; pp. 13459–13466. [CrossRef]

17. Sundermeyer, M.; Mousavian, A.; Triebel, R.; Fox, D. Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); IEEE: Xi'an, China, May 30 2021; pp. 13438–13444. [CrossRef]

18. Jiang, Y.; Moseson, S.; Saxena, A. Efficient Grasping from RGBD Images: Learning Using a New Rectangle Representation. In Proceedings of the 2011 IEEE International Conference on Robotics and Automation; May 2011; pp. 3304–3311. [CrossRef]

19. Shi, C.; Miao, C.; Zhong, X.; Zhong, X.; Hu, H.; Liu, Q. Pixel-Reasoning-Based Robotics Fine Grasping for Novel Objects with Deep EDINet Structure. *Sensors* **2022**, *22*, 4283, doi:10.3390/s22114283. [CrossRef]

20. Kumra, S.; Joshi, S.; Sahin, F. GR-ConvNet v2: A Real-Time Multi-Grasp Detection Network for Robotic Grasping. *Sensors* **2022**, *22*, 6208, doi:10.3390/s22166208. [CrossRef]

21. Wei, J.; Liu, H.; Yan, G.; Sun, F. Robotic Grasping Recognition Using Multi-Modal Deep Extreme Learning Machine. *Multidimens. Syst. Signal Process.* **2017**, *28*, 817–833, doi:10.1007/s11045-016-0389-0. [CrossRef]

22. Trottier, L.; Giguère, P.; Chaib-draa, B. Dictionary Learning for Robotic Grasp Recognition and Detection. *ArXiv Prepr. ArXiv160600538* **2016**, doi:10.48550/arXiv.1606.00538. [GoogleScholar]

23. Wang, Z.; Li, Z.; Wang, B.; Liu, H. Robot Grasp Detection Using Multimodal Deep Convolutional Neural Networks. *Adv. Mech. Eng.* **2016**, *8*, 1687814016668077, doi:10.1177/1687814016668077. [CrossRef]

24. Redmon, J.; Angelova, A. Real-Time Grasp Detection Using Convolutional Neural Networks. In Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA); May 2015; pp. 1316–1322. [CrossRef]

25. Ainetter, S.; Fraundorfer, F. End-to-End Trainable Deep Neural Network for Robotic Grasp Detection and Semantic Segmentation from Rgb. In Proceedings of the 2021 IEEE International Conference on Robotics and Automation (ICRA); IEEE, 2021; pp. 13452–13458. [CrossRef]

26. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2016; pp. 779–788. [CrossRef]

27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, 2015; pp. 234–241. [GoogleScholar]

28. Guo, J.; Han, K.; Wu, H.; Tang, Y.; Chen, X.; Wang, Y.; Xu, C. CMT: Convolutional Neural Networks Meet Vision Transformers. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: New Orleans, LA, USA, June 2022; pp. 12165–12175. [CrossRef]

29. He, T.; Zhang, Z.; Zhang, H.; Zhang, Z.; Xie, J.; Li, M. Bag of Tricks for Image Classification with Convolutional Neural Networks. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); IEEE: Long Beach, CA, USA, June 2019; pp. 558–567. [CrossRef]

30. Hendrycks, D.; Gimpel, K. Gaussian Error Linear Units (Gelus). *ArXiv Prepr. ArXiv160608415* **2016**, doi:10.48550/arXiv.1606.08415. [GoogleScholar]

31. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In Proceedings of the International conference on machine learning; pmlr, 2015; pp. 448–456. [GoogleScholar]

32. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ArXiv Prepr. ArXiv201011929* **2020**, doi:10.48550/arXiv.2010.11929. [GoogleScholar]

33. Aladago, M.M.; Piergiovanni, A.J. Compound Tokens: Channel Fusion for Vision-Language Representation Learning. *ArXiv Prepr. ArXiv221201447* **2022**, doi:10.48550/arXiv.2212.01447. [GoogleScholar]

34. Zhang, Y.; Choi, S.; Hong, S. Spatio-Channel Attention Blocks for Cross-Modal Crowd Counting. In Proceedings of the Proceedings of the Asian Conference on Computer Vision; 2022; pp. 90–107. [GoogleScholar]

35. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition; 2018; pp. 7132–7141. [PubMed]

36. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Proceedings of the Advances in Neural Information Processing Systems; Curran Associates, Inc., 2015; Vol. 28. [CrossRef]

37. Depierre, A.; Dellandréa, E.; Chen, L. Jacquard: A Large Scale Dataset for Robotic Grasp Detection. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS); October 2018; pp. 3511–3516. [CrossRef]