A Novel Fusion Network for Apple Image Classification and Quantity Recognition

Hanyu Jiang*†
Zhipeng Wang*
Jiahan chen*
GuanYuan Pan
YuDie Jin
22320324@hdu.edu.cn
Hangzhou Dianzi University HDU-ITMO Joint Institute
Hangzhou, ZheJiang, China

ABSTRACT

We focuses on identifying images containing apples from a large number of orchard fruit images and determining the number of apples in each filtered image. We propose a CV5Fnet model that combines traditional OpenCV image processing operations with the Watershed Algorithm and YOLO V5. We first build a high-precision, lightweight fruit classifier to accurately filter apple images from five fruit images in the dataset, and pass apple images to the red apple recognition module and the green apple recognition module based on YOLOV5, which are based on the filters, HSV color space conversion, masking operations, and Watershed Algorithm, respectively. The apple pictures are passed to the red apple recognition module based on filter, HSV color space conversion, mask operation, watershed algorithm and the green apple recognition module based on YOLOV5 to recognize the number of red apples and the number of green apples in the target pictures. and green apples in the target picture respectively, and finally sum up to the number of apples in the picture. In the publicly available dataset 2023APMCM_A_2, the accuracy of our fruit classifier is as high as 99.86%, and the final image processing results show that CV5Fnet has achieved good results in recognizing the number of apples.

CCS CONCEPTS

• **General and reference** → **Experimentation**.

KEYWORDS

image processing, watershed algorithm, YOLO V5, Apple Quantity Identification

ACM Reference Format:
Hanyu Jiang, Zhipeng Wang, Jiahan chen, GuanYuan Pan, and YuDie Jin. 2024. A Novel Fusion Network for Apple Image Classification and Quantity Recognition. In *The International Conference on Computer Vision and Deep Learning (CVDL 2024), January 19–21, 2024, Changsha, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3653781.3653794

1 INTRODUCTION

With the rise in population, the demand for fruits, which are rich in nutrients and vitamins, is increasing year by year[1–3],the traditional fruit industry chain is facing more and more challenges and needs the addition and updating of novel technologies. Currently computer vision technology has been widely used for fruit recognition tasks [4, 5], such as fruit picking robots, fruit sorters [6, 7], etc. Most of the datasets of these recognition systems are camera-generated RGB images, which become the training set for machine learning (ML) after a series of preprocessing and image enhancement operations. Machine learning allows machines to autonomously learn input features and output corresponding results, empowering computers to act without being explicitly programmed to do so [8–10]. Currently, Deep Learning (DL) is one of the most commonly used methods for machine learning, which automatically learns appropriate hyperparameters and improves the accuracy of problem solving through gradient descent algorithms. In particular, Convolutional Neural Network (CNN) [11] is the main deep learning architecture used for image processing [12, 13], e.g., AlexNet uses Convolutional Neural Network (CNN) to process spatial information in images.RCNN proposes a single-stage target detection method, while Fast-RCNN and Faster-RCNN achieve higher efficiency.The emergence of YOLO, which achieves fast target detection while being able to maintain good accuracy, has the widest application in the field of video recognition. However, in the task of fruit recognition in orchards, due to the complex background environment of orchards and the large number of objects to be recognized, the above methods still have some limitations. 1. Poor accuracy for specific tasks: It is difficult for the model to accurately localize and identify fruits due to the presence of leaf shading, fruit shading and mixed shading of orchard fruits. 2, Cost and complexity: Nowadays, most of the image processing tasks are performed with models based on convolutional neural networks. And along with the development of deep neural networks, the structure of the network model is becoming more and more complex, and the amount of

[*] Both authors contributed equally to this research.
[†] Corresponding author.

CVDL 2024, January 19–21, 2024, Changsha, China
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-1819-9/24/01
https://doi.org/10.1145/3653781.3653794

computation is getting bigger and bigger, lightweight equipment such as small fruit picking robots are difficult to carry, and high-end equipment increases the cost of the orchard. To solve this problem, we propose a new lightweight network model (V5FNet) based on OpenCV image feature extraction, Watershed Algorithm and YOLO V5 to solve the task of recognizing and counting apple pictures from an orchard with heavy occlusion phenomenon, taking apple recognition as an example. We will use a fruit classifier to recognize the input apple images, and the recognized images will go to the red apple recognition module and the green apple recognition module respectively. In the red apple recognition module, the model extracts the red features of the image through a series of operations such as filtering, color space transformation, double mask superposition, and so on, so as to exclude the interference of the green leaf occlusion . For the occlusion of the fruit itself, the model combines with the watershed algorithm to segment the attached red blocks, and finally discriminates by the shape of the edges combined with the background image information. The green apple recognition module is a YOLO V5 fine-tuning model trained using the green apple dataset, which is specifically used for green apple recognition. Finally, three images were randomly selected from the test set for wake-up test, and the classifier accuracy was as high as 99.86%, the green apple recognition accuracy reached 87%, and the fusion model accuracy was stable at 92.42% 97.38%.

In a nutshell, our main contributions are mainly as follows:1. A lightweight fruit classifier architecture is built to realize small volume and high accuracy (99.86%).2. After classifying the fruits, the idea of recognizing the number of apples in apple pictures is further proposed, and the traditional OpenCV image processing operations are combined with the Watershed Algorithm as well as the YOLO V5, and the CV5Fnet model is proposed, which has achieved good results. The model can be improved on the basis of this model to further carry out the tasks of apple location determination, quality prediction and ripeness classification.

2 PRELIMINARIES

We focuses on identifying images containing apples from numerous orchard fruit images and determining the number of apples in each screened image . For ease of description, red-green interspersed apples are treated as green apples in this paper. Our dataset and code will be open sourced at https://github.com/IceProgramer/AppleRecognition.

3 OPENCV - YOLO V5 FUSION NETWORK

Fig. 1 shows the whole structure of CV5Fnet, we can see from Fig. 1 that CV5Fnet consists of a fruit classifier, and two different color apple recognition modules, in the following, we will introduce the specific structure of this classifier and the two modules

3.1 Fruit Classifier

Fruit classifier consists of three main types of modules, which we call ConvBlock, ResidualBlock, and Fully Connected Layer (FC).The ConvBlock consists of a convolutional layer, a batch normalization layer, ReLU activation function, and a maximum pooling layer.The ResidualBlock layer consists of two convolutional layers, two batch normalization layers, and two activation functions. Finally, we output the results using FC. See Fig. 2 for detailed structure.

3.1.1 *ConvBlock.*

ConvBlock consists of one convolutional layer, one batch normalization layer, ReLU activation function, and one max pooling layer. Our input $X = [x_1, x_2, \ldots, x_B] \in R^{BCWH}$ will first go through a layer of convolution, and then we will obtain new data $X' \in R^{BC'WH}$. We then go through batch normalization and activation functions, and finally go through the max pooling layer to obtain the output $Y \in R^{BC'\frac{W}{2}\frac{H}{2}}$. We will present the entire process using the following formula:

$$Y = MaxPooling(ReLU(BatchNorm(Conv(X)))) \tag{1}$$

3.1.2 *ResidualBlock.*

The ResidualBlock layer consists of two convolutional layers, two batch normalization layers, and two activation functions. The input $X \in R^{BCWH}$ will first go through a convolutional layer, an activation function, and a max pooling layer to obtain the output $X' \in R^{BC'WH}$, Then, the input X' will first go through a convolutional layer, an activation function, and a max pooling layer to obtain the output $X'' \in R^{BC'WH}$. The last two outputs are added together to obtain the final output Y.

$$X' = \text{MaxPooling}(\text{ReLU}(\text{Conv}(X))) \tag{2}$$

$$X'' = \text{MaxPooling}(\text{ReLU}(\text{Conv}(X'))) \tag{3}$$

$$Y = X' + X'' \tag{4}$$

3.1.3 *FC Layer.*

We will put the previous final output $Y \in R^{BCWH}$ into FC. The final output dimension of the FC layer is 5, and then we take the index with the highest value among the 5 dimensions as the final classification category.

$$C = argmax(FC(Y), dim = -1) \tag{5}$$

3.2 Red Apple Recognition Module

Fig. 3 shows the architecture of the red apple recognition module, which is mainly composed of two sub-modules, namely image processing and masking operation, which successively performs filtering operation, HSV color space transformation, masking operation and red feature extraction, and finally marks all the red apples from the image with an edge frame. The following is a detailed description of each step with the example image in Fig. 4.

3.2.1 *Filtering operations.*

Wang et al.'s study[14] surfaces that the combination of Gaussian filtering and Non-local means filtering leads to better visualization and better evaluation of the model in terms of peak signal-to-noise ratio (PSNR), and we were inspired to try the combination of numerous filters (mean filtering, bilateral filtering, and median filtering) with Gaussian filtering, and ultimately found that the combination of Gaussian filtering first, followed by median filtering was effective in our data set lived the most effective visualization results.

Gaussian filter is a linear smoothing filter that chooses the weights according to the shape of Gaussian function[3].In image processing, zero-mean two-dimensional discrete Gaussian functions are often

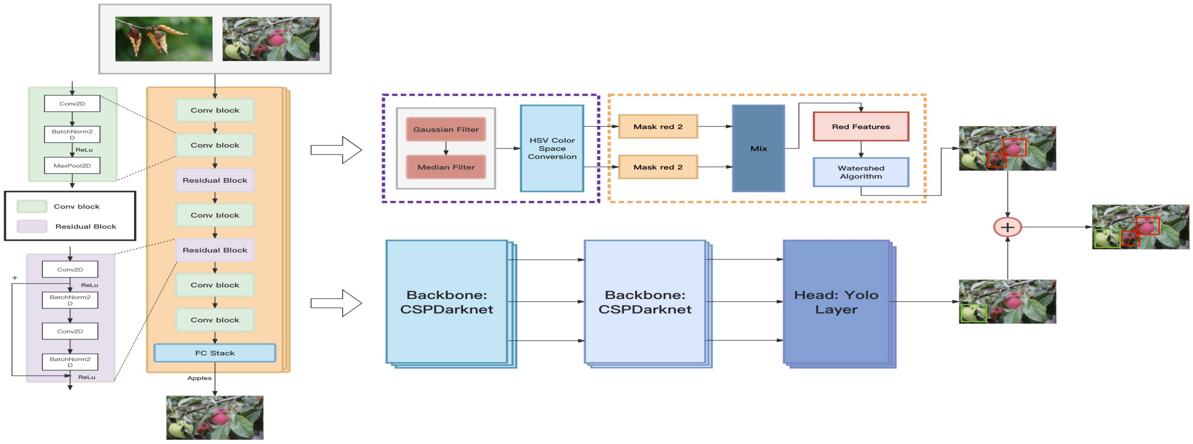

Figure 1: YOLO V5 Model

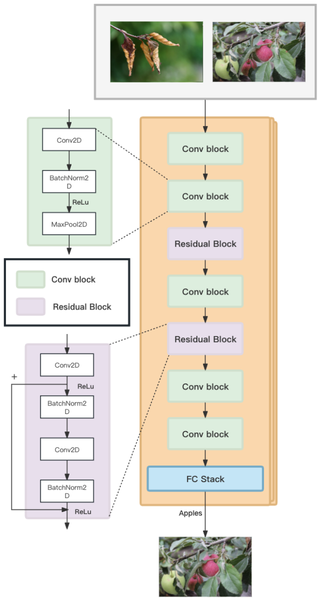

Figure 2: Fruit Classifier

used as smoothing filters[15].

$$g[i,j] = e^{\frac{i^2+j^2}{2\sigma^2}} \tag{6}$$

2-d discrete Gaussian function

where the Gaussian distribution parameter σ determines the width of the Gaussian function.We chose the most commonly used $\sigma = 5$.

Originally proposed as a time series analysis tool in 1971[5], median filtering has since been applied to image processing as well. Median filtering is done by moving a window over the image (or sequence) and replacing the value in the center of the window with the median of the original values within the window, resulting in a smoother image.

$$y_{ij} = Median[x_{i+r,j+s}; (r,s) \in A], (i,j) \in \mathbb{Z}^2 \tag{7}$$

two-dimensional median filter

where A is the filtering window, and we chose A = 5*5 corresponding to the Gaussian filtering.
The results of Gaussian filtering and median filtering are shown in Figs. 6, 7

3.2.2 *HSV Color Space Conversion.*

In HSV color space, color and luminance are separated, the representation of color is more stable, and the range of colors is easier to define, which will be more helpful for the model to identify and segment red apples.The formula for converting RGB images to HSV images is as follows:

$$\begin{aligned}
V &= \max(R,G,B) \\
S &= \begin{cases} \frac{V-\min(R,G,B)}{V}, & V \neq 0 \\ 0,\ \text{otherwise} & \end{cases} \\
H &= \begin{cases} \frac{60(G-B)}{V-\min(R,G,B)}, & V = R \\ \frac{120+60(B-R)}{V-\min(R,G,B)}, & V = G \\ \frac{240+60(R-G)}{V-\min(R,G,B)}, & V = B \end{cases} \\
H &= \begin{cases} H+360, H < 0 \\ H, H \leq\ 0 \end{cases}
\end{aligned} \tag{8}$$

3.2.3 *Masking operation.*

In order to exclude the green leaf occlusion as well as other environmental interferences, we need to extract the red features of the image and ignore the other colors. Masking operations are performed to extract the parts of interest in an image and are commonly used for tasks such as image segmentation and target detection. By creating a mask, we can filter out the uninteresting parts of the image and keep only the regions we care about. This helps to reduce the complexity of subsequent processing and improves the accuracy and efficiency of the algorithm. In the HSV color space, we define two ranges corresponding to two discrete red regions. Then, two masks mask_red1 and mask_red2

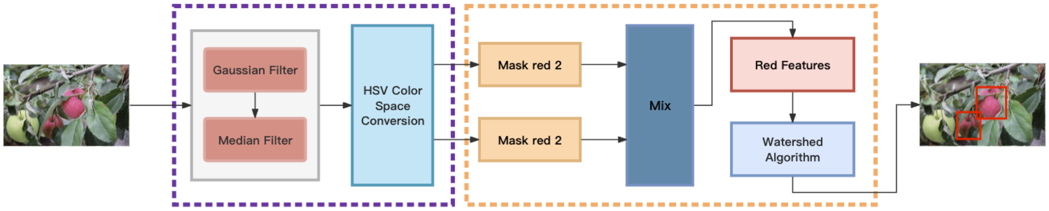

Figure 3: Red Apple Recognition Module

Figure 4: Original drawing

are created based on these two ranges, representing the two red regions respectively. We use these two masks to mask the image and extract the red part to get masked_img1 and masked_img2. finally, these two images are superimposed to get masked_mixed and the red part with the same pixel position as the white part of masked_mixed is intercepted from the original image to get the final result with the extracted red features. Fig. 5 shows the process of a complete masking operation.

3.2.4 Watershed algorithm. Watershed algorithm is an image segmentation technique that uses the gradients and grayscale values in an image to identify object boundaries. This algorithm treats the image as a topographic surface, with grayscale values representing elevations, and then simulates the process of flooding the image to create "watersheds," thereby achieving image segmentation. Our model uses the"Watershed by Flooding" approach first proposed by Vincent and Soille. After the "perforation" of each nadir, the entire terrain begins to be inundated by water. Water gradually fills all catchment basins from the minimum height of the lowest point. At the confluence of the different catchment basins, a watershed is formed. The process ends when the water reaches the highest point of the terrain, so that each catchment basin is covered by a watershed line.

3.3 Green Apple Recognition Module

Inspired by the 90.9% accuracy of Mao et al. in detecting the number of apples in nature using YOLOv5, we also decided to fine-tune the YOLO v5 model to detect the number of green apples in the dataset citess.YOLOv5 is an algorithm based on single-stage target detection, and the main network structure includes:

Backbone: YOLOv5 uses a lightweight but powerful convolutional neural network architecture: CSPDarknet53 as its backbone network for extracting image features.

Neck: A feature fusion module called PANet is used by the YOLO V5 to better capture target information at different scales and fuse feature maps at different levels.

Detection Head: The detection head section comprises multiple convolutional layers and a final prediction layer, which are responsible for generating bounding boxes and category predictions for the target.

Algorithmically, YOLOv5 employs a target detection method called "anchor-free", which directly predicts the centroid and bounding box of the target without the need for a predefined anchor box. In addition, YOLOv5 uses a structure called "Cross Stage Partial Network" (CSP) to accelerate training and improve model performance.

3.4 EXPERIMENTAL SETUPS

2023APMCM_A_2 and 2023APMCM_A_3 are used as our training set and test set, respectively. These datasets were officially provided by the 2023 APMCM competition.The 2023APMCM_A_2 dataset has been well categorized with fruits and stored in separate folders according to 5 categories: apples, popcorns, pears, plums, and tomatoes.The 2023APMCM_A_3 dataset contains the same total number of images of fruits with the same categories, but all the fruits are mixed in the same folder. The specific information about the number of various fruit categories in the dataset is as follows:

Table 1: Dataset

Dataset	Apple	Carambola	Pear	Plum	Tomato
2023APMCM_A_2	11144	2080	3012	2298	2171
2023APMCM_A_3	-	-	-	-	-

3.5 EXPERIMENTAL RESULTS

To validate the performance of the fruit classifiers, we selected AlexNet, VGG11, VGG16, ResNet18 and ResNet50 as baselines. These are all neural network models based on convolutional algorithms that can be used for image classification tasks.2023APMCM_A_2 is divided into training and validation datasets in the ratio of 8:2. 2023APMCM_A_3 is used as an unlabeled test dataset with a similar total sample size. However, we evaluate the model's metrics based on the validation set metrics size of the first dataset. Our model consists of a total of 5 ConvBlocks, 2 ResidualBlocks, and the FC. The color image is finally cropped to 256 * 256 size, partly because the original image size is 270 * 185, which is relatively close, and partly because it is easier to compute.

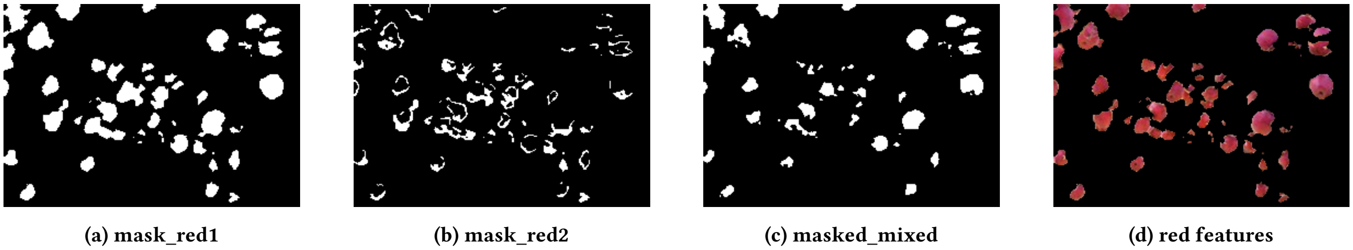
(a) mask_red1 **(b) mask_red2** **(c) masked_mixed** **(d) red features**

Figure 5: Masking operation

Figure 6: Gaussian filtered

Figure 7: Median filtered

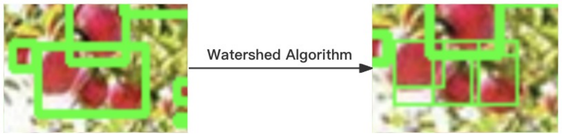

Figure 8: Watershed algorithm effect

The model is trained using the SGD optimizer with a learning rate of 0.0003. The batch size is set to 32, the training ephemeris is set to 100, and the loss function is CrossEntropyLoss.

Table 2 shows how the fruit classifier compares to 5 other baseline models on out datasets. The evaluation metrics we use include loss, accuracy, and macro. It can be seen that our model outperforms the other five models in terms of loss metrics, acc, and macro. On the loss metric, our model is 0.12510, 0.11597, 0.11190, 0.07304, and 0.05678 lower than other models, respectively. On the acc metric, our model is 2.9400%, 2.6892%, 2.5828%, 1.6900%, and 1.4000% higher than other models, respectively. In terms of macro metrics, our model outperforms other models by 0.0290, 0.0272, 0.0195, 0.0137, and 0.0121, respectively. Thanks to our use of batch normalization layers and the appropriate combination of convolutional blocks and residual connection blocks, it can be seen from the excellent performance of these indicators that our model has achieved the best performance.

Table 2: comparison test

Model	Loss	Accuracy	Macro
AlexNet	0.13612	0.9692	0.9603
vgg11	0.12699	0.9717	0.9621
vgg16	0.12292	0.9728	0.9698
ResNet18	0.08406	0.9817	0.9756
ResNet50	0.06780	0.9846	0.9772
Ours	0.01102	0.9986	0.9893

For the final model, we used 100 images extracted each time and repeated 10 times to test the accuracy of the model, and the experimental results show that the comprehensive accuracy of the model is maintained at 92.42% 97.38%. Fig. 10 shows some pictures of the results.

3.6 CONCLUSION

In order to solve the problem of orchard fruit images, the target is heavily obscured by leaves and fruits, which makes it difficult to recognize, we proposes a new lightweight network model (V5FNet) based on OpenCV image feature extraction, watershed algorithm and YOLO V5. The model provides a new lightweight, high-precision fruit classifier for the occlusion phenomenon, the model takes apples as an example and recognizes red apples and green apples respectively. The red features of the image are extracted using filtering, HSV color space transformation, and masking operations, thus eliminating the interference of background and green leaf occlusion. For fruit occlusion, the model incorporates a watershed algorithm to separate the adhering apples. For green apples, we singles out the green apple dataset from its own dataset, and uses this dataset to train a YOLO V5 model to recognize green apples. Eventually, the fruit classifier obtained 99.86% accuracy in the 2023APMCM_A_2 dataset, and the accuracy of the final model was stabilized at 92.42% 97.38%.

We hope that, like Weng et al [16] . we can use AI technology in our daily lives to solve problems for the benefit of mankind . In

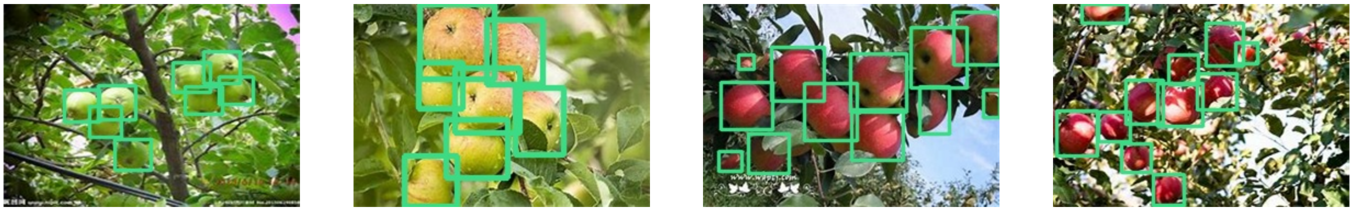

Figure 9: Picture showing the final result of the model (the number of target frames is the number of apples)

the future,this model can be further improved to realize the functions of apple location determination,weight prediction,ripeness classification and so on to further improve the model.

ACKNOWLEDGMENTS

Thanks to the Beijing Society of Image Graphics for providing the dataset.

REFERENCES

[1] Abdullahi, H.S.; Sheriff, R.; Mahieddine, F. Convolution neural network in precision agriculture for plant image recognition and classification. In Proceedings of the IEEE 2017 Seventh International Conference on Innovative Computing Technology (Intech), Porto, Portugal, 12–13 July 2017; pp. 1–3.

[2] Annabel, L.S.P.; Annapoorani, T.; Deepalakshmi, P. Machine Learning for Plant Leaf Disease Detection and Classification–A Review. In Proceedings of the IEEE 2019 International Conference on Communication and Signal Processing (ICCSP), Chennai, India, 4–6 April 2019; pp. 538–542.

[3] Agarwal, M.; Kaliyar, R.K.; Singal, G.; Gupta, S.K. FCNN-LDA: A Faster Convolution Neural Network model for Leaf Disease identification on Apple's leaf dataset. In Proceedings of the IEEE 2019 12th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 18 July 2019; pp. 246–251.

[4] Perez, R.M.; Cheein, F.A.; Rosell-Polo, J.R. Flexible system of multiple RGB-D sensors for measuring and classifying fruits in agri-food Industry. Comput. Electron. Agric. 2017, 139, 231–242.

[5] Rocha, A.; Hauagge, D.C.; Wainer, J.; Goldenstein, S. Automatic fruit and vegetable classification from images. Comput. Electron. Agric. 2010, 70, 96–104.

[6] Rachmawati, E.; Supriana, I.; Khodra, M.L. Toward a new approach in fruit recognition using hybrid RGBD features and fruit hierarchy property. In Proceedings of the 2017 IEEE 4th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), Yogyakarta, Indonesia,19–21 September 2017; pp. 1–6.

[7] Tao, Y.; Zhou, J. Automatic apple recognition based on the fusion of color and 3D feature for robotic fruit picking. Comput. Electron. Agric. 2017, 142, 388–396.

[8] Wang, W.; Siau, K. Artificial intelligence, machine learning, automation, robotics, future of work and future of humanity: A review and research agenda. J. Database Manag. 2019, 30, 61–79.

[9] Samuel, A.L. Some studies in machine learning using the game of checkers. IBM J. Res. Dev. 2000, 44, 206–226.

[10] Liu, W.; Wang, Z.; Liu, X.; Zeng, N.; Liu, Y.; Alsaadi, F.E. A survey of deep neural network architectures and their applications. Neurocomputing 2017, 234, 11–26.

[11] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. Proc. IEEE 1998, 86, 2278–2324.

[12] Goodfellow, I.; Bengio, Y.; Courville, A. Deep Learning; MIT Press: Cambridge, MA, USA, 2016.

[13] LeCun, Y.; Bengio, Y.; Hinton, G. Deep Learning. Nature 2015, 521, 436–444.

[14] Manyu Wang,A new image denoising method based on Gaussian filter,China Three Gorges University,Journal of Image Processing,2014.

[15] Qin-lan Xie. Adaptive Gaussian smoothing filter for image denoising [J].Computer Engineering and Applications. 2009(16)

[16] Weng, W.; Chen, Q.; Dai, Y.; Chen, J.; Chen, D. Multi-scale Fusion Dynamic Graph Neural Network For Traffic Flow Prediction.In Proceedings of the 2023 2nd International Conference on Algorithms, Data Mining, and Information Technology, 85–90, 2023.