Light Weight Apple Defect Detection by Gaussian Mixture Model and Attention Mechanism

Xiubo Ma* XIUBOMA@163.COM
Anhui Sanlian University, Hefei, 230601, China

Xiongwei Sun
Anhui Institute of Optics and Fine Mechanics, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei, 230031, China

Shaoqing Shi
Hefei Cigarette Factory of Anhui Zhongyan Industry Co., Ltd., Hefei, 230081, China

Editors: Nianyin Zeng and Ram Bilas Pachori

Abstract

To improve the speed and accuracy of online apple defect detection, a modified PP-PicoDet model based on mixed Gaussian color modeling and spatial attention mechanism is proposed. Firstly, in order to enhance the detection performance of small and low-contrast defects, the method constructs an offline color model based on GMM theory and builds a saliency numerical channel that integrates the saliency of anomalous targets on RGB data. Secondly, the SE mechanism of the network is optimized, integrating the CBAM (convolution block attention module) structure to enhance the perception of structural features in anomalous regions through strengthened spatial attention mechanisms. Furthermore, the abnormal loss function is adjusted by employing a Gaussian model to establish a label assignment loss function strategy, simplifying parameter tuning and improving the model's motivation for small proportion defects. This enhancement aims to improve the detection performance of the model under conditions of imbalanced samples. Experiments indicate that the algorithm, implemented on a self-built application platform using real-world datasets, achieved a 5.6% increase in accuracy and a 5.8% increase in recall at the cost of only 0.5% time delay. The algorithm meets the requirements of timeliness and reliability for detecting defects online, contributing to the enhancement of efficiency in apple quality grading and production.

Keywords: Defect detection, PP-PicoDet, GMM, Attention mechanism

1. Introduction

Apple is an important crop in China, with its yield and planting area ranking first in the world. Surface defect detection is an important part of apple quality inspection, but currently, the detection method still relies mainly on manual visual inspection, which is costly and inefficient, making it difficult to apply to large-scale production. Therefore, achieving automation and high-precision detection of apple surface defects is of great practical significance for ensuring apple quality and improving enterprise profits.

With the continuous promotion and application of deep learning theory in fruit detection, compared to traditional image processing algorithms, deep learning models do not need to design complex feature extractors for the detected object. They have important performance advantages in practical detection applications, attracting researchers to continuously introduce new network model structures into fruit defect detection applications. XUE et al. (2020) proposed using the Google Net model and using transfer learning for training to achieve apple surface defect detection. Wang et al. (2022) addressed the issue of overly complex existing plant disease identification models and used the Ghost module to lightweight improve VGG16. ZHANG et al. (2022) used a combination of GhostNet and transfer learning to improve the accuracy of network classification by using a Dropout layer. With the continuous development of object detection models, classic object detection frameworks with better performance such as R-CNN, SSD, YOLO (Redmon and Farhadi, 2018; Kaur and Singh, 2023) are constantly being introduced. However, online detection models often pursue speed improvement in online detection applications, leading to a significant performance decline in model recognition accuracy when faced with the diversity of surface defects in natural products such as apples. This problem is a common problem in online agricultural product testing, and the key to solving this problem is to balance the speed and accuracy of lightweight defect detection models.

In summary, classic lightweight deep learning networks need to further enhance the expression ability of feature extraction networks for various types of defects in order to improve the accuracy of target defect detection while maintaining detection speed. Therefore, this article takes apple defect detection as an example and proposes a visual attention mechanism based on offline GMM (WANG and ZHAO, 2014) color modeling using the PP-PicoDet (Yu et al., 2021) lightweight detection model. By integrating the network spatial attention mechanism, the detection performance of the lightweight defect detection network is optimized on small and weak contrast targets, improving the stable performance of the lightweight model when facing diverse defects.

2. Target saliency detection based on GMM

2.1. Effective color background modeling

Assuming the distribution of colors on the surface of healthy apples follows a certain probability distribution pattern, a probabilistic statistical model is used to fit the color sample space by approximating the distribution of data points using a weighted sum of multiple Gaussian distributions. GMM (WANG and ZHAO, 2014) (Gaussian Mixture Model) is one type of finite probability mixture model, which enhances the model's ability to express complex distributions by utilizing a linear combination of a finite number of independent Gaussian probability models. The GMM model is proposed to fit the distribution of complex apple surface color data, capturing the color distribution characteristics of normal areas on the apple surface for pre-screening color abnormal regions, as shown in Figure 1. An offline approach is employed to construct a color probability distribution field and establish significant features of target colors, enabling enhanced expression of features for abnormal defect targets while ensuring speed is not compromised.

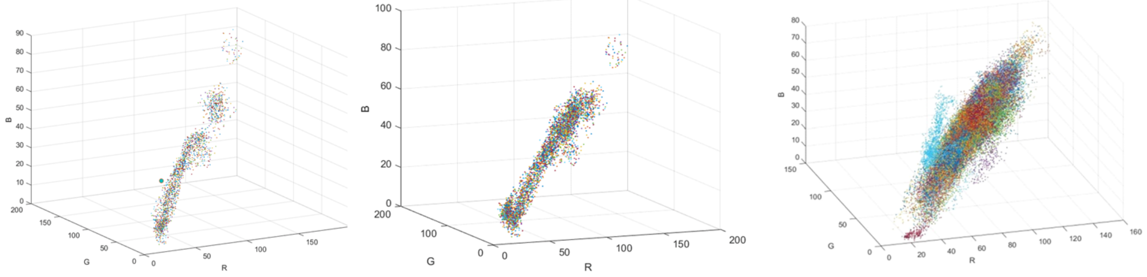


Figure 1: Accumulate color samples until the color feature space pattern stabilizes.

Assuming the entire color distribution can be represented by K Gaussian models, then the Gaussian probability density function for any color sample vector X belonging to the i-th color category can be expressed as:

$$\phi(x;\mu,\sum) = \frac{1}{(2\pi)^{(N/2)|\sum|^{1/2}}} e^{-\frac{(x-\mu)^T \sum^{-1}(x-\mu)}{2}} \tag{1}$$

where μ is the mean color vector for the case, Σ is the covariance matrix. $(x-\mu)^T$ denotes transpose, $|\Sigma|$ and Σ^{-1} represent determinant and inverse matrix operations. The Gaussian mixture color model can be considered as the probability distribution of color sample vectors belonging to K categories of colors. The joint probability density of the Gaussian mixture model is given by:

$$\begin{cases} \rho(X|\theta) = \sum\limits_{k=1}^{K} \alpha_k \phi(x;\theta_k) \\ \sum\limits_{k=1}^{K} \alpha_k = 1, a_k \geq 0, k = 1, ..., K \end{cases} \tag{2}$$

where α_k represents the weights of the mixture components, θ represents the parameter set of each sub-model in GMM, and θ_k represents the parameters of the k-th sub-model. Utilizing the Bayesian Information Criterion (Dellaert, 2000) (BIC) to strike a balance between the fitting ability and complexity of the model, the number of components in models like the Gaussian Mixture Model (GMM) is determined. The formula for model calculation is as described in Equation 3.

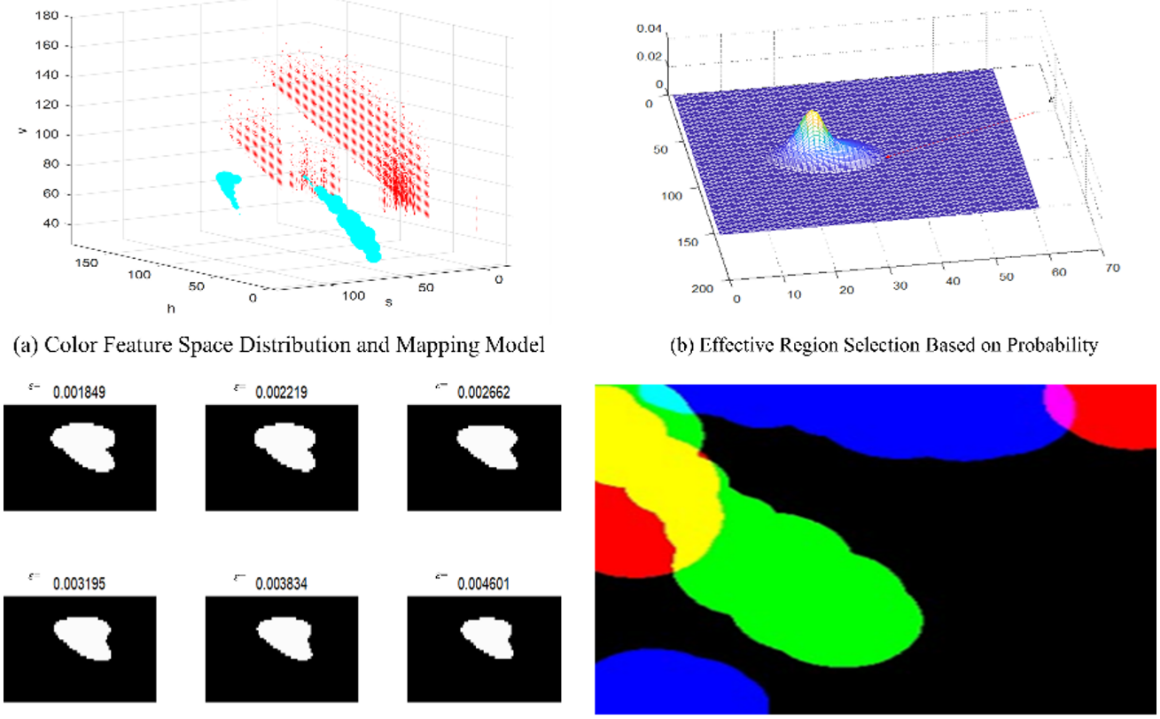
$$BIC = K\ln(n) - 2\ln(\hat{L}) \tag{3}$$

where $\hat{L}$ represents the maximum likelihood estimation of the fitting function, specifically maximizing the likelihood function to obtain the maximum likelihood probability value, K is the number of fitted sub-models, n is the actual number of sampled data points. A smaller BIC value indicates a better model fit. Under the condition of setting the parameter K, the Expectation-Maximization (Woo et al., 2018) (EM) algorithm is used for iterative estimation of parameters for each sub-model.

2.2. Generation of attention distribution field

To generate the attention distribution field for healthy apple surface colors, we map the three-dimensional color feature space onto three axes: *H* (Hue), *S* (Saturation), and *V* (Value). Through three sets of two-dimensional spatial distributions, we simplify the expression of color space distribution patterns. As shown in Figure 2, in the *H-S*, *S-V*, and *H-V* mapping spaces, we further enhance color feature stability. Here, we introduce empirical threshold adjustments to regulate the effective color feature confidence. Areas with low confidence are discarded to exchange for a more stable probability expression in the remaining regions.

As shown in Figure 2, the color distribution ranges in the *H-S*, *H-V*, and *V-S* planes are represented using a mapping image of size [256*256*3]. Each time a suspected surface defect area is extracted, the presence of the area in the effective color mapping image is evaluated pixel by pixel. This process enables the statistical calculation of color confidence. The statistical results of confidence are then used to generate the target attention information channel, denoted as T, which facilitates the inference data of the fusion attention mechanism in *RGB-T* form.



(a) Color Feature Space Distribution and Mapping Model

(b) Effective Region Selection Based on Probability

(c) Color Effective Regions Based on Empirical Thresholds

(d) Color Feature 2D Mapping Distribution

Figure 2: Color mapping table generation process based on Gaussian Mixture.

3. Lightweight model integrating attention mechanism

3.1. Spatial feature fusion

The PP-PICO network utilizes the ESNet to extract target features, where the SE (Hu et al., 2018) module adjusts the weights of different channels to enhance features channel-wise. As surface defects on apples often exhibit large size distributions and rich texture distributions on surface, enhancing the spatial texture information of features can improve the network's ability to represent features effectively. Here, the CBAM (Woo et al., 2018) module, which integrates spatial and channel attention mechanisms, is introduced into the feature extraction module of the network, replacing the ES Block network structure, as shown in Figure 3. This strengthens the network's sensitivity to texture differences and enhances the model's detection performance for surface defects on apples.

The channel attention module of the CBAM structure is utilized to spatially optimize the abnormal color attention distribution channels in the *RGB-T* data, obtaining channel weights. Specifically, average pooling reinforces the structural information of regions with significant colors in the T channel, while max pooling is used to characterize the impact of channels on the overall detection when the color specificity of the detected target is insufficient. The average pooling factor *AvgPool(F)* and maximum pooling factor *MaxPool(F)* are connected to a shared two-layer perceptrons, generating channel attention weights *Mc(F)*. The first layer is set to C/r neurons, where r is the compression ratio, and the number of neurons in the second layer is C. Then multiply the coefficients with the

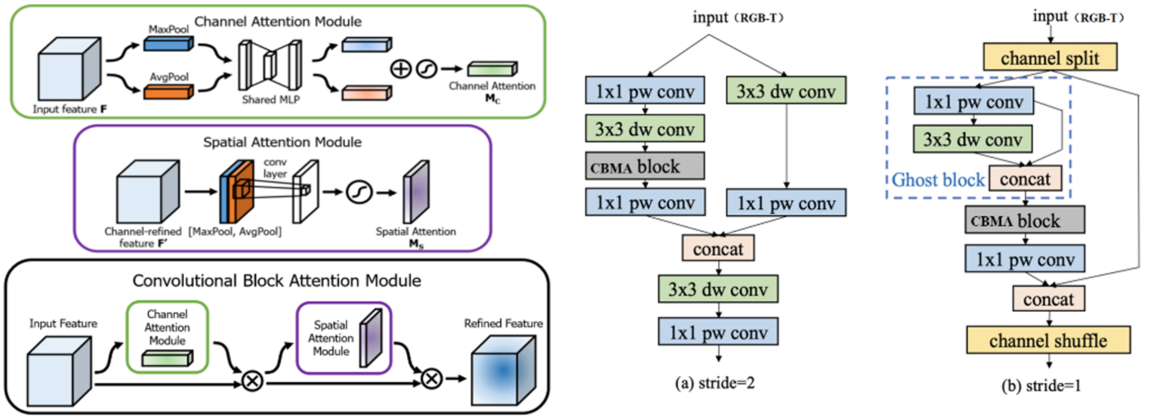


Figure 3: Spatial Attention-Enhanced Network Structure.

original feature F to form the output features.

$$\begin{aligned} \mathrm{M}_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{\max})) \end{aligned} \tag{4}$$

The spatial attention mechanism describes the spatial distribution aggregation of features and complements the channel weights. The feature extraction network utilizes the connection between average pooling and maximum pooling to obtain feature descriptors, and generates spatial attention maps through a convolutional layer $Ms(F) \in C \times H \times W$. The spatial weights are formed through convolutional structure encoding, as follows:

$$\begin{aligned} M_s(F) &= \sigma(f^{7\times7}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{7\times7}([F^c_{avg}; F^c_{\max}])) \end{aligned} \tag{5}$$

where σ is the sigmoid function and $f^{7\times7}$ is the empirical value which implemented as a large kernel convolution. As shown in Figure 3, the network utilizes a connection through average pooling and max pooling along the channels to obtain feature descriptions. A convolutional layer is used to generate spatial attention maps.

3.2. Loss Functions

In the PicoDet network model, SimOTA utilizes a combination of weighted VFLoss (Lin et al., 2020) and GIoUloss (Tian et al., 2019) to construct the loss matrix. While VFLoss (Lin et al., 2020) addresses dense object detection problems using an asymmetric processing structure, it does not offer intuitive advantages in practical parameter tuning. Hence, this paper selects the symmetrically structured FocalLoss loss in combination with GIoUloss (Tian et al., 2019) to construct the loss matrix.

$$\begin{cases} cost = loss_{fl} + \lambda loss_{giou} \\ loss_{fl} = -\alpha_t(1-\rho_t)^\gamma \log(\rho_t) \end{cases} \tag{6}$$

The Focal Loss is used to adjust the difficulty of sample classification, giving higher weight to hard-to-classify samples and lower weight to easy samples in the loss function. The classical

Focal Loss function utilizes hyperparameters to control the degree of loss attenuation, with a larger parameter leading to more significant loss attenuation.

However, the contrast between the weights of easy and hard cases in the exponential function parameter tuning is not intuitive. Here, a probability density function constructed using a standard Gaussian function is used to describe the distribution of easy and hard samples. As shown in Equation 7, the Gaussian function is employed to adjust the degree of weight attenuation for easy and hard samples based on probability indicators.

$$\mathrm{loss}_{fl} = \alpha_t \frac{1}{\sqrt{2\pi}\sigma} e^{-(p_t)^2/(2\sigma^2)} \tag{7}$$

Assuming that the distribution of positive and negative samples follows an independent and identically distributed model, when computing the cross-entropy loss, the imbalance of small sample cases in apple defect samples is adjusted by increasing the balance parameter α_{t}. Adjusting the parameters using the Gaussian function provides a smoother model. By utilizing a standard probability distribution model and combining it with sample proportions, the relative distribution ratio of easy-to-separate areas to difficult-to-separate areas is directly calculated. This yields modulation results for comprehensive probability model parameters. Empirical parameters are obtained through specific experiments: $\lambda = 3.0$ $\sigma = 0.25$.

4. Experiments

4.1. Dataset and evaluation criteria

4.1.1. DATASET

The experiment utilized Fuji apples, with a total of 3000 apple image samples collected using a Hikvision MVCA013-A0UC industrial camera. The resolution of the target fruits was 540×540. The defects were annotated into 7 categories using Labelimg annotation software. The dataset was randomly divided into training, validation, and testing sets in a ratio of 7:2:1. The distribution of data labels and categories is shown in Table 1. From Table 1, it can be observed that the distribution of cases in each category is uneven. During specific model training, samples from each category need to be proportionally selected for model training.

Table 1: Categories of apple defects.

Code	Defect Category	Number of Instances	Number of Labels
1	Disease	1002	1756
2	Stem Cuts	805	964
3	Rust Spots	604	814
4	Blemishes	403	814
5	Insect Damage	146	160
6	Bruising	1460	1708
7	Black Spots	873	1068

4.1.2. EVALUATION CRITERIA

In classical object detection tasks, the mean values of evaluation metrics such as mAP@0.5 and mAP@0.5:0.95 are used to evaluate detection performance. The mAP is computed as shown in Equation 8. FPS represents the number of images detected per second and serves as a standard measure of detection performance.

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \\ AP = \int \genfrac{}{}{0pt}{}{1}{0} P(R)dR \\ \text{m}AP@0.5 = \frac{1}{N} \sum_{k=1}^{N} AP@0.5_k \\ \text{m}AP@0.5 : 0.95 = \frac{1}{10}(mAP@0.5 + mAP@0.55 + ... + mAP@0.95) \end{cases} \tag{8}$$

In the equation, *TP* represents the number of detection boxes with an Intersection over Union *(IOU)* greater than or equal to the specified threshold, *FP* represents the number of detection boxes with an *IOU* less than the specified threshold, and *FN* represents the number of missed targets.

4.2. Analysis of ablation experiment results

The training environment for the experiment consists of a system with 16 cores Intel(R) Core^{TM} i7-12700F CPU @ 2.60GHz, 32GB of RAM, and an NVIDIA GeForce RTX 3070 GPU with 32GB of memory. The coding environment utilizes Python 3.7.4 and PaddlePaddle V2.1. The learning rate for the PP-PicoDet model is uniformly set to 0.0057. The image input size is set to 540x540, and the batch size is set to 4. Regarding model parameter training, the algorithm is set to run for 100 epochs. Every 10 epochs, a validation is performed, and the model with the best performance on the validation set is saved at the end.

To validate the effectiveness of the modifications, the optimization components of the model will be systematically combined here to test the actual effects of different modifications. The default PP-PICODDet model is based on the ESNet_m (Yu et al., 2021) framework. Four optimization strategies will be implemented on this framework: ① The base network is enhanced with a color attention mechanism RGB-T. ② The base network is enhanced with a CBAM attention mechanism. ③ The base network is optimized with a modified loss function. ④ The base network is augmented with data preprocessing before augmentation.

In the experiment, the detection results of the 7 defect categories are averaged for evaluation. The effects of different optimization strategies on the model are compared, and metrics such as accuracy and recall are calculated, as shown in Table 2.

4.3. Comparison and evaluation of different methods

Comparing the performance of the model with the dataset used in this paper against classical online lightweight object detection network models, where the learning rates for YOLOv3 (Redmon and Farhadi, 2018), YOLOv5s (Kaur and Singh, 2023), and YOLOv5n (Kaur and Singh, 2023) are uniformly set to 0.0052. The number of iterations is set to 100, and the batch size is set to 4. The comparative experimental results are shown in Table 3.

Table 2: Performance Comparison of Model's Optimization.

Model	P	R	mAP@0.5	FPS
PP-PICO	0.871	0.896	0.925	54.6
① PP-PICO+RGBT	0.924	0.929	0.973	53.5
② PP-PICO+CBAM	0.897	0.941	0.935	53.8
③ PP-PICO+ Optimization of the loss function	0.886	0.912	0.928	54.5
④ PP-PICO+ Data preprocessing	0.887	0.903	0.921	54.6
PP-PICO+①+②+③+④	0.935	0.928	0.976	52.5

Table 3: The performance comparison between the proposed method and classical object detection networks.

Model	P	R	mAP@0.5	FPS
YOLOv5s	0.931	0.897	0.873	50.1
YOLOv5n	0.851	0.858	0.704	67.6
YOLOv3	0.837	0.886	0.914	68.3
PP-PICO_s	0.821	0.813	0.764	50.9
PP-PICO_l	0.902	0.902	0.985	58.8
PP-PICO_s +①+②+③+④	0.864	0.869	0.904	56.4
PP-PICO_m +①+②+③+④	0.934	0.947	0.920	52.5
PP-PICO_l+①+②+③+④	0.958	0.960	0.987	50.1

From Table 3, it can be observed that compared to classical lightweight network models, the method proposed in this paper achieves a unified improvement in the detection accuracy and recall of apple defects through color modeling and complementary spatial feature saliency enhancement. Although the improved model's detection speed does not reach its optimum, with additional time consumption controlled within 5%, it still satisfactorily meets the time requirements of the detection task.

5. Conclusion

This paper proposes a rapid modification model based on the lightweight network framework PP-PicoDet. Firstly, a color attention mechanism based on offline Gaussian Mixture Model (GMM) is constructed to build saliency channel mechanisms. This enhances the detection effectiveness of small and weak defect features on apple surfaces using the RGB-T data mode under limited resources. Secondly, in conjunction with the saliency channel mechanism, the CBAM spatial attention network structure is incorporated to strengthen the spatial features of small-sized and weak-textured targets. Lastly, considering the limited number and uneven distribution of defect samples in the measured dataset, a weighted label loss reset strategy based on Gaussian functions is introduced to improve the efficiency of parameter weight modulation under uneven sample conditions, further enhancing detection stability. Experimental results demonstrate that the proposed method is suitable for rapid detection of complex defect morphologies on apple surfaces. It effectively optimizes the detection performance of various and complex defects, exhibiting good stability and accuracy. This method holds practical significance for improving online apple defect detection applications.

Acknowledgments

This study was supported by the Natural Science Research Project of Anhui Educational Committee (grant number: 2022AH051988, 2022AH051996) , the Foundation for Distinguished Young Talents in Higher Education of Anhui (grant number: YQYB2023077) and the Quality Engineering Project of Anhui(grant number: 2022jyxm491).

References

Frank Dellaert. *The Expectation-Maximization Algorithm*, pages 387–390. Springer New York, New York, NY, 2000. doi: 10.1007/978-1-4419-0300-6_22.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. doi: 10.1109/CVPR.2018.00745.

Ravpreet Kaur and Sarbjeet Singh. A comprehensive review of object detection with deep learning. *Digital Signal Processing*, 132:103812, 2023. ISSN 1051-2004. doi: 10.1016/j.dsp.2022.103812.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020. doi: 10.1109/TPAMI.2018.2858826.

Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement, 2018.

Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9626–9635, 2019. doi: 10.1109/ICCV.2019.00972.

JQ Wang, X Ji, and HF Mo. Plant disease detection based on lightweight vgg. *Journal of Chinese Agricultural Mechanization*, 43(04):25–31, 2022. doi: 10.13733/j.jcam.issn.2095-5553.2022.04.005.

Siming WANG and Wei ZHAO. Brightness feature autocorrelation and gmm combined target detection. *Computer Engineering*, (5):219–223, 1 2014.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 3–19, Cham, 2018. Springer International Publishing.

Yong XUE, Liyang WANG, Yu ZHANG, and Qun SHEN. Defect detection method of apples based on googlenet deep transfer learning. *Transactions of the Chinese Society for Agricultural Machinery*, 51(7):30, 2020. doi: 10.6041/j.issn.1000-1298.2020.07.004.

Guanghua Yu, Qinyao Chang, Wenyu Lv, Chang Xu, Cheng Cui, Wei Ji, Qingqing Dang, Kaipeng Deng, Guanzhong Wang, Yuning Du, Baohua Lai, Qiwen Liu, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. Pp-picodet: A better real-time object detector on mobile devices, 2021.

Qianru ZHANG, Yunfei WANG, Shuaichao LV, Lei SONG, Yuying SHANG, and Huaibo SONG. Integrity classification of wheat straw epidermis based on improved ghostnet. *Journal of Nanjing Agricultural University*, (004):045, 2022.