ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

Improved Tea Recognition Algorithm Based on RT-DETR

Jianhao Liang[1,*]

[1]School of Mechanical Engineering, Xihua University, Chengdu, 611730, China

[*] Corresponding author:(Email: 1466169996@qq.com)

Abstract

An improved RT-DETR model is proposed to solve the challenges of complex texture details, diverse target scales and high computational efficiency in tea recognition tasks. The convolutional gated linear unit (CGLU) is introduced to achieve dynamic weight adjustment, which enhances the robustness to complex background and variable target. OmniKernel module is designed to integrate multi-direction and multi-scale convolution kernel to improve the modeling ability of tea texture directivity and scale diversity. Combined with frequency-domain feature enhancement module (FSAM), the global context and local detail features are modeled jointly to suppress the interference of background noise. A feature segmentation and fusion module (SPDConv) is proposed to optimize the global consistency of features through space segmentation and channel fusion. Experiments show that the improved model, under the synergistic effect of enhanced input features and multi-scale feature pyramid network, significantly improves the accuracy of tea recognition and adaptability to complex scenes, while maintaining high computational efficiency, providing a high-precision and robust solution for tea detection tasks.

Keywords

RT-DETR model, Tea identification, Object detection.

1. INTRODUCTION

Tea is an important cash crop in the world, and its quality detection and automatic sorting technology is of great significance to enhance the added value of tea industry[1]. Traditional manual sorting relies on experience and is inefficient. The object detection technology based on computer vision provides a new idea for intelligent classification of tea. However, tea detection is faced with unique challenges: 1. Tea has an irregular and flaky structure, and the surface texture is highly directional and has complex details such as folds and curls[2]; 2. Tea stacking and occlusion are common in the processing scene, and the dynamic range of the target scale is large; 3. Production line deployment needs to balance detection accuracy and real-time requirements, and the existing model is difficult to take into account efficiency and complex scenario adaptability[3].

The detection model based on CNN initially realizes the localization of tea targets through multi-scale feature pyramid, but the lack of directional texture modeling results in the serious loss of fine-grained features. Although Transformer architecture can capture the global context, it is difficult to meet real-time detection requirements due to its high computational complexity and insufficient sensitivity to local details[4].. The newly proposed RT-DETR model strikes a balance between speed and accuracy through hybrid architecture design, but it still has limitations in tea detection tasks[5].

To solve the above problems, an improved RT-DETR model is proposed in this paper. The core innovations are as follows: 1) Convolutional gated linear units (CGLU) are designed to enhance

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

the robustness of the model to complex textures and background interference through dynamic weight allocation mechanism[6]; 2) OmniKernel module was constructed and multi-directional gradient convolution kernel was integrated to realize adaptive sensing of tea edge direction; 3) A frequency-domain feature enhancement module (FSAM) is proposed to separate noise components in the frequency domain and enhance local detail characterization in the spatial domain[7]. 4) The feature segmentation and fusion module (SPDConv) is introduced to optimize the global consistency of the feature pyramid through spatial dimension cutting and cross-channel interaction. The block diagram of the model structure is shown in Figure 1.

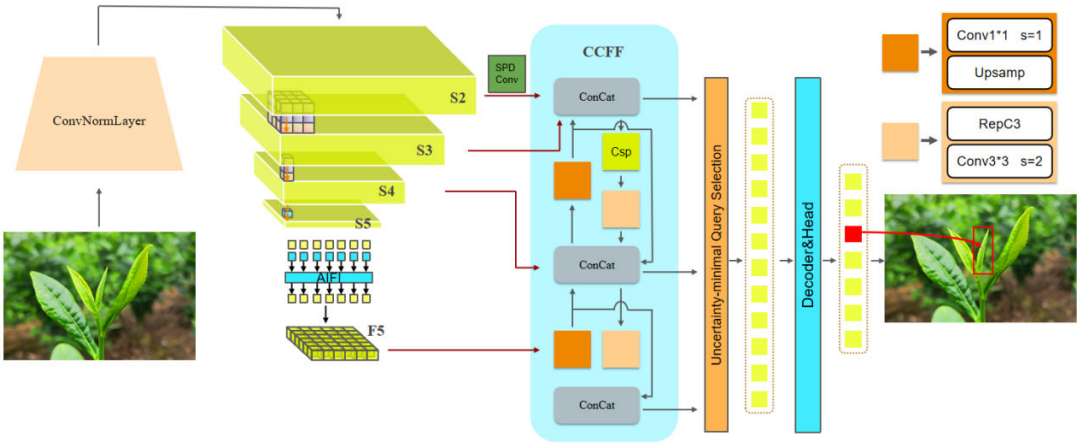


Figure 1. RT-DETR Improved Model Architecture Diagram

2. CORE MODULE DESIGN

Due to the dense texture of the vein of tea tree, large scale variation range and complex background, the target detection and recognition of tea is very challenging, and the expressive power of the model is very high. According to the characteristics of tea images, four basic modules are constructed in the aspects of feature extraction and fusion: BasicBlock_Faster_CGLU module, frequency-spatial domain feature enhancement module FSAM, OmniKernel module and SPDConv feature segmentation and fusion module. These modules can effectively combine global information and local information in feature representation and improve the computational efficiency and adaptability of the model.

2.1. BasicBlock Faster CGLU module

In order to further improve the feature learning ability of residual block, BasicBlock_Faster_CGLU replaces branch 2b in the original residual block with Faster_CGLU, and its structure is shown in Figure 2.Faster_CGLU implements parallel modeling of global channels and local Spaces through partial convolution and dynamic adaptive CGLU mechanism. The two convolution layers of the standard residual block only perform simple linear operations. After adding Faster_CGLU, the network can learn the importance weights between channels to improve the modeling ability of the residual block, which realizes an innovative method for constructing efficient deep networks.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

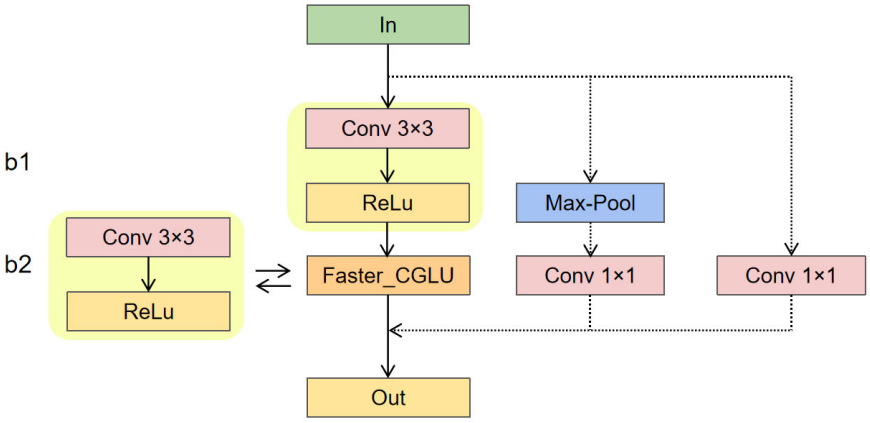


Figure 2. BasicBlock_Faster_CGLU Module

2.2. Design and mathematical analysis of CGLU

Aiming at the global and local feature modeling requirements in tea recognition, this paper designed and introduced the CGLU module, and the structure process is shown in Figure 3.

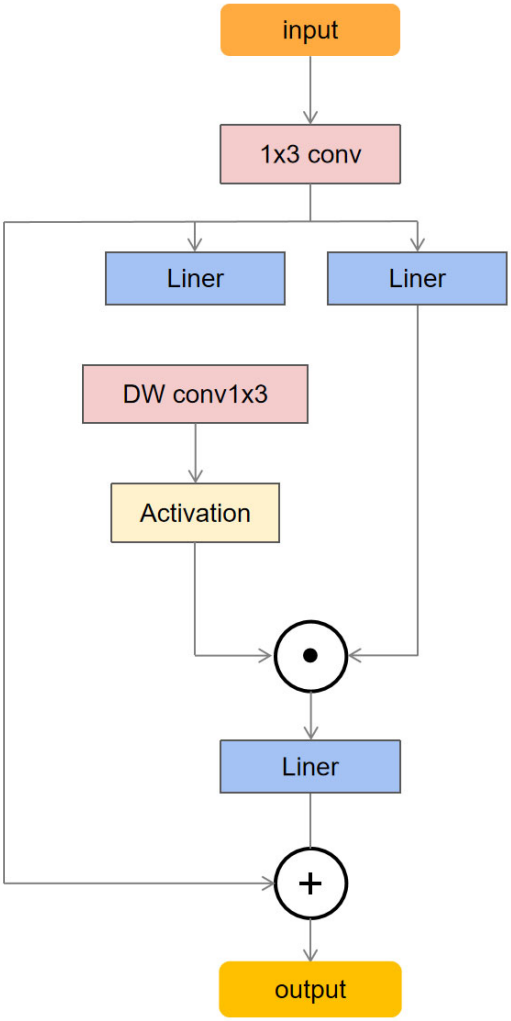


Figure 3. CGLU Module

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

Feature separation is the first step of the CGLU module, dividing input features into two branches for different purposes: the main feature branch can be used for further spatial feature extraction, while the gated branch computes dynamic weights to adjust the weights of features. This architecture not only increases the expressiveness of the network, but also provides a good foundation for the subsequent deep convolution and channel mixing operations. The formula for this process is as follows:

$$X_1, X_2 = Conv1 \times 1(X).chunk(2, dim = 1) \tag{1}$$

Where$chunk(2, \mathrm{dim} = 1)$indicates that the feature is evenly divided into two parts in the channel dimension.In the feature separation operation, 1×1 convolution is introduced to realize the efficient transformation of features through the operation only in the channel dimension, and the number of channels of output features is adjusted by the number of convolution cores:

$$X'[:, i, h, w] = \sum_{j=1}^{C_{in}} X[:, i, h, w] \cdot W_{j,i} + b_i \tag{2}$$

Dynamic feature adjustment realizes the selective enhancement or suppression of feature importance through the adaptive adjustment of dynamic weights. The local feature $X_{spatial}$ of the main feature branch is combined with the dynamic weight X_{gate} element-by-element multiplication to form the dynamically adjusted feature $X_{dynamic}$. This enables CGLU to have powerful dynamic modeling capabilities and adaptively adjust the output according to the content of the input features. The mathematical formula of dynamic feature adjustment is expressed as follows:

$$X_{dynamic} = X_{spatial} \odot X_{gate} \tag{3}$$

The overall operation of CGLU can be summarized in the following formula:

$$\begin{aligned} CGLU(X) = Conv1 \times 1(DWConv(Conv1 \times 1(X)[:, : C_{hidden}]) \\ \odot \sigma(Conv1 \times 1(X)[:, C_{hidden} :])) \end{aligned} \tag{4}$$

2.3. OmniKernel module

In order to further improve the model's ability to capture multi-directional and multi-scale features, the OmniKernel module is introduced in this paper, as shown in Figure 4. OmniKernel module uses multiple convolution operations in different directions and scales to participate in multi-level modeling of input features, which greatly improves the sensitivity of the model to feature details, direction information and texture in tea images.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

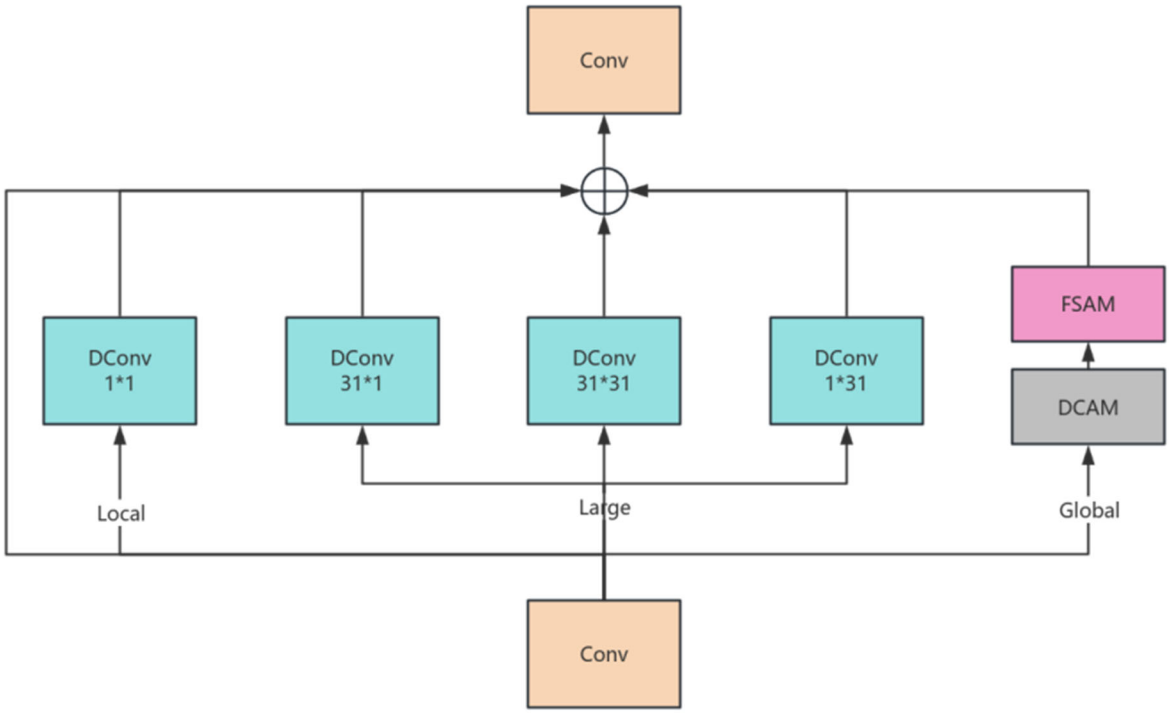


Figure 4. OmniKernel Module

The core idea of OmniKernel module is to use 31×31、31×1、1×31、1×1 convolution kernel shapes to process input features at the same time, and introduce multi-direction and multi-scale receptive fields into feature representation. Given input feature $W_{FCA} \in \mathbb{R}^{C \times H \times W}$, feature compression and nonlinear activation are first performed through a convolution layer:

$$X_{in} = GELU(Conv_{1\times1}(X)) \tag{5}$$

In order to further enhance the nonlinear expression of features, ReLU activation function is added to the module after the final fusion. The OmniKernel module has the following design advantages: large kernel convolution is used to describe global features, strip convolution is used to describe long-range dependencies in some directions, and point convolution is used to describe interactions between channels. Although the module introduces multiple convolution branches, the packet convolution and point convolution architectures reduce the computational complexity well while guaranteeing high performance, combine input features with residual connections, reduce the possibility of feature degradation, and show stronger robustness when dealing with tea complex background and texture noise.

2.4. Frequency - space feature enhancement module

In order to describe the global semantic information and local texture details of tea objects more effectively, a frequency domain - spatial feature enhancement module FSAM was proposed, as shown in Figure 5. The module fuses the information of frequency domain and spatial domain, obtains the global features of frequency domain by Fourier transform, and combines the attention mechanism of frequency domain and spatial domain to realize the adaptive enhancement of multi-scale features and build a more expressive feature representation.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

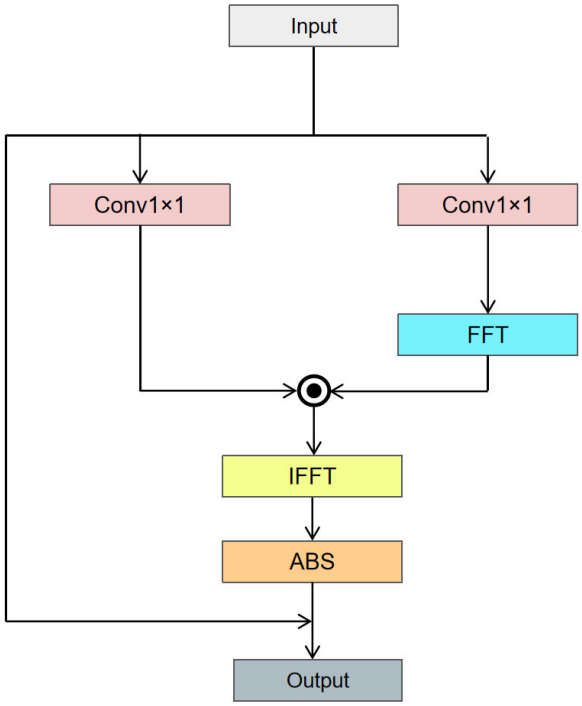


Figure 5. FSAM Module

Given the input feature graph $X \in \mathbb{R}^{C\times H\times W}$, where C is the number of channels, H and W are the height and width of the feature graph respectively, the input feature is compressed by dimension and then mapped to the frequency domain, and the frequency domain feature representation is obtained through the fast Fourier transform:

$$X_{FFT} = FFT(X, norm = 'backward') \tag{6}$$

In order to further improve the distinguishing ability of Frequency domain features, Frequency Channel Attention (FCA) is introduced. In the FCA module, firstly, global average pooling of input features is carried out to extract the global features of each channel:

$$W_{FCA} = Conv_{1\times1}(GAP(X)) \tag{7}$$

Apply attention weights to frequency-domain features:

$$X_{FCA} = X_{FFT} \cdot W_{FCA} \tag{8}$$

Use the inverse Fourier transform IFFT to restore it to the spatial domain:

$$X_{IFFT} = IFFT(X_{FCA}, dim = (-2,-1), norm = 'backward') \tag{9}$$

In order to further optimize the local feature representation of Spatial domain, Spatial Channel Attention (SCA) is introduced. In the SCA module, the input features are firstly

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

dimensionally reduced by global average pooling to extract the global information of the spatial domain:

$$W_{SCA} = Conv_{1\times1}(GAP(X_{IFFT})) \tag{10}$$

Apply attention weights to spatial domain features:

$$X_{SCA} = X_{IFFT} \cdot W_{SCA} \tag{11}$$

By fusing the information of frequency domain and spatial domain, a more expressive feature representation is formed:

$$X_{out} = \alpha \cdot X_{SCA} + \beta \cdot X \tag{12}$$

Compared with traditional spatial feature extraction methods, FSAM module provides global context information for tea recognition process through frequency domain modeling, and spatial modeling enhances local detail features. The combination of the two makes the feature expression more comprehensive. With the help of frequency domain and spatial attention mechanism, the interference effect of background noise on feature learning can be dynamically suppressed, and the accuracy of tea recognition can be improved.

2.5. Feature segmentation and fusion module SPDConv

In order to further improve the optimization of feature expression and the diversification of feature distribution, this paper introduces the feature segmentation and fusion module SPDConv to interweave and separate features from different spatial regions by reducing the spatial dimension of feature maps, extract local information and improve the consistency of global features [8].

The input feature X is segmented along the spatial dimension to generate four sub-feature graphs:

$$\begin{aligned} X_1 = X[\ldots,::2,::2], X_2 = X[\ldots,::2,::2] \\ X_3 = X[\ldots,::2,1::2], X_4 = X[\ldots,1::2,1::2] \end{aligned} \tag{13}$$

Where, $::2$ means sampling at intervals.

In order to establish connections between local features, the four sub-feature graphs obtained by segmentation are spliced in the channel dimension:

$$X_{split} = Concat([X_1, X_2, X_3, X_4], dim = 1) \tag{14}$$

By spatial segmentation of input features, the SPDConv module reduces the spatial dimension of feature graphs by half, while maintaining high feature representation, significantly reducing the computational complexity of subsequent convolution operations and preserving key input information.

This paper proposes to combine RT-DETR model with BasicBlock_Faster_CGLU module, FSAM module, OmniKernel module and SPDConv module to achieve breakthroughs in global

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

and local feature fusion and multi-scale feature modeling, which brings higher accuracy and robustness to tea recognition.

3. EXPERIMENTAL RESULTS AND ANALYSIS

3.1. Experimental Environment

The experiment was conducted on the Windows 10 platform in Python 3.8 programming language, using the deep learning framework PyTorch for object detection tasks. PyTorch with dynamic computation graphs, automatic differentiation, and good GPU acceleration is the most widely used framework in modern deep learning. In terms of hardware configuration, this paper uses the 13th generation Intel Core i9-13900HX CPU, NVIDIA GeForce RTX 4080 Laptop GPU 12 GB video memory graphics card, CUDA version 12.0, which provides powerful computing support for model training and reasoning. In terms of software environment, this paper uses PyCharm as integrated development environment for coding and debugging. All experiments of the comparison model were carried out under the following computing platform and environment configuration, the specific experimental Settings are shown in Table 1:

Table 1. Experimental Environment

Index	Parameter
Operating system	Windows10
CPU	13th Gen Intel(R) Core(TM) i9-13900HX
GPU	NVIDIA GeForce RTX 4080 Laptop GPU
CUDA	12.0
Video memory	12G
Training framework	PyTorch
Programming language	Python3.8
Compiling environment	Pycharm

3.2. Comparative result analysis

In this experiment, multiple target detection models were compared in tea detection tasks, namely, YOLOv5, YOLOv8, YOLOv11, RT-DETR-R18, RT-DETR-R50 and the improved model, RT-DETR-SOEP-Faster-CGLU. The experimental indexes are accuracy rate, recall rate, mAP50 (%), MAP50-95 (%), and the comparison results are shown in Table 2:

Table 2. Comparative Experiment Results

Model	Precision	Recall	mAP50(%)	mAP50-95(%)
Yolov5	0.815	0.78	0.781	0.413
Yolov8	0.836	0.807	0.812	0.445
Yolov11	0.823	0.798	0.783	0.429
RTDETR-R18	0.827	0.842	0.843	0.478
RTDETR-R50	0.837	0.847	0.854	0.482
RTDETR-SOEP-Faster-CGLU	0.856	0.850	0.871	0.504

The experimental results show that the improved RT-Dettr-SOEP-faster-CGLU model has better advantages than other models in tea target detection tasks, and the evaluation indexes such as accuracy, recall rate and mAP have been greatly improved, indicating that the model can

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

efficiently detect tea targets. In the case of complex background environment and uneven shape of tea leaves, the target can still be detected and recognized, and the overall performance of the model can be improved.

In order to visually demonstrate the performance of each model on the tea recognition task, we visually compared the mAP(%) detection performance of different models on the self-built tea test set, and the relevant results are shown in Figure 6.

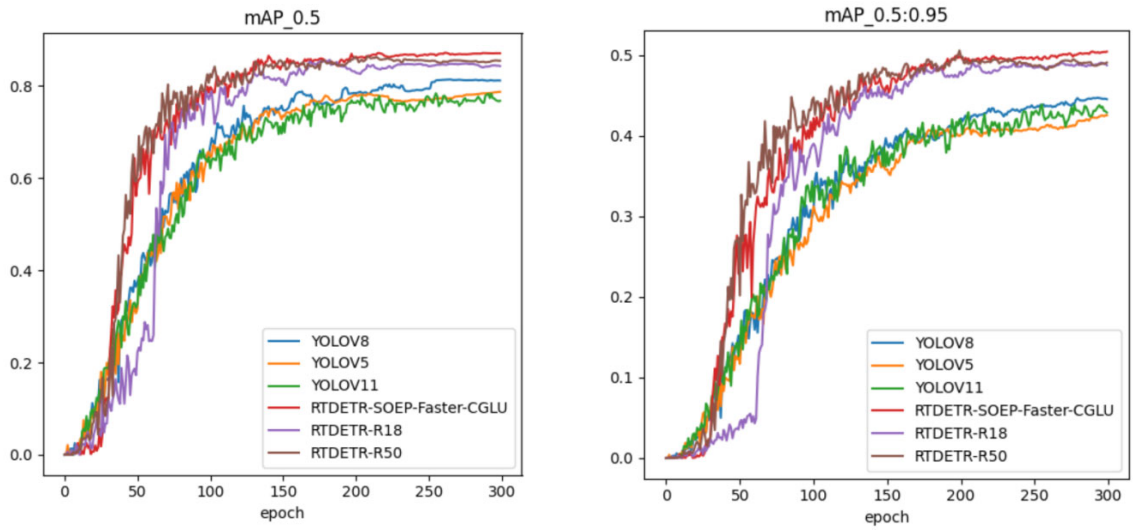


Figure 6. Comparison of Tea Leaf Recognition mAP(%) for Different Models

4. CONCLUSION

In this paper, an improved RT-DETR model is proposed to solve the problems of complex background, compact texture information and large scale variability of tea images. By adding BasicBlock_Faster_CGLU module, frequency-space domain feature enhancement module FSAM, OmniKernel and SPDConv module to the basic model, the feature expression ability is enhanced.The experimental results show that compared with the traditional YOLO series model and RT-DETR, the improved RT-DETR-SOEP-faster-CGLU significantly improves the recognition accuracy, making the model more stable and reliable in solving difficult tea detection tasks, and providing a guarantee for improving the accuracy and performance of tea target detection.

REFERENCES

[1] J.W. Zhao, Q.S. Chen, X.Y. Huang, C.H. Fang, Qualitative identification of tea categories by near infrared spectroscopy and support vector machine, J. Pharm. Biomed. Anal. 41(4) (2006) 1198-1204. https://doi.org/10.1016/j.jpba.2006.02.053.

[2] J. Zhou, H. Cheng, J.M. Zeng, L.Y. Wang, K. Wei, W. He, W.F. Wang, X. Liu, Study on Identification and Traceability of Tea Material Cultivar by Combined Analysis of Multi-Partial Least Squares Models Based on Near Infrared Spectroscopy, Spectrosc. Spectr. Anal. 30(10) (2010) 2650-2653. https://doi.org/10.3964/j.issn.1000-0593(2010)10-2650-04.

[3] W. Jian, Z. Xianyin, D. ShiPing, IDENTIFICATION AND GRADING OF TEA USING COMPUTER VISION, Appl. Eng. Agric. 26(4) (2010) 639-645.

[4] C. Zhang, J. Wang, G.D. Lu, S.M. Fei, T. Zheng, B.C. Huang, Automated tea quality identification based on deep convolutional neural networks and transfer learning, J. Food Process Eng. 46(4) (2023) 16. https://doi.org/10.1111/jfpe.14303.

ISSN: 2472-3703 DOI: 10.6911/WSRJ.202504_11(4).0017

[5] C.Y. Yan, Z.H. Chen, Z.L. Li, R.X. Liu, Y.X. Li, H. Xiao, P. Lu, B.L. Xie, Tea Sprout Picking Point Identification Based on Improved DeepLabV3+, Agriculture-Basel 12(10) (2022) 15. https://doi.org/10.3390/agriculture12101594.

[6] D.J.I. Shi, TransNeXt: Robust Foveal Visual Perception for Vision Transformers, (2023).

[7] Y.N. Cui, W.Q. Ren, A. Knoll, Omni-Kernel Modulation for Universal Image Restoration, IEEE Trans. Circuits Syst. Video Technol. 34(12) (2024) 12496-12509. https://doi.org/10.1109/tcsvt.2024.3429557.

[8] C.Y. Wang, A. Bochkovskiy, H.Y.M.J.a.e.-p. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, (2022).