Academic Journal of Science and Technology
ISSN: 2771-3032 | Vol. 9, No. 1, 2024

The Study on Recognition and Detection of Express Package Grabbing Based on Machine Vision

Weixiang Qi, Xiangguo Sun

School of Mechanical Engineering, Sichuan University of Science & Engineering, Yibin, 644002, Sichuan, China

Abstract: This research addresses the automated supply of parcels in logistics by employing a robotic arm to replace manual intervention. The process involves the suction-based retrieval of disorderedly stacked parcels and their placement onto a conveyor belt, presenting an economically significant engineering challenge. This paper designs a visual recognition system to assist the robotic arm in identifying the grasping surface of disorderedly stacked parcels. Leveraging transfer learning, a pre-trained model constructs a multimodal network framework. Pre-trained models of ResNet 169, DenseNet121, ResNet101, and ResNet50 serve as backbone networks to train and test on custom parcel datasets. Post comparative testing of model performance, DenseNet169 is chosen to construct the visual recognition network. Specifically, the RGB and Depth data of parcels are separately fed into DenseNet169 for feature extraction. Post-feature extraction, multimodal fusion is applied, integrating a lightweight attention mechanism, CBAM, to enhance semantic segmentation accuracy. Subsequently, post-processing of image features filters out the background, achieving precise identification of the parcel grasping region. Ultimately, the constructed network achieves an average TOP1 true class accuracy of 95.86% on the test set. Thus, the designed parcel visual recognition system based on this methodology exhibits robustness, meeting the demand for autonomous parcel retrieval by the robotic arm. It offers significant support towards resolving challenges in automated parcel supply within logistics.

Keywords: ResNet, DenseNet, Deep learning, Convolutional neural networks, Courier parcels, Visual recognition of gripping surfaces.

1. Introduction

In recent years, with the rapid development of e-commerce and internet technology, the volume of express parcels has continued to soar, leading to an urgent demand for rapid and efficient sorting [1]. In the realization of automated supply of packages in the logistics front end, the identification and detection of express parcels have become a critical link. The rapid advancement of deep learning has further propelled the progress of target detection technology [2-4], offering more promising prospects for logistics automation [5].

He K et al. [6] introduced the concept of residual expression into the architecture of convolutional neural networks, proposing a deep residual convolutional neural network framework to alleviate the difficulty of training deep neural networks. The residual network enhances classification accuracy by increasing network depth, effectively avoiding the problem of decreased accuracy after reaching a certain number of layers. Huang et al. [7] proposed the DenseNet network architecture, where each layer's input is a combination of feature maps from preceding layers, and the feature maps of that layer are used as inputs for all subsequent layers. This architecture mitigates the vanishing gradient problem to some extent through feature reuse and skip connections, enhancing feature propagation, and significantly reducing the number of parameters.

To enhance the robustness of target recognition algorithms, Eitel et al. [8] proposed a multi-modal target recognition network architecture consisting of two relatively independent neural network streams. Each network stream primarily processes a specific modality of information and integrates different modality features through a late fusion approach. During the training of this network, depth data is rendered into three-channel color images, enabling model training without the need for large-scale depth datasets. Simultaneously, augmenting data is achieved by corrupting depth images with noise patterns from real-world scenarios, thus improving the robustness of deep neural networks. This approach demonstrated improved performance on multi-modal datasets and showcased recognition capabilities in actual noisy environments.

Due to the capability of capturing both RGB images and depth images of objects simultaneously [9], RGBD cameras enable the acquisition of three-dimensional positional information of objects [10,11]. Consequently, RGBD cameras are increasingly being applied in the field of robotic arm visual grasping.

Dealing with cluttered and disorganized items becomes challenging when inferring their relative poses from textured color images due to factors such as sensor noise, occlusion, and obstruction [12,13]. Depth image data contains rich geometric information. Hence, when picking up objects with planar surfaces in crowded environments, using depth images is more suitable. This approach reduces reliance on target image and texture feature extraction, thereby enhancing the robustness of visual recognition systems.

Zeng et al. [14] proposed an 'grasp-first-recognize-later' approach to address the grasping problem of unknown objects in cluttered scenes, significantly reducing the complexity of recognition. They utilized a full convolutional residual network, ResNet101, to process multi-modal information from RGB images and depth maps. This network outputted labels for grasping regions. Simultaneously, the depth map was transformed into a 3D point cloud, enabling the calculation of surface normal vectors for grasping areas, thereby providing information about grasp angles.

2. The Collection and Annotation of Datasets

Given the absence of publicly available datasets for express parcels, we independently collected images of 15 different types of express parcels using an RGBD camera, comprising 8 types of cardboard boxes, 2 types of bubble-wrapped flat items, and 5 types of soft pouches (refer to Figure 1). Throughout this process, we considered combinations of parcels with varying quantities, types, and orientations, amassing a total of 713 images. These collected images were annotated. Subsequently, we augmented the dataset through operations such as rotation and mirroring, resulting in an expanded dataset containing 2,842 images.

Figure 1. Different Types of Packages

The annotated images were generated based on the collected RGB images (detailed in Figures 2 and 3). Unlike the conventional coordinate-based annotation methods used in traditional object detection, the labeling results in this study's model are presented in grayscale format. Three colors (black, gray, white) represent three labels (non-graspable area, graspable area, background area). Typically, the non-graspable area refers to parcels unsuitable for grasping, such as those mostly covered by other parcels or positioned at the edges.

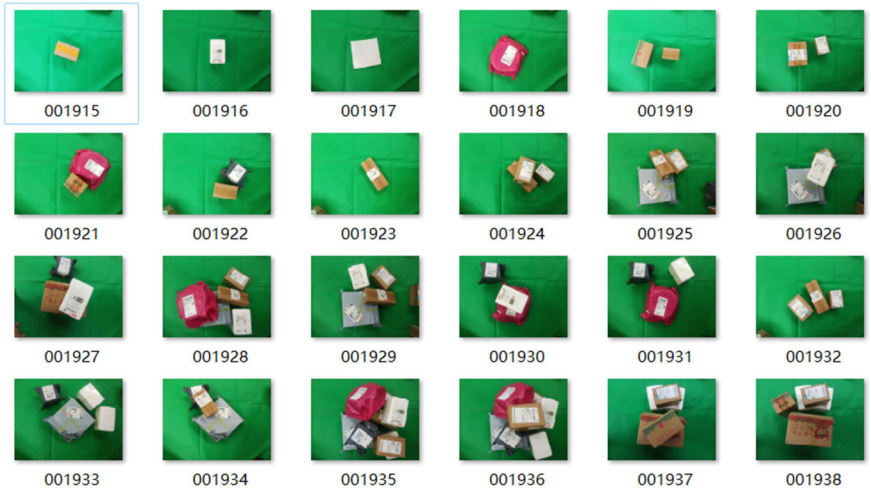

Figure 2. Captured RGB Image

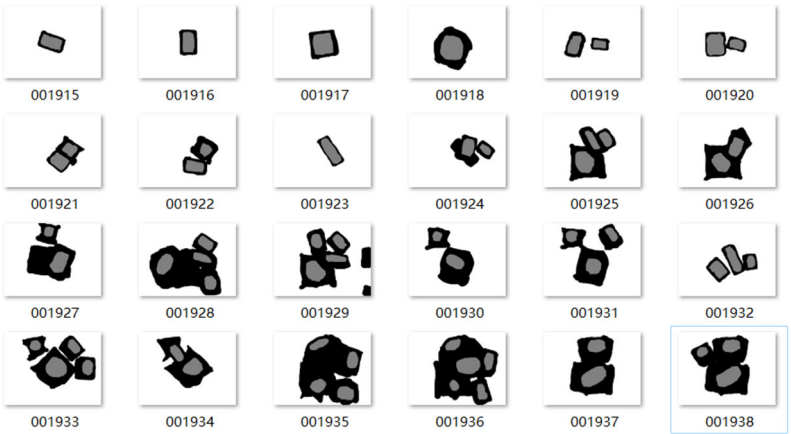

Figure 3. Annotated Image

3. The Model Design

In this study, for the suction cup grasp surface recognition of express parcels, we propose a detection and identification network called ParcelNet. The structure of the network is illustrated in Figure 4.

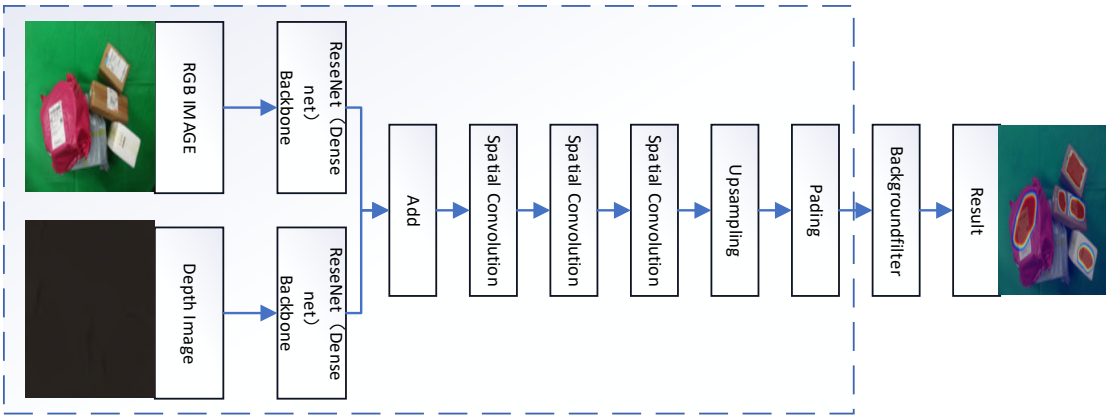

Figure 4. Parcel Recognition Network ParcelNet Model Architecture

The visual recognition network model adopts a step-by-step processing approach to handle the collected express parcel data. Initially, RGB and Depth data are separately fed into selected pre-trained residual networks (ResNet) or densely connected networks (DenseNet) to extract feature information from the images. Subsequently, these extracted features undergo multimodal fusion to obtain a more comprehensive and enriched information representation. The specific process involves multimodal concatenation followed by a series of convolutional operations to fuse the concatenated features, including adjustments in channel numbers, upsampling, padding, and other operations, completing the feature extraction process. Once the feature extraction is completed, a background filtering is applied to the grasp surface features, yielding the foreground grasp surfaces. This process effectively integrates the multimodal information of the data and extracts the target grasp surface features that meet the grasping requirements.

In the rapidly evolving field of deep learning, transfer learning plays a crucial role in two main aspects: feature extraction and fine-tuning. In feature extraction, we leverage knowledge acquired by pre-trained models on standard datasets (such as ImageNet) but excluding the top-level classification structure. Subsequently, a new classifier is built on this foundation to perform specific classification tasks. Removing the top-level classifier from the pre-trained model renders it as a universal feature extractor, capable of extracting effective feature information from new datasets. Meanwhile, in fine-tuning, the weights of the pre-trained model are considered initial values for the new task and are adjusted and updated during training. This approach aims to fine-tune the general feature maps towards specific features relevant to the new dataset, enabling the adaptation of universal functionalities to specific tasks without overriding general learning.

The transfer learning method employed in the study is the second approach, utilizing pre-trained weights from ImageNet to compensate for the relatively smaller training dataset. Through transfer learning, it helps prevent model overfitting due to the smaller dataset size.

In our study, for the Backbone depicted in Figure 4, we introduced several mainstream models pre-trained on ImageNet, including ResNet50, ResNet101, DenseNet121, and DenseNet169. To clearly distinguish between these models, we assigned unique names to the networks built based on different pre-trained models. For instance, the network constructed using ResNet50 is referred to as ParcelNet(R50), the one corresponding to ResNet101 is named ParcelNet(R101), the network based on DenseNet121 is labeled ParcelNet(D121), and the network derived from DenseNet169 is designated as ParcelNet(D169). This naming convention helps accurately identify and reference the source and specific attributes of the utilized models.

The parameter comparison results for these network models are shown in Table 1.

Table 1. Comparison of model parameters

Net	Parameter amount
ParcelNet(D169)	14254979
ParcelNet(D121)	8068995
ParcelNet(R101)	44663875
ParcelNet(R50)	25671747

The table clearly demonstrates that the express parcel detection and recognition network model constructed using DenseNet has fewer parameters compared to the one built using ResNet. This disparity arises from the skip connections employed by ResNet, which increase the overall parameter count. These connections necessitate additional weights to learn residuals for gradient propagation and alleviation of gradient vanishing issues. In contrast, DenseNet utilizes a dense connectivity structure, directly linking each layer with all other layers by concatenating outputs from preceding layers as inputs to subsequent layers. This facilitates more efficient parameter sharing and reuse, allowing DenseNet to effectively utilize parameters to a greater extent. This design enables DenseNet to achieve greater computational efficiency despite having several times more layers than ResNet. Although DenseNet boasts multiple layers compared to ResNet, it exhibits lower computational overhead.

4. Model Training and Testing

Utilizing the previously annotated 2,842 images along with their corresponding labels, we randomly partitioned the data into training and test sets in a ratio of 8:2. Subsequently, these

datasets were fed into the model for training, followed by testing post-training to evaluate the performance of each model.

4.1. Model Training

In this study's experiments, we utilized the Intel(R) Xeon(R) Gold 6234 processor and executed operations on an NVIDIA GTX 2080 Ti GPU. The chosen deep learning framework was PyTorch 1.12, with CUDA version 12.0, and the model training was conducted using Python version 3.8.

The cross-entropy loss function, sometimes referred to as pixel-wise cross-entropy loss function, was employed in our approach. The primary objective of this loss function is to compare the model's predicted class probabilities for each pixel position with the actual pixel-level labels, ensuring accurate pixel-wise classification by the model. The mathematical expression for the cross-entropy loss function is as follows:

$$\mathrm{CE}(p,q) = -\sum_i p_i \log(q_i)$$

In this study, we utilized hyperparameters set at a learning rate of 0.001, an assumed training cycle of 200 epochs, and a weight decay coefficient of 1e-4. To prevent model overfitting, an early stopping strategy was implemented. We trained the training dataset using ParcelNet(R50), ParcelNet(R101), ParcelNet(D121), and ParcelNet(D169). After 150 training batches, the training loss for all models dropped to 1% and stabilized. Notably, DenseNet169 exhibited the lowest loss rate, decreasing to 0.5% (refer to Figure 5). Subsequently, we tested these trained models on a test set containing 527 images, obtaining the TOP1 graspable surface detection accuracy for each model, as presented in Table 2. These results offer significant empirical evidence supporting our choices.

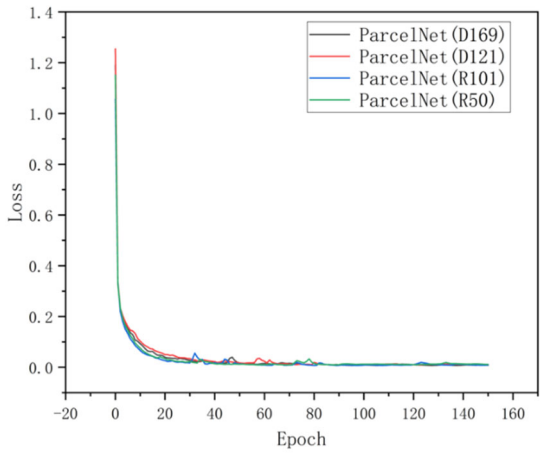

Figure 5. Training Loss Graph

4.2. Testing and Evaluation

In evaluating the model's performance, various metrics come into play, including overall classification accuracy, recall (also known as sensitivity), and precision. In the context of recognizing and grasping express parcels with a robotic arm, as the arm typically handles one parcel at a time, the robustness of the object detection network lies in its continuous ability to correctly detect at least one graspable surface. Reflecting this, our evaluation metric is the precision of inference proposals against manually annotated accuracy. If the center pixel of a proposal is manually labeled as a suction-cup graspable area, it's considered a true positive (TP). If labeled as a non-graspable area but detected as graspable, it's a false positive (FP). Among the detected graspable regions, the one with the largest area is designated as TOP1. Therefore, we focus on the precision of TOP1 graspable areas as the performance metric, utilizing Top1 True Positive Accuracy to assessre cognition performance ,expressedas:

$$Precision_{(TOP1)} = \frac{TP}{TP + FP}$$

After completing model training, we evaluated the models using a test set comprising 568 images to validate their accuracy. Below are the average accuracies from the tests conducted for each model, as detailed in Table 2.

Table 2. Comparison of test set prediction

Net	Top1
ParcelNet(D169)	93.39%
ParcelNet(D121)	88.21%
ParcelNet(R101)	90.56%
ParcelNet(R50)	82.32%

The experimental results indicate that ParcelNet(D169) exhibits the best performance, achieving a TOP1 accuracy of 93.39%. Following closely is ParcelNet(R101) with a TOP1 accuracy of 90.56%. Notably, ParcelNet(D169) comprises 14,254,979 parameters, while ParcelNet(R101) consists of 44,663,875 parameters. Considering the trade-off between model performance and parameter count, the detection and recognition network constructed based on DenseNet169, ParcelNet(D169), was ultimately selected.

5. Model Optimization

In order to further improve the system's robustness and optimize the accuracy of semantic segmentation, we introduced CBAM (Convolutional Block Attention Module) [15], which is a lightweight attention mechanism. We compared the model parameters before and after introducing CBAM (detailed in Table 3).

Table 3. Comparison of model parameters

Net	Parameter amount
ParcelNet(D169)+CBAM	14289991
ParcelNet(D169)	14254979

CBAM, as a lightweight convolutional attention module, integrates channel and spatial attention mechanisms without significantly increasing the model parameters (refer to Figure 6). Its implementation involves several steps: Firstly, global MaxPool and AvgPool operations are applied to the feature maps to obtain two distinct feature vectors. Subsequently, these two feature vectors pass through a shared perceptron for processing and are merged into a completely new feature vector. Finally, a Sigmoid operation is performed on this new feature vector to obtain attention weights in the channel domain. This approach by CBAM, through extraction and fusion of different features, effectively enhances the performance of deep learning models.

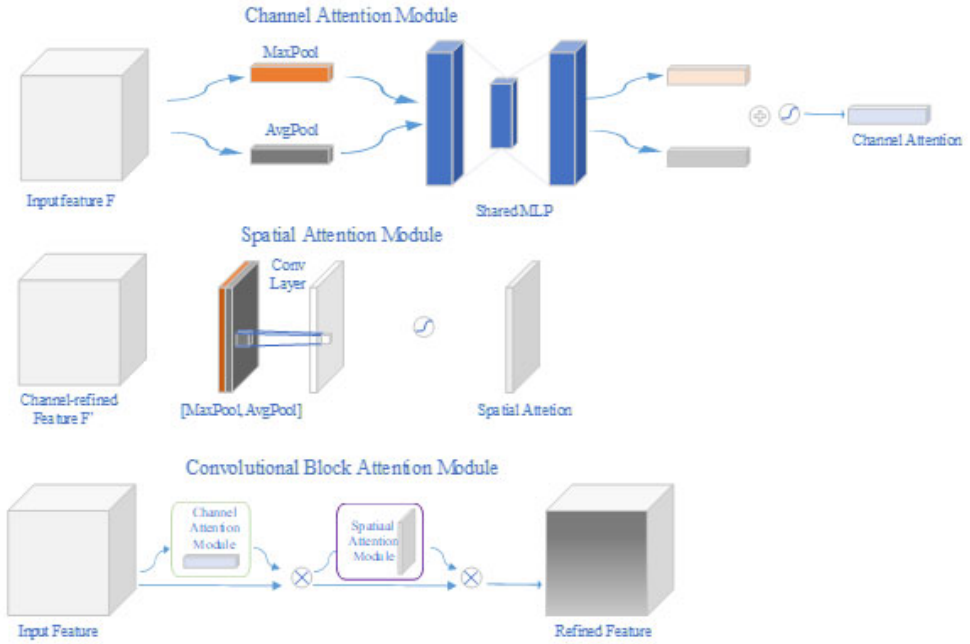

Figure 6. Schematic Diagram of CBAM Attention Module

After the multimodal concatenation in the model, we further optimize it by inserting the Convolutional Block Attention Module (CBAM) attention mechanism between convolutional layers. The specific structure is illustrated in Figure 7.

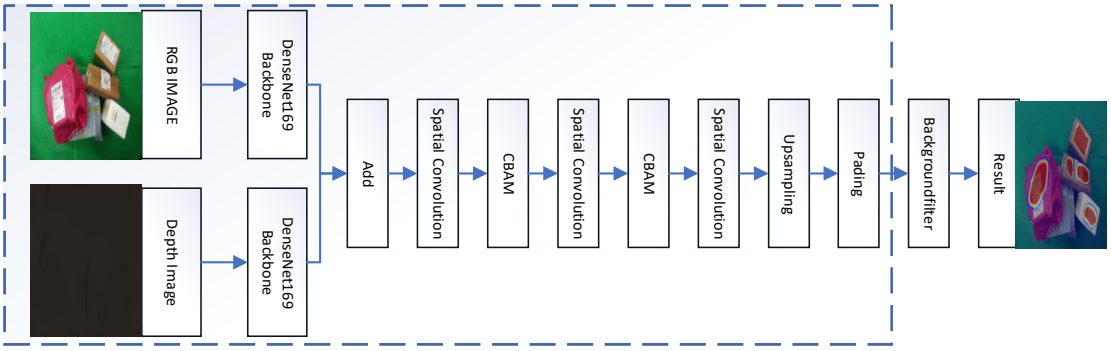

Figure 7. ParcelNet(D169)+CBAM

After integrating the attention mechanism into the model, we employed the previous training approach and conducted training for 150 epochs on the training dataset. During this process, we observed the training loss function graph* (refer to Figure 8) and noted that the training loss also decreased to a level of 0.5% and remained stable.

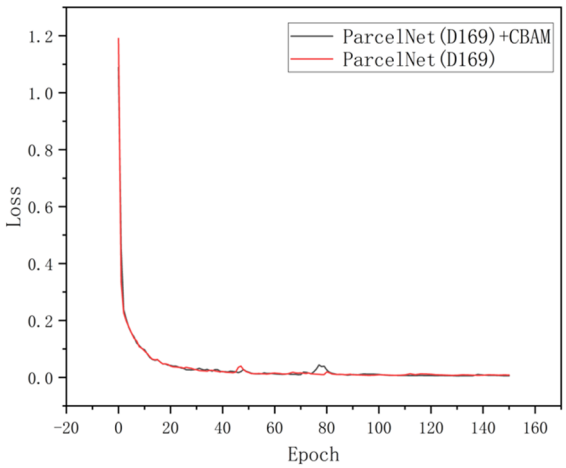

Figure 8. Training Loss Function Graph*

After completing the model training, we tested its performance using a test set containing 568 images, with detailed results provided in Table 4. Upon integrating the CBAM module, ParcelNet (D169) achieved a TOP1 accuracy of 95.86%. This represents a gain of over 2% in accuracy compared to the model without the CBAM module. We can distinctly observe the impact of CBAM on the model's performance. The introduction of this attention mechanism helps the model better capture and utilize contextual information within images, thereby enhancing the accuracy of semantic segmentation. Consequently, it improves the detection accuracy of ParcelNet (D169).

Table 4. Comparison of test set prediction

Net	Top1
ParcelNet(D169)+CBAM	95.86%
ParcelNet(D169)	93.39%

6. Conclusion

In this study focusing on unstructured package grab area recognition, we constructed a multimodal recognition network using different pre-trained deep learning models. Through a series of training and testing experiments and

considering comprehensive results, we selected ParcelNet (D169) built upon DenseNet for our approach. Concurrently, we optimized the network by introducing CBAM, a lightweight attention mechanism, resulting in a significant improvement in recognition accuracy, elevating its TOP1 accuracy to 95.86%. This enhancement greatly boosted the network's robustness, providing a significant solution to unstructured package grab area recognition.

References

[1] Bai Wenjie. Research on Logistics Sorting and Planning System Based on Deep Learning of Express Delivery Waybill [D]. Anhui University of Science and Technology, 2021.

[2] Wang C, Bai X, Wang X, et al. Self-supervised multiscale adversarial regression network for stereo disparity estimation[J]. IEEE Transactions on Cybernetics, 2020, 51(10): 4770-4783.

[3] Zhang J, Xie Z, Sun J, et al. A cascaded R-CNN with multiscale attention and imbalanced samples for traffic sign detection[J]. IEEE access, 2020, 8: 29742-29754.

[4] Alasmari N, Alohali M A, Khalid M, et al. Improved metaheuristics with deep learning based object detector for intelligent control in autonomous vehicles[J]. Computers, 2023, 108: 108718.

[5] Localization Algorithm and System Design for Vision-Based Navigation AGV Journal of Qingdao University(Natural Science Edition), 2022, 35: 83-91.

[6] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016: 770-778.

[7] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 4700-4708.

[8] Eitel A, Springenberg J T, Spinello L, et al. Multimodal deep learning for robust RGB-D object recognition[C]. 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015: 681-687.

[9] Wang Y, Wang C, Long P, et al. Recent advances in 3D object detection based on RGB-D: A survey[J]. Displays, 2021, 70: 102077.

[10] Rahman M M, Tan Y, Xue J, et al. Notice of violation of IEEE publication principles: Recent advances in 3D object detection in the era of deep neural networks: A survey[J]. IEEE Transactions on image processing, 2019, 29: 2947-2962.

[11] Arnold E, Al-Jarrah O Y, Dianati M, et al. A survey on 3d object detection methods for autonomous driving applications[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 20(10): 3782-3795.

[12] Kruglyak L, Lander E S. Complete multipoint sib-pair analysis of qualitative and quantitative traits[J]. American journal of human genetics, 1995, 57(2): 439.

[13] Vidal J, Lin C-Y, Martí R. 6D pose estimation using an improved method based on point pair features[C]. 2018 4th international conference on control, automation and robotics (iccar), 2018: 405-409.

[14] Zeng A, Song S, Yu K-T, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching[J]. International Journal of Robotics Research, 2022, 41(7): 690-705.

[15] Woo S, Park J, Lee J-Y, et al. Cbam: Convolutional block attention module[C]. Proceedings of the European conference on computer vision (ECCV), 2018: 3-19.