Hindawi
Journal of Control Science and Engineering
Volume 2021, Article ID 6614672, 11 pages
https://doi.org/10.1155/2021/6614672

*Research Article*

# Fresh Tea Sprouts Detection via Image Enhancement and Fusion SSD

**Bin Chen, Jili Yan, and Ke Wang**

*College of Mathematics Physics and Information Engineering, Jiaxing University, Jiaxing 314000, China*

Correspondence should be addressed to Bin Chen; chenbin@zjxu.edu.cn

Received 9 December 2020; Revised 13 April 2021; Accepted 19 April 2021; Published 26 April 2021

Academic Editor: Radek Matušů



The accuracy of Fresh Tea Sprouts Detection (FTSD) is not high enough, which has become a big bottleneck in the field of vision-based automatic tea picking technology. In order to improve the detection performance, we rethink the process of FTSD. Meanwhile, motivated by the multispectral image processing, we find that more input information can lead to a better detection result. With this in mind, a novel Fresh Tea Sprouts Detection method via Image Enhancement and Fusion Single-Shot Detector (FTSD-IEFSSD) is proposed in this paper. Firstly, we obtain an enhanced image via RGB-channel-transform-based image enhancement algorithm, which uses the original fresh tea sprouts color image as the input. The enhanced image can provide more input information, where the contrast in the fresh tea sprouts area is increased and the background area is decreased. Then, the enhanced image and color image is used in the detection subnetwork with the backbone of ResNet50 separately. We also use the multilayer semantic fusion and scores fusion to further improve the detection accuracy. The strategy of tea sprouts shape-based default boxes is also included during the training. The experimental results show that the proposed method has a better performance on FTSD than the state-of-the-art methods.

## 1. Introduction

Automatic tea picking by machine is an effective way to solve the tea picking labor problem. However, the existing vision-based tea picking robots cannot meet the requirement of a high-quality tea picking task due to the poor fresh tea sprouts detection and the uncontrollable imaging condition [1–3]. Tang et al. [4] proposed a Local Binary Pattern (LBP) and Gray-Level Co-Occurrence-Matrix- (GLCM-) based method for green tea leaves classification. A fresh tea leaves quantity assignment method by estimating total polyphenols using near infrared spectroscopy is introduced in [5]. Unfortunately, most existing studies focused on tea classification and quantity assignment [6, 7], and the imaging condition is limited to indoor or laboratory [8]. Therefore, the fresh tea sprouts detection in outdoor condition is an urgent problem needed to be solved.

Vision-based fresh tea sprouts detection belongs to the category of object detection [9, 10]. In recent years, object detection has developed quickly with the aid of Convolutional Neural Networks (CNNs). Generally, the CNN-based object detectors can be divided into two types: a one-stage object detector and two-stage object detector. The one-stage object detector is usually modeled as a simple regression problem and encapsulated all the computation in a single feed-forward CNN [11], which can effectively increase the detection speed. The two-stage object detector is mainly based on the proposal-driven stage and detection refining stage. Detection speed is the key advantage of the one-stage-based methods, but the accuracy is insufficient for detecting small objects. Especially in the fresh tea sprouts detection task, the tea sprouts are much smaller than the ripe leaves, and it is hard to detect the tea sprouts quickly and efficiently via the existing methods. Moreover, the requirement of detection speed and accuracy is high when the object detector is used in the vision-based tea picking robots.

Most existing research studies focused on how to extract more robust features, such as the improvement of the feature extraction network. In this paper, we rethink the process of fresh tea sprouts detection and aim to propose a new

improvement strategy for object detection. Motivated by the multispectral image processing, we find that more input information can lead to a better detection result. With this in mind, we propose a novel Fresh Tea Sprouts Detection method via Image Enhancement and Fusion SSD (FTSD-IEFSSD). The main contributions of our paper are as follows:

(1) We propose a novel improvement strategy for object detection using input information increase via image enhancement. The added input information obtained by the image enhancement usually contains more salient features, which help to improve the accuracy of object detection. In the task of FTSD, we observed that the gray-level features in the enhanced image have a significant difference between the fresh tea sprouts area and the background area. Therefore, the FTSD performance can be improved efficiently via the proposed method.

(2) An RGB-channel-transform-based image enhancement algorithm is proposed to increase the input information, and the parameters used in the transform algorithm are analyzed in the experimental section. In order to further improve the detection performance, we use the multilayer semantic fusion and adaptive scores fusion to enhance the feature maps. The designed semantic fusion part consists of four convolutional layers with different semantic depths, and the balance of detection speed and accuracy is also considered in our algorithm.

(3) To the best of our knowledge, the proposed FTSD-IEFSSD is the first using image enhancement-based input information increase, multilayer semantic fusion, and adaptive scores fusion for fresh tea sprouts detection with outdoor imaging condition. Our method is an end-to-end object detection method; the input is the original color image directly obtained from the outdoor, and the output is the detection results. It is a challenging task due to the uncontrollable imaging condition and the high requirement of detection speed and accuracy.

The remainder of the paper is organized as follows: the related works about tea sprouts detection and the CNN-based object detectors are presented in Section 2. Section 3 describes the conventional SSD [12] and the proposed FTSD-IEFSSD, and the proposed RGB-channel-transform-based image enhancement algorithm and semantic fusion are also included. The experimental results and tea sprouts detection performance are shown in Section 4. Section 5 gives the conclusions of our work.

## 2. Related Works

According to different emphases on the prior knowledge, most existing tea sprouts detection methods can be categorized into two groups, the manually designed feature-based method and the CNN-based method.

(i) The manually designed feature-based method uses the designed features to obtain the location of tea sprouts, and the detection accuracy is highly depended on the designing of features. Wu et al. [13] introduce a $G$ and G-B components feature-based method to find the tea sprouts and background, but the accuracy is low as the obtained features are not highly semantic. A rapid watershed algorithm via the color information is proposed to segment the tea sprouts in [14]. However, the segmentation performance is not well as the low detection speed and accuracy. The robustness of the manually designed features is limited, which is mainly dependent on the experience of the designer.

(ii) The CNN-based method obtains the features automatically via the convolutional layers, which is a kind of data-driven method. Wang et al. [15] utilize the segmented samples to train a deep learning model to identify the tea sprouts, but the effect of background is not included during the training. The accuracy is low when the input is the fresh tea sprouts in outdoor condition. A method of recognizing the picking points of the tender tea shoots with the YOLOv3 [16] deep convolutional neural network algorithm was given in [17], but it also uses the segmented samples as the input. The CNN-based fresh tea sprouts detection method belongs to the category of object detection, which can be divided into two types: a one-stage object detector and two-stage object detector. The one-stage-based methods became gradually popular from the YOLO [18], and then, the YOLO-v2 [19], SSD [12], FSSD [20], DSSD [21], and ASSD [22] occurred. The two-stage-based methods are famous with the R–CNN [23], Fast R–CNN [24], Faster R–CNN [25], and SPPnet [26].

The conventional SSD [12] approach performs the detection via the feed-forward convolutional network and scores for the presence of object class instances in fixed-size bounding boxes. Figure 1 shows the network architecture of SSD; the backbone of the network is based on the VGG-16, and some auxiliary structure is added. Convolutional feature layers are added to obtain more multiscale feature maps in the SSD. So, there are more fixed sets of detection predictions by the added layers. Then, the object is detected via the default bounding boxes with a fixed position in each feature map cell and the detection predictions. During the training part, the bounding boxes matching strategy is implemented by the location, aspect ratio, and scale.

The one-stage-based methods model the detection task as a regress problem, and all the computation is encapsulated in a feed-forward convolutional network [27–29]; thereby, the detection speed is highly improved. The two-stage-based methods are proposal driven, and the second stage is used to refine the detection [30–32]. The existing SSD-based methods usually use the shallow layer to achieve the high detection speed. However, the shallow layer suffers the semantic information lacking problem which leads to a low accuracy on detecting small objects. A straightforward way to solve that problem is to increase the number of network layers, but the detection speed decreases quickly, and the

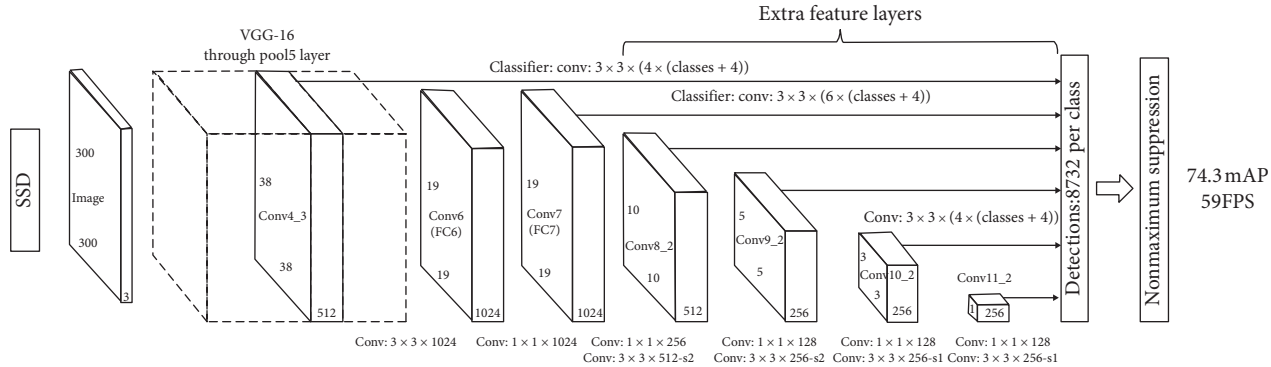



Figure 1: Network architecture of the conventional SSD method by Liu [12]; the VGG-16 is used as the backbone.

advantage of one-stage object detector no longer exists either.

In this paper, we rethink the process of fresh tea sprouts detection, where the process contains input data, detection, and output result. Many CNN-based improvement strategies [33–36] for object detection are focused on detection network structure or feature fusion. There are also a few improvement works on output result, such as label distribution learning [37–39], but the improvement strategies on input data for FTSD are not published. Meanwhile, motivated by the multispectral image processing, we find that more input information can lead to a better detection result. With this in mind, a novel Fresh Tea Sprouts Detection method via Image Enhancement and Fusion SSD (FTSD-IEFSSD) is proposed in this paper. In our method, we obtain an enhanced image via RGB-channel-transform-based image enhancement algorithm with the input of an original fresh tea sprouts color image. The enhanced image can provide more input information, where the contrast in fresh tea sprouts area is increased and the background area is decreased. Then, the enhanced image and color image are used in the detection subnetwork separately, where the ResNet50 is employed as the backbone. We also use the multilayer semantic fusion and adaptive scores fusion to further improve the detection accuracy. The strategy of tea sprouts shape-based default boxes is also included during the training. The novel improvement strategy using input information increase via image enhancement and the semantic fusion operation can balance the calculation speed and object detection accuracy. The experimental results show that the novel method leads to good detecting results; meanwhile, the proposed method can be used in many other fresh sprouts detecting tasks.

## 3. Methods

*3.1. The Proposed FTSD-IEFSSD.* The network architecture of our method is shown in Figure 2; the input is a color image with fresh tea sprouts, and the output is the corresponding results of FTSD. In order to improve the detection performance, a novel improvement strategy for object detection using input information increase via image enhancement is proposed in our method. We develop an RGB-channel-transform-based image enhancement algorithm to get the enhanced image, which can provide more input information. The details of the image enhancement algorithm are introduced in Section 3.2. Therefore, the proposed network architecture is mainly composed of two subnetworks: an enhanced image subnetwork and color image subnetwork, which can extract more useful features and default boxes. Motivated by ASSD [22], we use the ResNet50 (conv1-5) as the backbone (see Table 1) in the subnetworks and build the pyramid convolutional blocks (conv6-9) following the same design of the conventional SSD.

In Figure 2, conv1_e means the first convolutional layer in the enhanced image subnetwork and conv1_c is the corresponding layer in the color image subnetwork. We use the conv3–9 to detect the fresh tea sprouts with different scales and also utilize the ReLU and batch normalization in hidden layers. The conv3 is enhanced via the feature map fusion of conv3–6 to obtain more semantic information, and the details are shown in Section 3.3. The prediction layer in each subnetwork is the same as the conventional SSD. In order to obtain the final detection result, the score fusion layer is used to merge the two subnetwork detection confidence scores with equal weights of 0.5, and the nonmaximum suppression is also used to remove the impact of overlapping boxes.

The setting of default boxes usually affects the accuracy of object detection directly in the SSD-based method. We observed that the scales and aspect ratios of fresh tea sprouts are different as the camera view and the individual difference in sprout. It also can be observed that the size of fresh tea sprouts is much smaller than the ripe leaves. Therefore, we reset the default boxes and optimize the scales and aspect ratios according to the size characteristics of fresh tea sprouts. Comparing with the method of building more CNN layers, the computational cost in our method is much lower, but the detection results are better.

*3.2. The Proposed RGB-Channel-Transform-Based Image Enhancement Algorithm.* In order to further improve the detection performance, we propose an RGB-channel-transform-based image enhancement algorithm to obtain the enhanced image, which can increase the input information. The enhanced image by our algorithm is calculated as

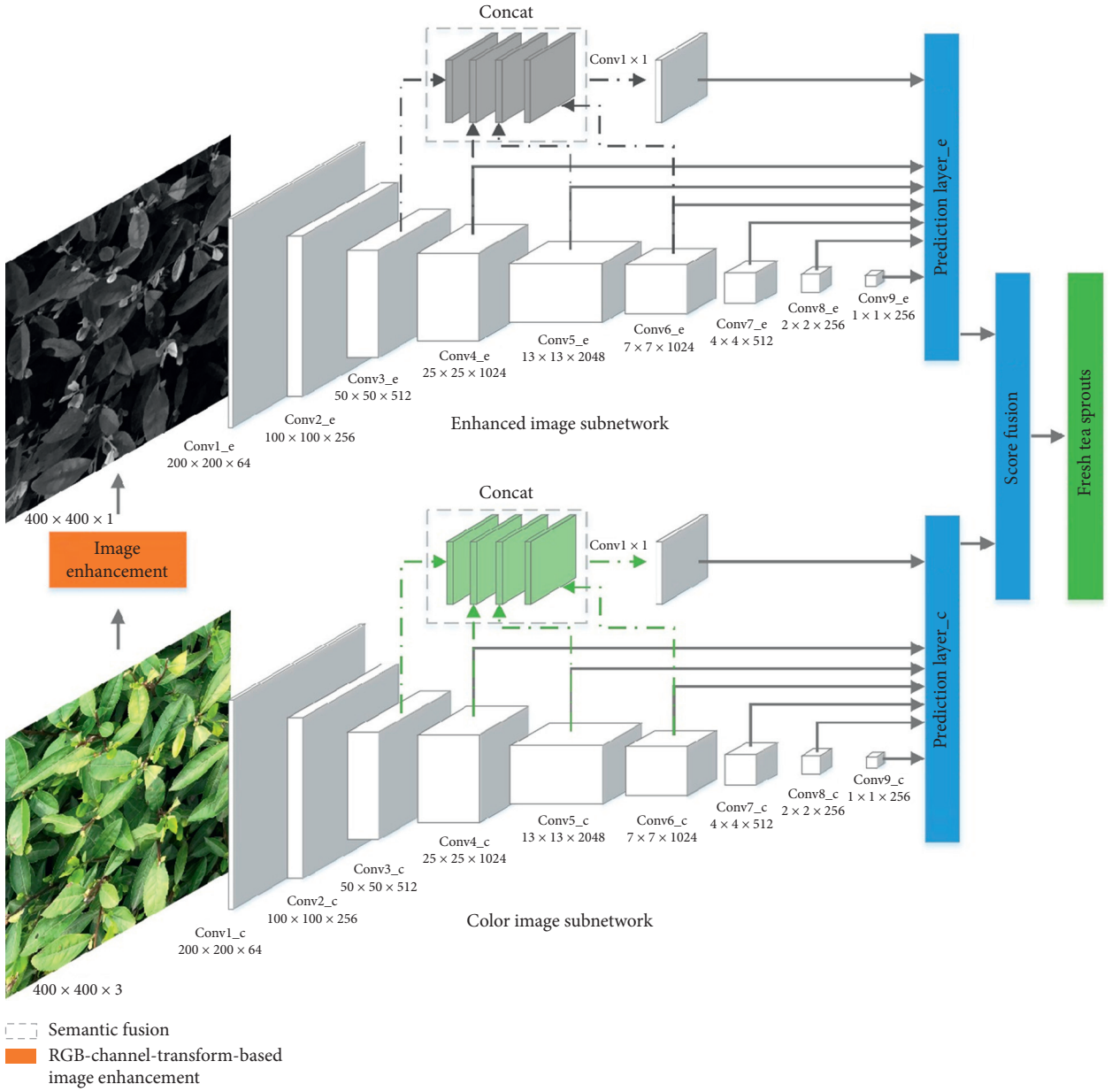



Figure 2: Network architecture of the proposed FTSD-IEFSSD method; the ResNet50 is used as the backbone of each subnetwork (conv1–5), and the extra convolutional layers and prediction layer follow the same design as the conventional SSD. An RGB-channel-transform-based image enhancement algorithm is proposed to get the enhanced image, which can provide more input information. The fusion features (enhanced conv3) are obtained from the semantic fusion of conv3–6, and the bilinear interpolation is employed in conv4–6 to align the same size of conv3. The score fusion is also employed to achieve the fresh tea sprouts detection.

$$\mathbf{f}_{\mathrm{En}} = \alpha\left|\mathbf{f}_R - \mathbf{f}_B\right| + \beta\mathbf{f}_G, \tag{1}$$

where $\mathbf{f}_{\mathrm{En}}$ is the enhanced image, $\mathbf{f}_R$, $\mathbf{f}_G$, and $\mathbf{f}_B$ are the corresponding RGB channels in the color image, $\alpha$ and $\beta$ are the weight coefficients and set as $\alpha + \beta = 1$, and  is the absolute value operation. In equation 1, the information of the fresh tea sprouts is mainly enhanced by the absolute value operation, and the $\mathbf{f}_G$ is used to adjust the brightness of the whole image. Therefore, we set the value of $\alpha$ much larger than $\beta$, and the analysis of the weight coefficients is introduced in Section 4.3. The input information is increased via the proposed image enhancement, and the detection accuracy is also increased in the proposed FTSD-IEFSSD.

*3.3. Semantic Fusion.* Fresh tea sprouts detection is a challenging task as the outdoor imaging condition and detection speed. The SSD utilizes the shallow layer to achieve the high detecting speed, but the insufficient semantic information causes the low detecting accuracy of small objects (fresh tea sprouts belong to a small object in the object detection task). Building more CNN layers is a straightforward way to solve that problem, but the detecting speed is

Table 1: Architecture of FTSD-IEFSSD with ResNet50 backbone.

| Layer name | Output size | Specification |
|---|---|---|
| Conv1 | $200 \times 200$ | $7 \times 7$, 64, stride 2 |
| Conv2 | $100 \times 100$ | $\begin{bmatrix} 1 \times 1 & 64 \\ 3 \times 3 & 64 \\ 1 \times 1 & 256 \end{bmatrix} \times 3$ |
| Conv3 | $50 \times 50$ | $\begin{bmatrix} 1 \times 1 & 128 \\ 3 \times 3 & 128 \\ 1 \times 1 & 512 \end{bmatrix} \times 4$ |
| Conv4 | $25 \times 25$ | $\begin{bmatrix} 1 \times 1 & 256 \\ 3 \times 3 & 256 \\ 1 \times 1 & 1024 \end{bmatrix} \times 6$ |
| Conv5 | $13 \times 13$ | $\begin{bmatrix} 1 \times 1 & 512 \\ 3 \times 3 & 512 \\ 1 \times 1 & 2048 \end{bmatrix} \times 3$ |

significantly reduced. The extracted feature fusion-based method is a better way as the good balance of calculation speed and object detection accuracy. Therefore, we fuse the contextual information from the shallow layers to enrich its semantics and utilize the tea sprouts shape-based default boxes to improve the performance of FTSD. The semantic fusion is employed in the proposed FTSD-IEFSSD. However, there are many feature maps can be used during the semantic fusion, such as layers 2–7 in Figure 2. The feature maps obtained from high levels usually have more global semantic information, and the low levels contain more local detailed features. It is hard to detect the fresh tea sprouts by the conventional SSD as the context information and detailed features need to be merged in one detector.

Motivated by FSSD [20], we fuse the contextual information from layers 3–6 to enhance the semantics of layer 3, which is shown by the gray dotted box in Figure 2. Layers 7–9 are not included in our semantic fusion part due to the small size and limited information to merge. The process of semantic fusion can be formulated as

$$\mathbf{x}^3 = \mathbf{W}^3 \text{Concat}\left\{\mathbf{x}^3, \mathbf{x}^4, \mathbf{x}^5, \mathbf{x}^6\right\} + \mathbf{b}^3, \quad (2)$$

where $\mathbf{x}^n$ is the feature map at convolutional layer $n$, $\mathbf{W}$ and $\mathbf{b}$ are the corresponding weight and coefficient separately, and Concat is the concatenation operation. Layers 4–6 are upsampled via bilinear interpolation in order to obtain the same size as layer 3 during the concatenation operation.

## 4. Results

*4.1. Implementation Details.* The experiments are run with 4 NVIDIA Titan *X* GPUs, the FTSD-IEFSSD model is trained on TensorFlow, and other steps are implemented using MATLAB. The input image resolution is $400 \times 400$ with top view. The dataset contains 6000 images, which were acquired in Hangzhou, China, from March 20 to April 4, 2019. Each fresh tea sprout in the dataset has one leaf or two (see 2). We use 80% of fresh tea sprouts images for training and 20% for evaluation, and the 10-fold cross validation is also used in the experiment. The enhanced image subnetwork and color image subnetwork are trained, respectively. The main evaluation metric is average precision (AP), which is widely used in the object detection task.

During the training, the min size ($s_{\min}$) and max size ($s_{\max}$) of the default boxes used in our method are one-seventh ratio of the SSD as the small tender tea shoot shape. We use the same strategy as the SSD to generate the default box and use the hard negative mining to solve the positive-negative box class imbalance problem. The aspect ratio ($a_r$) for default boxes in layers 3, 8, and 9 is $\{1, 2, 1/2\}$ and in layers 4–7 is $\{1, 2, 1/2, 3, 1/3\}$. The width ($w$) and height ($h$) normalized calculation of default box is also the same as the conventional SSD. We utilize the Stochastic Gradient Descent (SGD) algorithm to optimize the weights, with a decay of 0.001 and initial learning rate of 0.001. The $\alpha$ and $\beta$ used in equation 1 are 0.95 and 0.05 separately. The overall objective loss function consisting of localization loss and confidence loss is also employed.

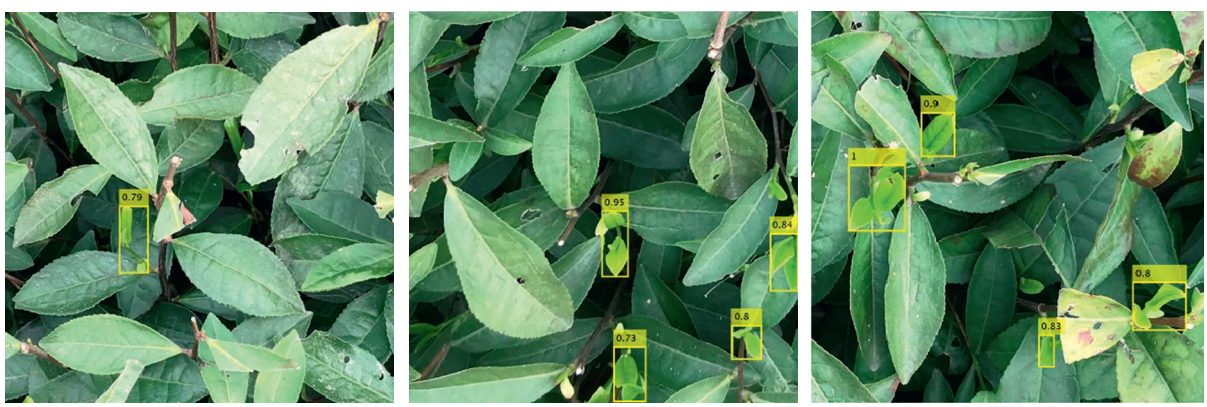
*4.2. Fresh Tea Sprouts Detection Results.* Comparisons of different methods for FTSD on the testing dataset are shown in Table 2, where the AP at different Intersection over Union (IoU) thresholds 0.5 (AP50) and 0.75 (AP75), averaged over thresholds between 0.5 and 1(AP) and Frames Per Second (FPS), are used as the evaluation metrics. We retrained the network models following the conventional training strategy as the pretrained models are no longer suitable for the task of FTSD. The min size and max size of the default boxes used in the comparative methods are the same as in our method. From Table 2, it can be observed that the semantic feature fusion-based methods (FSSD and the proposed FTSD-IEFSSD) have better performance than other models on AP. The main reason is that the semantic information lacking problem in shallow layers can be solved by the semantic fusion with different layers. We also observed that the one-stage object detector (YOLOv3 and SSD-based methods) has a big improvement on detection speed (FPS). Meanwhile, the corresponding detection accuracy is close to the two-stage object detector. Compared with the SSD method (ResNet50 backbone), our method has better performance on AP, AP50, and AP75, which are increased to 83.9, 92.8, and 88.9 separately. The FPS of our method is lower than the conventional SSD due to the two subnetworks and the semantic fusion operation. However, the detection accuracy by the proposed FTSD-IEFSSD is increased significantly, which can further improve the success rate of automated tea picking.

Figure 3 shows part of FTSD results obtained by the conventional SSD (ResNet50 backbone), R–FCN (ResNet50 backbone), FSSD, and the proposed FTSD-IEFSSD. In Figure 3, the IoU threshold with a score of 0.5 or above is drawn. We observed that the FTSD results gained by the conventional SSD contain more tea sprout branches, while some small tea sprouts are lost during the detection. The abovementioned phenomenon leads to a low FTSD detection accuracy compared to the other methods in Figure 3. The R–FCN has better detection accuracy than the conventional SSD, but the FPS is quickly dropped. The detection accuracy obtained by the FSSD is better than the conventional SSD and R–FCN, but it is not high to achieve the

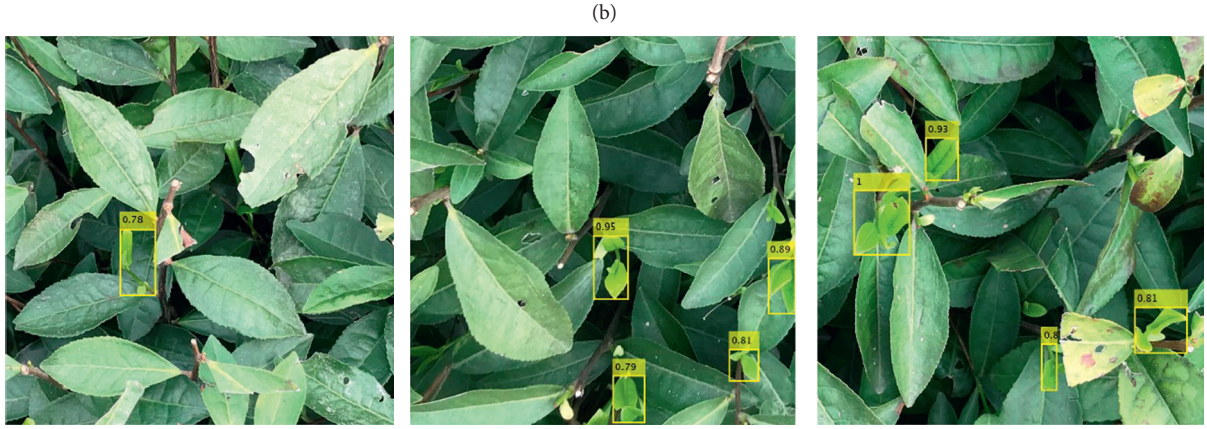
Table 2: Comparisons of different methods for FTSD on the testing dataset.

| Method | Backbone network | FPS | AP | AP50 | AP75 |
|---|---|---|---|---|---|
| Faster R–CNN | VGG16 | 11.0 | 77.9 | 86.3 | 81.4 |
| Faster R–CNN | ResNet50 | 6.3 | 80.4 | 89.0 | 84.7 |
| R–FCN | ResNet50 | 21.5 | 80.9 | 89.3 | 85.1 |
| R–FCN | ResNet101 | 14.7 | 81.1 | 90.0 | 86.1 |
| YOLOv3 | Darknet-53 | 31.0 | 80.9 | 89.8 | 85.0 |
| SSD | VGG16 | 37.2 | 79.6 | 88.9 | 84.1 |
| SSD | ResNet50 | 25.4 | 80.7 | 89.1 | 85.0 |
| DSSD | ResNet50 | 22.0 | 81.5 | 90.1 | 86.9 |
| FSSD | VGG16 | 50.2 | 81.0 | 89.9 | 86.0 |
| FTSD-IEFSSD | ResNet50 | **15.1** | **83.9** | **92.8** | **88.9** |


(a)




(b)




(c)

Figure 3: Continued.

(d)

Figure 3: Detection results obtained by the R–FCN (ResNet50 backbone), conventional SSD (ResNet50 backbone), FSSD, and the proposed FTSD-IEFSSD; the leaves in the yellow boxes are the detected fresh tea sprouts, and an IoU threshold with a score of 0.5 or above is drawn. (a) Detection results obtained by the conventional SSD (ResNet50 backbone). (b) Detection results obtained by the R–FCN (ResNet50 backbone). (c) Detection results obtained by the FSSD. (d) Detection results obtained by the proposed FTSD-IEFSSD.

requirement of vision-based tea picking robots. In Figure 3, it can be observed that more small fresh tea sprouts are detected via our method, especially the tea sprouts are partly covered by the mature leaves. The main reason is the fused contextual information from layers 3–6 can provide more details about the fresh tea sprouts. The comparisons in Table 2 and 3 indicate that the proposed FTSD-IEFSSD has a better overall performance on AP and FPS than the state-of-the-art methods, which is an effective support to the high-quality tea picking task.

### 4.3. Model Analysis

*4.3.1. Effects of Min Size and Max Size for Default Boxes.* To understand the proposed FTSD-IEFSSD better, we carried out some comparative experiments to analyze how the key components affect the tea sprouts detection performance. The effects of min size ($s_{\min}$) and max size ($s_{\max}$) for default boxes on FTSD-IEFSSD are shown in Table 3, where the aspect ratio ($a_r$) for default boxes in layers 3, 8, and 9 is $\{1, 2, 1/2\}$ and in layers 4–7 is $\{1, 2, 1/2, 3, 1/3\}$, and the FPS and AP are the results of the testing dataset. From Table 3, we find that the AP in row 3 has a significant improvement compared with row 2. The main reason is that some fresh tea sprouts have the outline sizes between the default boxes $s_{\min} = 0.03$ and $s_{\min} = 0.04$. We also find that the improvement of AP from row 3 to row 8 is limited as the fresh tea sprouts with the outline sizes between the default boxes $s_{\max} = 0.13$ and $s_{\max} = 0.14$ are rare. Therefore, the above-mentioned phenomenon indicates that the default boxes with min size 0.03 and max size 0.13 can cover most outline sizes of fresh tea sprouts.

*4.3.2. Effects of Aspect Ratio for Default Boxes.* The effects of aspect ratio ($a_r$) are shown in Table 4, where the min size ($s_{\min}$) and max size ($s_{\max}$) for default boxes are 0.03 and 0.13 separately, and the FPS and AP are the results of the testing dataset too. From Table 4, it can be observed that the AP is increased with more default box shapes, but the FPS is decreased due to the computational cost of the added default boxes. If we remove the default boxes with 3 and 1/3 aspect ratio in all layers, the AP drops 3.2 and the FPS raises 5.9 compared with the second row in Table 4. In order to achieve the requirement of vision-based automatic tea picking, the detection accuracy and speed should be balanced. Therefore, we set the aspect ratio ($a_r$) for default boxes in layer 3, 8, and 9 to $\{1, 2, 1/2\}$ and in layers 4–7 to $\{1, 2, 1/2, 3, 1/3\}$. In the proposed FTSD-IEFSSD, we also utilize the contextual information from layers 3–6 to enhance the semantics of layer 3 and the shape-based default boxes to obtain more information about the tea sprouts. Then, the computational cost of the proposed method is reduced, and the detection accuracy is increased by these optimizing strategies.

*4.3.3. Effects of Weight Coefficients for Image Enhancement.* The image enhancement results by different weight coefficients ($\alpha$ and $\beta$) are shown in Figure 4, and the value of $\alpha$ is, respectively, 0.6, 0.7, 0.8, 0.9, and 0.95. It can be observed that the contrast of the fresh tea sprouts is improved with the increase of $\alpha$ in Figure 4; meanwhile, the negative effects on FTSD caused by the background are gradually reduced. We also set $\alpha$ to 1, but the texture details of the fresh tea sprouts are easy to be lost. Therefore, we use 0.95 as the value of $\alpha$ in the experiment. The enhanced image obtained by our method can provide more input information and more salient features, which leads to better FTSD performance comparing with the original color image. In addition, the enhanced image is a kind of gray image, and little background information may be lost during the process of image enhancement. Therefore, we use the enhanced image subnetwork and color image subnetwork to solve the input information and salient features lacking problem. The multilayer semantic fusion and scores fusion are also used to further improve the fresh tea sprouts detection accuracy.

*4.3.4. Effects of Different Fusion Layers for Detection Accuracy.* The semantic fusion used in our method is an

Table 3: Effects of min size and max size for default boxes on FTSD-IEFSSD.

| Min size ($s_{\min}$) | Max size ($s_{\max}$) | FPS | AP |
|---|---|---|---|
| 0.05 | 0.13 | 21.2 | 78.5 |
| 0.04 | 0.13 | 16.6 | 81.2 |
| 0.03 | 0.13 | **15.1** | **83.9** |
| 0.02 | 0.13 | 6.0 | 84.2 |
| 0.01 | 0.13 | 3.6 | 83.7 |
| 0.03 | 0.11 | 20.8 | 81.6 |
| 0.03 | 0.12 | 17.1 | 82.6 |
| 0.03 | 0.14 | 13.6 | 84 |
| 0.03 | 0.15 | 12.7 | 83.8 |

Table 4: Effects of aspect ratio for default boxes on FTSD-IEFSSD.

| Aspect ratio ($a_r$) | | | | FPS | AP |
|---|---|---|---|---|---|
| Layer 3 | Layer 4–7 | Layer 8 | Layer 9 | | |
| {1, 2, 1/2} | {1, 2, 1/2} | {1, 2, 1/2} | {1, 2, 1/2} | 21 | 80.7 |
| {1, 2, 1/2} | {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2} | {1, 2, 1/2} | **15.1** | **83.9** |
| {1, 2, 1/2} | {1, 2, 1/2} | {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2} | 20 | 80.8 |
| {1, 2, 1/2} | {1, 2, 1/2} | {1, 2, 1/2} | {1, 2, 1/2, 3, 1/3} | 20.5 | 80.7 |
| {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2} | {1, 2, 1/2} | {1, 2, 1/2} | 18.3 | 81.8 |
| {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2} | {1, 2, 1/2} | 12.6 | 84.2 |
| {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2, 3, 1/3} | {1, 2, 1/2, 3, 1/3} | 6.7 | 84.3 |

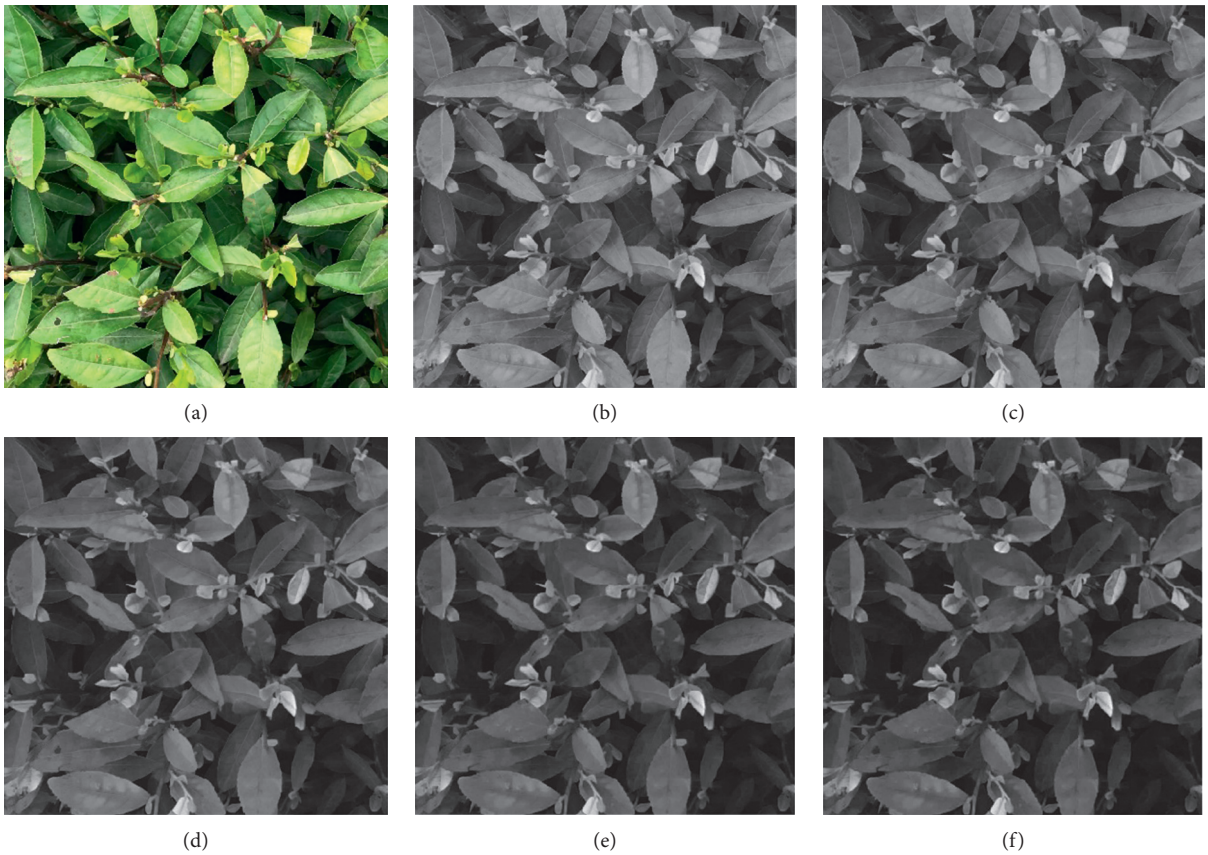


Figure 4: The image enhancement results by different weight coefficients ($\alpha$ and $\beta$); $\alpha$ is, respectively, 0.6, 0.7, 0.8, 0.9, and 0.95. (a) Color image, (b) enhanced image$\alpha = 0.6$, $\beta = 0.4$, (c) enhanced image $\alpha = 0.7$, $\beta = 0.3$, (d) enhanced image $\alpha = 0.8$, $\beta = 0.2$, (e) enhanced image $\alpha = 0.9$, $\beta = 0.1$, and (f) enhanced image $\alpha = 0.95$.$\beta = 0.05$.

Table 5: Effects of different fusion layers for detection accuracy on FTSD-IEFSSD.

| Semantic fusion layers | | | | FPS | AP |
|---|---|---|---|---|---|
| Layer3 | Layer 4 | Layer 5 | Layer 6 | | |
| ✓ | ✓ | ✗ | ✗ | 15.5 | 82.8 |
| ✓ | ✓ | ✓ | ✗ | 15.3 | 83.0 |
| ✓ | ✓ | ✓ | ✓ | **15.1** | **83.9** |

Table 6: Comparisons of different network structures in the FTSD-IEFSSD for FTSD.

| Method | Backbone network | Semantic fusion | FPS | AP | AP50 | AP75 |
|---|---|---|---|---|---|---|
| Enhanced image subnetwork | ResNet50 | ✓ | 28.5 | 81.4 | 90.7 | 87.1 |
| Color image subnetwork | ResNet50 | ✓ | 24.9 | 81.5 | 90.8 | 87.3 |
| FTSD-IEFSSD | ResNet50 | ✓ | **15.1** | **83.9** | **92.8** | **88.9** |
| Enhanced image subnetwork | ResNet50 | ✗ | 28.7 | 80.5 | 89.1 | 84.9 |
| Color image subnetwork | ResNet50 | ✗ | 25.4 | 80.7 | 89.6 | 85.0 |
| FTSD-IEFSSD | ResNet50 | ✗ | 15.6 | 82.2 | 91.5 | 87.6 |
| FTSD-IEFSSD | ResNet34 | ✓ | 17.0 | 82.1 | 91.4 | 87.6 |
| FTSD-IEFSSD | ResNet101 | ✓ | 11.2 | 84.5 | 93.2 | 89.7 |
| FTSD-IEFSSD | VGG-16 | ✓ | 35.6 | 82.0 | 90.9 | 87.5 |

efficient way as the good balance of calculation speed and object detection accuracy. In the experimental part, we evaluate the effects of different fusion layers for detection accuracy, and the details are shown in Table 5. The FPS and AP are the corresponding results of the testing dataset too, and the strategies of aspect ratio and weight coefficients ($\alpha$ and $\beta$) used in different fusion layers follow the baseline of the proposed FTSD-IEFSSD.

From Table 5, it can be observed that the AP is raised with the adding of fusion layers, but the FPS only changes a little. Specifically, the decrease of 0.4 in FPS can bring an increase of 1.1 in detection accuracy, which is an efficient way to improve the detection accuracy. The main reason is more semantic features are obtained by more convolutional layers. Then, the differences between the fresh tea sprouts and ripe leaves are obtained and enhanced. Therefore, in our task, the detection accuracy is increased via the addition of fusion layers.

The feature maps obtained from high levels usually have more global semantic information, but the layers 7–9 in ResNet50 are not included in our semantic fusion part due to the small size and limited information to merge. Therefore, we fuse the contextual information from layers 3–6 to enhance the semantics of layer 3, which has better performance on FTSD. Meanwhile, the detection speed satisfied the requirement (the FPS is usually larger than 13) of the vision-based automatic tea picking robot.

*4.3.5. Comparisons of Different Network Structures in the FTSD-IEFSSD.* In the experimental part, we also analyze the effect of the network structures on FTSD. Comparisons of different network structures (single subnetwork, include semantic fusion operation or not, and different backbones in the feature extraction part) are shown in Table 6, and the FPS and AP are the corresponding results of the testing dataset. The semantic fusion operation and strategy of aspect ratio used in the single subnetwork structure (ResNet34, ResNet50, ResNet101, and VGG-16 backbone based) are the same as the FTSD-IEFSSD. The min size ($s_{\min}$) and max size ($s_{\max}$) for default boxes used in all network structures are 0.03 and 0.13 separately. The weight coefficients ($\alpha$ and $\beta$) used in the enhanced image subnetwork are 0.95 and 0.05 separately. The VGG-16 backbone network structure follows the baseline of the conventional SSD [12], but the semantic fusion operation is the same as the proposed FTSD-IEFSSD.

From the comparisons of different network structures in Table 6, we find that the two subnetwork-based methods (row 3 and 6–9 in Table 6) have an average improvement of 1.92 on AP than the single subnetwork-based methods (row 1, 2, 4, and 5 in Table 6), which indicates that the FTSD accuracy is improved by the strategy of two subnetworks. The main reason is that the two subnetwork-based methods can obtain more input information via the enhanced image subnetwork. Meanwhile, more useful features can be extracted by the added input information, which leads to a better performance on FTSD. From Table 6, we also find that the detection methods (row 1–3 in Table 6) using semantic fusion improve the AP by 1.13 compared with the methods without semantic fusion (row 4–6 in Table 6). More local detailed features with global semantic information are obtained during the semantic fusion operation, where more small fresh tea sprouts can be detected. Therefore, the detection performance is improved via the semantic fusion and added input information in the proposed FTSD-IEFSSD.

The backbone network used in the feature extraction part usually has little impact on the detection accuracy. In the experimental part, we also compare different backbone networks to find a more suitable network structure to further improve the detection performance. In Table 6, the detection performance of our method with different backbone networks (ResNet34, ResNet50, ResNet101, and VGG-16) are shown in row 2 and 7–9. It can be observed that the VGG-16-based method has the highest detection speed but the detection accuracy is low. Meanwhile, the ResNet101 has the best performance on detection accuracy, but the detection speed cannot meet the requirement of the vision-based automatic tea picking robot. Therefore, the detection accuracy and speed

should be balanced to obtain better performance on FTSD. In the proposed FTSD-IEFSSD, we use the ResNet50 as the backbone network which has the best comprehensive detection performance. The detection accuracy of our method is improved by 1.9 compared with the VGG-16-based method. From the comparison of different network structures in Table 6, it also can be observed that the AP obtained by our method is higher than the others. The main reason is that the proposed network structures in our method are more suitable for the fresh tea sprouts detection. Meanwhile, the idea of adding input information via image enhancement is another important factor.

## 5. Conclusions

In this paper, we develop the FTSD-IEFSSD, a novel fresh tea sprouts detection method, via image enhancement and improved fusion SSD. The strategy for object detection using input information increase via image enhancement is proposed in our method, which is a new way to improve the detection performance. The multilayer semantic fusion operation, adaptive scores fusion, and the tea sprouts shape-based default boxes are also employed to balance the calculation speed and detection accuracy. The experimental results indicate that the proposed method leads to good performance on FTSD. The testing dataset AP of our method is 83.9, which has better performance than the state-of-the-art methods. The detection speed (FPS) of our method is 15.1, which can basically meet the requirement of high-quality tea picking task. FTSD is still a challenging task due to the uncontrollable imaging condition and the small size of tea sprouts. Future work includes the following aspects. More fresh tea sprouts images should be collected. The semantic fusion operation could be further optimized to improve the detection performance. The model compression operation could be included to increase the detection speed.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## Acknowledgments

This research was funded by the Science and Technology Planning Project of Jiaxing, China, Grant no. 2018AY11001, and Zhejiang Provincial Natural Science Foundation of China under Grant nos. LQ20F020026, LQ18F020007, and LY18F020021.

## References

[1] W. Pei and X. Wang, "The two-dimension coordinates extraction of tea shoots picking based on image information," *Acta Agriculturae Zhejiangensis*, vol. 28, pp. 522–527, 2016.

[2] G. Ya, "Research on the application of automation software control system in tea garden mechanical picking," in *Proceedings of the International Conference on Applications and Techniques in Cyber Security and Intelligence*, pp. 1830–1836, Glasgow, UK, July 2019.

[3] H. Yang, L. Chen, M. Chen et al., "Tender tea shoots recognition and positioning for picking robot using improved YOLO-V3 model," *IEEE Access*, vol. 7, pp. 180998–181011, 2019.

[4] Z. Tang, Y. Su, M. J. Er, F. Qi, L. Zhang, and J. Zhou, "A local binary pattern based texture descriptors for classification of tea leaves," *Neurocomputing*, vol. 168, pp. 1011–1023, 2015.

[5] A. K. Hazarika, S. Chanda, S. Sabhapondit et al., "Quality assessment of fresh tea leaves by estimating total polyphenols using near infrared spectroscopy," *Journal of Food Science and Technology*, vol. 55, no. 12, pp. 4867–4876, 2018.

[6] N. Bhattacharyya, R. Bandyopadhyay, M. Bhuyan, B. Tudu, D. Ghosh, and A. Jana, "Electronic nose for black tea classification and correlation of measurements with "Tea Taster" marks," *IEEE Transactions on Instrumentation and Measurement*, vol. 57, no. 7, pp. 1313–1321, 2008.

[7] M. B. Banerjee, R. B. Roy, B. Tudu, R. Bandyopadhyay, and N. Bhattacharyya, "Black tea classification employing feature fusion of E-Nose and E-Tongue responses," *Journal of Food Engineering*, vol. 244, pp. 55–63, 2019.

[8] T. Yiping, W. Weiyang, Z. Wei et al., "Tea ridge identification and navigation method for tea-plucking machine based on machine vision," *Transactions of the Chinese Society for Agricultural Machinery*, vol. 47, pp. 45–50, 2016.

[9] C. C. Wu, "Developing situation of tea harvesting machines in Taiwan," *Engineering, Technology & Applied Science Research*, vol. 5, no. 6, pp. 871–875, 2015.

[10] Z.-Q. Zhao, P. Zheng, S.-T. Xu, and X. Wu, "Object detection with deep learning: a review," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.

[11] J. Yi, P. Wu, and D. N. Metaxas, "Single shot multibox detector," *Computer Vision and Image Understanding*, vol. 18, 2017.

[12] W. Liu, D. Anguelov, D. Erhan et al., "Ssd : single shot multibox detector," in *Proceedings of theEuropean conference on computer vision*, pp. 21–37, Glasgow, UK, December 2016.

[13] X. Wu, F. Zhang, and J. Lv, "Study on the method of tea tender leaf recognition based on image color information," *Journal of Tea Science*, vol. 33, pp. 584–589, 2013.

[14] M. Shao, *Research on Computer Vision Based Recognition Methods of Longjing Tea Sprouts*, China Jiliang University, Jiliang, China, 2013.

[15] K. Wang and D. Liu, "Intelligent identification for tea state Based on deep learning," *Journal of Chongqing Institute of Technology*, vol. 29, pp. 120–126, 2015.

[16] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, http://arxiv.org/abs/1804.02432.

[17] H. Yang, L. Chen, M. Chen et al., "Tender tea shoots recognition and positioning for picking robot using improved YOLO-V3 model," *IEEE Access*, vol. 7, pp. 180998–181011, 2019.

[18] J. Redmon, S. Divvala, R. Girshick et al., "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, Seattle, WA, USA, May 2016.

[19] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the 2017 IEEE conference on*

*computer vision and pattern recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[20] Z. Li and F. Zhou, "FSSD: feature fusion single shot multibox detector," 2018, http://arxiv.org/abs/1712.00960.

[21] C. Y. Fu, W. Liu, A. Ranga et al., "Dssd: deconvolutional single shot detector," 2017, http://arxiv.org/abs/1701.06659.

[22] J. Yi, P. Wu, and D. N. Metaxas, "ASSD: attentive single shot multibox detector," *Computer Vision and Image Understanding*, vol. 189, Article ID 102827, 2019.

[23] R. Girshick, J. Donahue, T. Darrell et al., "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, 2014.

[24] R. Girshick, "Fast r-cnn," in *Proceedings of the 2015 IEEE International Conference on Computer Vision*, pp. 1440–1448, Santiago, Chile, 2015.

[25] S. Ren, K. He, R. Girshick et al., "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, pp. 91–99, 2015.

[26] P. Purkait, C. Zhao, and C. Zach, "SPP-Net: deep absolute pose regression with synthetic views," 2017, http://arxiv.org/abs/1712.03452.

[27] X. Wu, D. Sahoo, and S. C. H. Hoi, "Recent advances in deep learning for object detection," *Neurocomputing*, vol. 396, pp. 39–64, 2020.

[28] Z. Tian, C. Shen, H. Chen et al., "Fcos: fully convolutional one-stage object detection," in *Proceedings of the IEEE international conference on computer vision*, pp. 9627–9636, Seoul, Korea, August 2019.

[29] H. Li, Z. Wu, C. Zhu et al., "Learning from noisy anchors for one-stage object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10588–10597, Seattle, WA, USA, May 2020.

[30] K. Duan, L. Xie, H. Qi et al., "Corner proposal network for anchor-free, two-stage object detection," 2020, http://arxiv.org/abs/2007.13816.

[31] Y. Liu, J. Han, Q. Zhang et al., "Salient object detection via two-stage graphs," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, pp. 1023–1037, 2018.

[32] P. Soviany and R. T. Ionescu, "Optimizing the trade-off between single-stage and two-stage deep object detectors using image difficulty prediction," in *Proceedings of the 2018 20th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing*, pp. 209–214, IEEE, Timisoara, Romania, August 2018.

[33] M. Liang, B. Yang, Y. Chen et al., "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7345–7353, Long Beach, CA, USA, June 2019.

[34] Y. Zhou, P. Sun, Y. Zhang et al., "End-to-end multi-view fusion for 3d object detection in lidar point clouds," in *Proceedings of the Conference on Robot Learning. PMLR*, pp. 923–932, Long Beach, CA, USA, January 2020.

[35] N Wang and X Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, pp. 55277–55284, 2019.

[36] T. Saba, A. S. Mohamed, M. El-Affendi et al., "Brain tumor detection using fusion of hand crafted and deep learning features," *Cognitive Systems Research*, pp. 221–230, 2020.

[37] X. Geng, "Label distribution learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, pp. 1734–1748, 2016.

[38] J. Wang and X. Geng, "Theoretical analysis of label distribution learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 5256–5263, New York, NY, USA, July 2019.

[39] B. Chen and J. L. Yan, "Fresh tea shoot maturity estimation via multispectral imaging and deep label distribution learning," *IEICE Transactions on Information and Systems*, pp. 2019–2022, 2020.