Proceedings of the 4th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/46/20241180

CNN advancements and its applications in image recognition: A comprehensive analysis and future prospects

Man Xiao

International Department, Wuhan University of Science and Technology, Wuhan, 430081, China

jlucky80619@student.napavalley.edu

Abstract. With the rapid advancement of industrialization, it has become increasingly evident that conventional image recognition methods are inadequate to meet the burgeoning demands for processing vast quantities of engineering image data. In response to this challenge, convolutional neural networks (CNNs), a cornerstone of deep learning algorithms, have emerged as powerful tools for feature extraction within the realm of image recognition. This paper endeavors to elucidate the underpinnings of image recognition theory and the intricacies of convolutional neural networks. It not only provides an in-depth exploration of the structural characteristics of three iconic CNN models, along with illustrative real-world applications but also delves into a comparative analysis of the practical efficacy and constraints associated with various advanced optimization techniques applied to CNN models. The paper's central focus lies in assessing the real-world applications of CNNs, particularly in the domains of medical diagnosis, agricultural disease identification, and transportation systems. It scrutinizes the impact and limitations of these advanced optimization techniques in each context. Subsequently, it culminates in a comprehensive discussion of the challenges that lie at the nexus of CNN technology and image recognition, providing insights into the prospects of this dynamic intersection.

Keywords: Convolutional Neural Network, Image Recognition, Deep Learning, Unsupervised Learning.

1. Introduction

In the past decade, the field of deep learning has experienced a rapid and transformative evolution, and at the forefront of this progress stands the convolutional neural network (CNN) as a venerable and highly regarded architectural design. This enduring framework has garnered substantial recognition and adoption, particularly in the domain of image recognition. Its popularity can be attributed to its remarkable aptitude for addressing a formidable challenge: the analysis and processing of vast volumes of industrialized image data, a task that has often posed insurmountable obstacles for traditional approaches [1].

This paper begins by describing the conventional image recognition steps and the constraints associated with its feature extraction, and thus puts forward the advantages and significance of convolutional neural network processing image recognition. Subsequently, the architectural features of convolutional neural network are elaborated in detail, with a focus on comparing and analysing three

DOI: 10.54254/2755-2721/46/20241180

classical convolutional neural network models and their application examples. Next, three popular areas for applying convolutional neural networks in image recognition are highlighted, focusing on comparing the latest case studies in each area to illustrate the effects and limitations of the optimization models employed. Lastly, challenges and the future outlook of convolutional neural networks within image recognition are discussed.

2. Theoretical Foundations

2.1. Image Recognition

When the human brain performs image recognition, it will match the object it sees with the same or similar things in memory, which is analogous to the image recognition principle of computer.

Traditional image recognition comprises four distinct stages: image acquisition, preprocessing, feature extraction, and matching [2]. Image preprocessing, a pivotal step within this workflow, involves a series of operations such as localization, normalization, enhancement, and segmentation applied to the original image. These operations collectively prepare the image for subsequent feature extraction while simultaneously enhancing recognition accuracy. Feature extraction is to selectively consider highly distinguishable, stable and reliable image features for extraction, such as the natural features carried by the image content itself: brightness, size, texture, edge contour, etc.; the artificial features obtained from the analysis and processing of the image: color histogram, moment features, etc. Feature matching involves regularly comparing the extracted feature information with existing templates in the database and outputting the template information with the highest similarity.

Feature extraction stands out as a pivotal stage, intricately tied to the ultimate success of image recognition. Conventional methods, however, suffer from a notable drawback: they heavily rely on manually crafted feature extractors, necessitating a high level of expertise and intricate tuning procedures. In contrast, deep learning offers a solution to this limitation by embracing a data-driven approach to feature extraction. This approach harnesses the power of extensive training data, enabling the automatic acquisition of deep, dataset-specific feature representations. Consequently, this methodology yields a more efficient and precise dataset representation, facilitating the extraction of abstract features that exhibit heightened robustness and generalization. Furthermore, it seamlessly integrates these learned features with the classifier, effectively fulfilling the task of image recognition. Figure 1 illustrates the flowchart of image recognition based on the deep learning model.

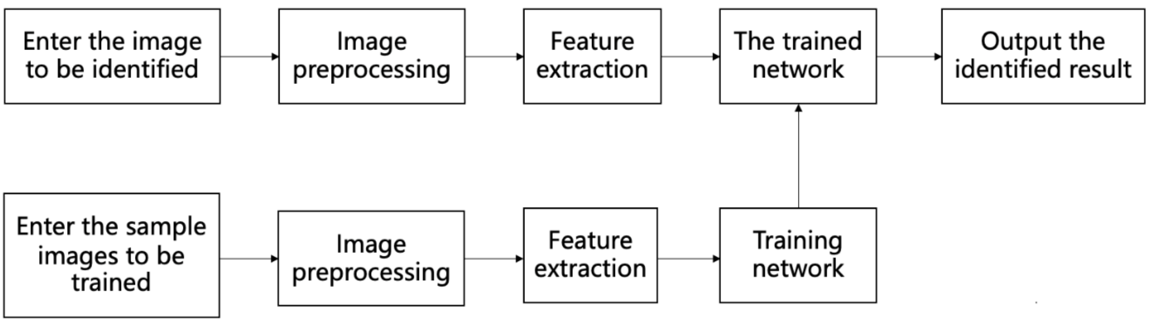

Figure 1. Flowchart of image recognition (Photo/Picture credit: Original).

In deep learning, image recognition is greatly facilitated by the convolutional neural network. It adeptly overcomes the issue of sparse essential data in images that may impede convergence if the model overly emphasizes superfluous data. Moreover, as it does not necessitate extensive preprocessing of the initial image, image recognition can be undertaken by directly inputting the original image and extracting image features via convolution operation. This method offers the benefits of proficient feature extraction, high accuracy of recognition, and robust implementation, resulting in its widespread usage.

DOI: 10.54254/2755-2721/46/20241180

2.2. Convolutional Neural Network

2.2.1. The Network Structure

The convolutional layer comprises convolutional kernels which execute convolutional operations on the pixel values transmitted from the input layer to derive feature information. In practical image processing assignments, a collection of convolutional kernels extracts only one feature, whilst multiple sets of kernels extract several features concurrently.

The result of convolution needs to be activated by the corresponding activation function so that the network satisfies a nonlinear relationship between the input and output, with the aim of enhancing the expressive power of the neural network. Among them, the most commonly used activation models are ReLU, Sigmod, tanh and other functions.

The role of the pooling layer is to downsize the feature data following convolution, eliminating surplus information, and thereby diminishing the computational burden of the model. There are three prevalent techniques for pooling: maximum, minimum, and average computations.

Conversely, the fully connected layer typically resides in the network's concluding layers. Its primary function lies in categorizing the extracted features, computing the likelihood that the input image pertains to each category, identifying the category with the highest likelihood, and deriving the response value output. This intricate process culminates in the accomplishment of image recognition. Taking the MNIST handwritten digit recognition dataset as an example, Figure 2 shows a complete CNN architecture.

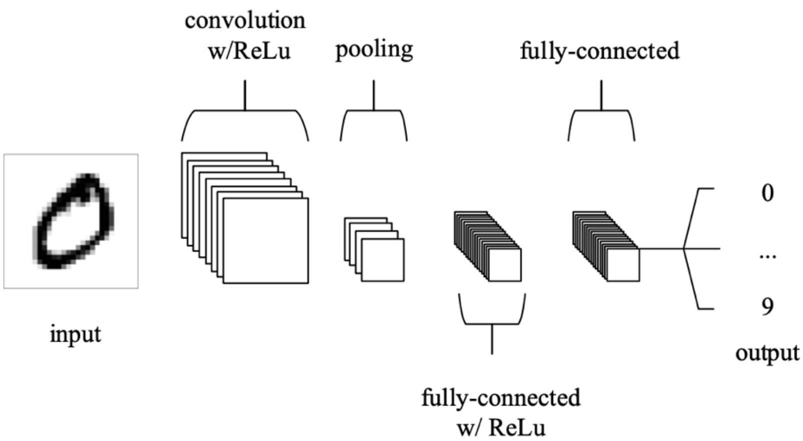

Figure 2. The CNN architecture [3].

The function of convolutional neural networks in image recognition can be expressed abstractly as follows: convolutional layers and activation layers are clustered together to form a feature extractor, and multiple fully connected layers are cascaded to form a classifier.

2.2.2. Training

The training of a convolutional neural network encompasses two distinct phases: forward propagation and back propagation [4]. During the forward propagation phase, data flows from lower to higher levels in the network. This process entails processing an input image through multiple layers of convolution and pooling to extract the feature vectors, which are then passed into the fully connected layer for classification and recognition. We evaluate the classification based on our expectations. The results are output if the it matches our expectation. When the result obtained from forward propagation fails to meet the expectation, meaning the error exceeds the accepted threshold, backpropagation is carried out to transmit the computational error information from higher to lower level. This permits the network to update and optimize the parameters of each layer by reducing the error value, and thereby enhancing the model's accuracy in image recognition.

3. Three Classical Models

3.1. LeNet-5

LeNet-5 was proposed by Yann LeCun et al. in 1998, which is regarded as the seminal network model in the field. It is also the first neural network extensively used for digital image recognition [5]. it contains a 7-layer network structure, including 2 stacked convolution-pooling layers, 2 fully connected layers and the output layer, which can be seen in Figure 3.

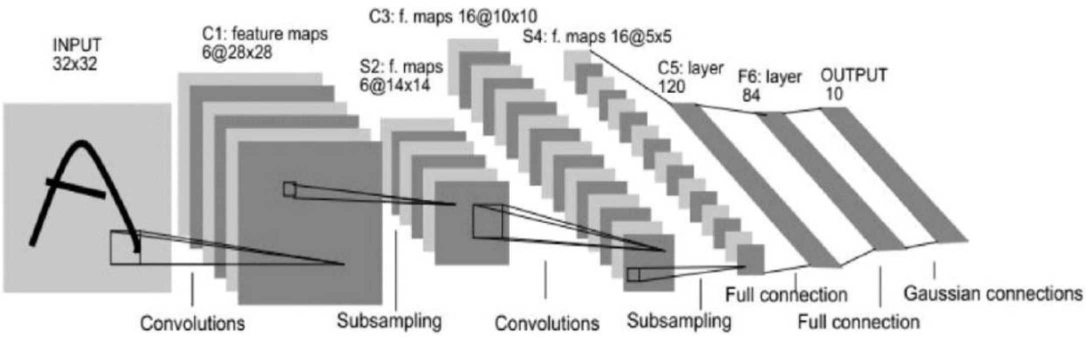

Figure 3. The model of LeNet-5 [5].

Among them, the convolutional layer extracts the spatial features of the image by utilizing a convolutional kernel of size 5*5. The activation function used is tanh. For reducing the dimensionality of the image data, average pooling is employed by the pooling layer. The full connectivity layer accumulates the fusion feature data from the previous layer, connecting all the nodes before passing them to the output layer. And the ten neurons in the output layer correspond to the numbers zero to nine in turn, outputting the final classification results.

Therefore, this model was primarily used for early handwritten digit recognition and classification. The recognition accuracy rate for the MNIST handwritten digit recognition dataset can reach up to 95% [6]. Due to the classical nature of the LeNet-5 model, it is simple to enhance the accuracy and generalizability of the model by implementing some improvements. For instance, Deng et al. eliminated the C5 layer from LeNet-5, swapped the activation function tanh with ReLU, downsized the input image (from 32*32 to 28*28), streamlined the number of neurons in each layer, and regulated the convolution kernel step [7]. This led to a significant reduction in computational amount, accelerated model convergence, and shortened the gradient descent training time. The results showed that the misidentification rate of MNIST dataset is 1.18%, with faster recognition speed and higher recognition accuracy.

3.2. ResNet

The Residual Neural Network (ResNet), introduced by Kaiming He's team at Microsoft Research, emerged victorious in the ILSVRC (ImageNet Large Scale Visual Recognition Challenge) of 2015, achieving a remarkable Top-5 error rate of 3.57 percent—nearly half that of the GoogLeNet error rate [8]. Its primary breakthrough lies in the proposal of the deep residual learning framework, designed to counteract the degradation phenomenon. This phenomenon entails training errors failing to decrease as network depth increases, effectively mitigating the challenges associated with excessively deep neural networks. ResNet offers a range of network architectures, each varying in depth, with common models encompassing 34, 50, 101, and 152 layers.

DOI: 10.54254/2755-2721/46/20241180

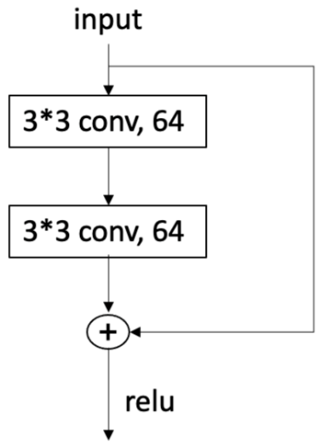

Figure 4. The residual learning block of ResNet [8].

Taking the 34-layer architecture as an example, among them, the most important residual module is composed of two 3 * 3 convolution layers in series, with two data channels. The purpose is to add the input data after two 3*3 convolutions and the original input data for output, and it is activated by the ReLU function, as shown in Figure 4.

ResNet has strong generalization ability, for example, in water conservancy engineering, Chen et al. in order to improve the recognition effect of concrete dam surface operation scene, constructed ResNet50-SEMSF model with ResNet50 as the backbone network structure [9]. The experimental results showed that the accuracy of the validation set was 98.00%, 1.20% higher than the original ResNet50 model, while the accuracy of VGG16 was relatively lower (94.49%).

3.3. MobileNet

MobileNet is an efficient and lightweight network architecture proposed by Google for mobile devices, which can substantially reduce the model size and computational requirements while maintaining high efficiency. Taking MobileNet V1 as an example, the architecture is shown in table 1 [10]. Its core structure is: (1) Replacing the standard convolutions with depthwise separable convolutions. There are two steps: the first step is depth convolution, a convolution kernel has only one channel and is only responsible for operating with one channel of the input layer; the second step is point-by-point convolution, i.e., 1×1 convolution, which upgrades and downgrades the feature map. The combination of these two operations achieves the purpose of reducing the number of parameters. (2) Two hyperparameters α and ρ. α is used to control the number of input and output channels, and ρ is used to control the resolution of the input image. Based on V1 in 2017, google successively introduced MobileNet V2 and V3 in 2018 and 2019 with higher accuracy and smaller models, the highlight of V2 is the introduction of the inverted residual with linear bottleneck, and V3 introduces the squeeze and excitation structure as the lightweight attention model (SE) and the use of h-swish instead of swish function [11, 12].

Table 1. MobileNet Body Architecture [10].

Type/Stride	Filter Shape	Input Size
Conv / s2	3×3×3×32	224×224×3
Conv dw / s1	3×3×32 dw	112×112×32
Conv / s1	1×1×32×64	112×112×32
Conv dw / s2	3×3×64 dw	112×112×64
Conv / s1	1×1×64×128	56×56×64
Conv dw / s1	3×3×128 dw	56×56×128

Table 1. (continued).

Conv / s1	1×1×128×128	56×56×128
Conv dw / s2	3×3×128 dw	56×56×128
Conv / s1	1×1×128×256	28×28×128
Conv dw / s1	3×3×256 dw	28×28×256
Conv / s1	1×1×256×256	28×28×256
Conv dw / s2	3×3×256 dw	28×28×256
Conv / s1	1×1×256×512	14×14×256
5×Conv dw / s1	3×3×512 dw	14×14×512
5×Conv / s1	1×1×512×512	14×14×512
Conv dw / s2	3×3×512 dw	14×14×512
Conv / s1	1×1×512×1024	7×7×512
Conv dw / s2	3×3×1024 dw	7×7×1024
Conv / s1	1×1×1024×1024	7×7×1024
Avg Pool / s1	Pool 7×7	7×7×1024
FC / s1	1024×1000	1×1×1024
Softmax / s1	Classifier	1×1×1000

Guo et al. conducted a comparative analysis of the accuracy and efficiency of network models, including MobileNet, GoogLeNet, and VGG16, using the ImageNet dataset [13]. The study showed that MobileNet exhibited the highest prediction accuracy, fastest speed and best performance, while requiring fewer computations and parameters than the other models.

4. Three applications

4.1. Medical Diagnosis

Medical imaging diagnosis is a crucial tool in modern medicine. As the amount and complexity of imaging data increase, the use of deep learning to expedite medical image recognition to achieve efficient diagnosis of the sick parts of the human body is becoming increasingly popular. In this field, the CNN is widely utilized by training large-scale image datasets and optimizing algorithm models to identify medical images quickly and accurately. In recent years, noteworthy CNN models for detecting medical images include Faster RCNN and YOLOs, while the prominent backbone models for classifying medical images consist of EfficientNet, ResNeXt, Xception, and MobileNets.

For example, in the field of image detection, Salman et al. retrained 500 real prostate tissue biopsy images, used a powerful target recognition model YOLO to automatically locate the important regions and graded the diagnosis of the regions where prostate cancer was detected, and has achieved 89% detection accuracy by testing 137 completely different real prostate tissue biopsy images [14]. In the field of image classification, Cao et al. in response to the growing demand for early diagnosis of thyroid cancer, researched to enhance the automatic grading method of thyroid cancer pathology images with EfficientNet network as the core [15]. On the basis of EfficientNet-B0, they introduced Coordinate Attention to replace the original SE-Net attention mechanism, and feature aggregation in both directions of channel relationship and precise location information to enhance the neural network's attention to cell location information. The experimental results showed that the accuracy of CA-EfficientNet reached 96.6%, which was 0.4% higher than that of EfficientNet-B0 network. Its application to the actual thyroid auxiliary diagnosis system is comparable to the diagnosis results of experienced pathologists, which can greatly reduce the workload of doctors and reduce the rate of manual misdiagnosis.

4.2. Agricultural Pest and Disease Identification

Identifying pests and diseases represents the primary challenge in agriculture. Traditional detection approaches rely on farmers' expertise or professional guidance. The rapid development of convolutional neural network models has greatly improved the efficiency of crop pest and disease identification.

Xu et al. built a lightweight model STM32-MobileNet based on MobileNet V2, deployed on a cost-effective edge device, STM32 [16]. They further reduced the secondary residual connections of the network structure to relieve the RAM pressure without altering the model's accuracy and speed, and compressed the model through quantisation-aware training techniques to improve the model's portability. Finally, the performance was evaluated using the X-CUBE-AI development platform. The results of testing on 5,278 images from the five-classified apple leaf disease dataset showed that the CPU inference runtime for STM32-MobileNet is only 513 s, compared to approximately 661 s for MobileNet and 978 s for MobileNet V2, with both accuracy rates of approximately 90%. Tests on 1651 images from the seven-classified tomato leaf disease dataset also showed that the improved model inference runs faster than the original model (approximately 158 s for STM32-MobileNet. 217 s for MobileNet V2 and 99 s for FD_MobileNet). This successfully ensures the accuracy of image classification while reducing the cost of edge devices and improving agricultural productivity.

However, the agricultural diseases are complex and varied, and obtaining sufficiently large training sets is difficult, thus impeding the CNN network model's recognition accuracy. To tackle this issue, Yao suggested implementing L2/L2M regularization into the Xception network [17]. She conducted recognition experiments on a 1560 seven-classified Peach tree disease dataset using AlexNet, ResNet50, Xception, SENet154, DenseNet169, HRNet-w48, and MobileNet V3 that were incorporated for regularization. The outcomes demonstrated that Xception achieved an accuracy of 92.23%, which was significantly greater than the other six models.

4.3. Transport

The application of CNN algorithms in licence plate recognition and traffic sign recognition is becoming increasingly sophisticated in the field of transport.

For instance, Liu constructed a three-stage number plate recognition model consisting of vehicle detection, number plate detection, and number plate recognition based on optimised CenterNet and ResNet models, which achieved an accuracy of 98.1% on the PKU dataset, proving that the model can achieve highly accurate number plate recognition in natural scenes [18]. The highlights are that the vehicle detection stage improves the model speed by simplifying the Hourglass-104 structure of CenterNet, while using the dilated convolutional network instead of the traditional convolutional neural network, and adding an output branch to the non-linear activation function of the residual block to optimise the network accuracy. In addition,the number plate recognition stage corrects the number plate distortion caused by shooting by adding the WPOD-NET network before the ResNet model.

Wu et al. used the YOLOv7 model to train the Chinese traffic sign detection dataset, and selected three different road scene models, ordinary, occluded and fuzzy, to verify the generalisation ability of the model, and compared it with three popular target detection algorithms, namely CenterNet, Faster R-CNN and SSD [19]. The test results showed that the YOLOv7 algorithm was twice as fast as CenterNet, ten times as fast as Faster R-CNN, and had a higher accuracy than CenterNet and Faster R-CNN, while the SSD algorithm had a higher speed than YOLOv7, but it had a leakage or misdetection on complex roads and blurred traffic symbols. YOLOv7 is able to achieve a higher detection rate than CenterNet and Faster R-CNN in the three scenarios. YOLOv7 can guarantee high detection accuracy with high detection speed in three scenarios, which provides a future research and development direction for traffic symbol detection, and is of great significance for automated driving technology.

5. Conclusion

This paper describes the theoretical basis of image recognition and convolutional neural network architecture, selects three classical models combined with examples, and summarises their architectural features. Finally, by comparing various current examples of medical diagnosis, agricultural disease recognition, and transportation, analyzing the pros and cons of their optimization models, and indicating practical significance of their applications. The previous section has demonstrated that the combination of convolutional neural networks and image recognition has yielded remarkable results, but there is still a wide range of progress to be made in the development of its applications, such as: (1) Strengthening the research on the recognition accuracy of unsupervised learning, which is able to reduce the current large number of image annotation work of supervised learning, thus improving the recognition efficiency. (2) Conducting research on the utilization of transfer learning methods to identify small sample datasets is necessary to overcome difficulties in obtaining large sample datasets in real-life situations. (3) Finding suitable models to solve specific engineering problems to avoid the unsatisfactory results of concrete recognition caused by the algorithm's generalization ability. Further research into these matters will have a pronounced impact on the advancement of engineering applications utilizing convolutional neural networks.

References

[1] Wang Xiao, Zhu Enzhao, Ai Zicheng. The principle of convolutional neural network and its application in medical imaging diagnosis. Chinese Journal of Medical Physics, 2022,39 (12): 1485-1489.
[2] Qi Yanrong. Acceleration and optimization of deep learning image recognition based on FPGA. Zhengzhou University, 2022.
[3] O'Shea K, Nash R. An introduction to convolutional neural networks. arXiv preprint arXiv: 1511.08458, 2015.
[4] Li Z, Liu F, Yang W, et al. A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems, 2021.
[5] LeCun Y. LeNet-5, convolutional neural networks. URL: http://yann. lecun. com/exdb/lenet, 2015, 20(5): 14.
[6] Dulari B, Chirag P, Hardik T, et al. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. Electronics, 2021, 10(20).
[7] Deng Changyin, Zhang Jie. Handwritten digit recognition based on improved LeNet-5 model. Information and Communication, 2018(01): 109-112.
[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.
[9] Cheng Shu, Sun Mengwen, Chen Yun, et al. A ResNet50-SEMSF method for intelligent recognition of concrete dam face operation scenarios. Journal of Hydropower Generation, 2023, 9(15): 1-10.
[10] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv: 1704.04861, 2017.
[11] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510-4520.
[12] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 1314-1324.
[13] Guo Junliang, Zhang Hongchuang. Research and Analysis of Convolutional Neural Network Models. Science, Technology, Innovation and Application, 2021, 11(23): 16-18+22.
[14] Emin M S,Gözde Ç Ç,Jahongir A, et al. Automated prostate cancer grading and diagnosis system using deep learning-based Yolo object detection algorithm. Expert Systems With Applications, 2022, 201.
[15] Cao Liling, Jiang Kehong, Cao Shouqi, et al. A deep learning-based method for grading thyroid cancer pathology images. Chinese Journal of Medical Physics, 2023, 40(05): 580-588.
[16] Xu Paitao, Cheng Xiang. Research on image recognition algorithm of pests and diseases in agricultural Internet of Things based on STM32. Journal of Internet of Things, 2023, 9(8): 1-10.

[17] Yao Na. Research on Intelligent Processing of Peach Tree Disease Images Based on Deep Learning. Huazhong Agricultural University, 2022.
[18] Liu Yanjiao. Three-stage Natural Scene Licence Plate Recognition Based on CenterNet and ResNet. Lanzhou University, 2023.
[19] Wu Xiao, Liu Jiajia, Duan Ping, et al. Road identifier detection based on YOLOv7. Laser Magazine, 2023, 9(8): 1-9.