

Research Article

BFP Net: Balanced Feature Pyramid Network for Small Apple Detection in Complex Orchard Environment

Meili Sun,^{1,2} Liancheng Xu,¹ Xiude Chen,³ Ze Ji,⁴ Yuanjie Zheng,¹ and Weikuan Jia ^{1,2}

¹School of Information Science and Engineering, Shandong Normal University, Jinan, China

²Key Laboratory of Facility Agriculture Measurement and Control Technology and Equipment of Machinery Industry, Zhenjiang 212013, China

³National Engineering Research Center for Apple, Shandong Agriculture University, Taian 271018, China

⁴School of Engineering, Cardiff University, Cardiff CF24 3AA, UK

Correspondence should be addressed to Weikuan Jia; jwk_1982@163.com

Received 1 April 2022; Accepted 31 August 2022; Published 26 September 2022

Copyright © 2022 Meili Sun et al. Exclusive Licensee Nanjing Agricultural University. Distributed under a Creative Commons Attribution License (CC BY 4.0).

Despite of significant achievements made in the detection of target fruits, small fruit detection remains a great challenge, especially for immature small green fruits with a few pixels. The closeness of color between the fruit skin and the background greatly increases the difficulty of locating small target fruits in the natural orchard environment. In this paper, we propose a balanced feature pyramid network (BFP Net) for small apple detection. This network can balance information mapped to small apples from two perspectives: multiple-scale fruits on the different layers of FPN and a characteristic of a new extended feature from the output of ResNet50 conv1. Specifically, we design a weight-like feature fusion architecture on the lateral connection and top-down structure to alleviate the small-scale information imbalance on the different layers of FPN. Moreover, a new extended layer from ResNet50 conv1 is embedded into the lowest layer of standard FPN, and a decoupled-aggregated module is devised on this new extended layer of FPN to complement spatial location information and relieve the problem of locating small apple. In addition, a feature Kullback-Leibler distillation loss is introduced to transfer favorable knowledge from the teacher model to the student model. Experimental results show that AP_s of our method reaches 47.0%, 42.2%, and 35.6% on the benchmark of the GreenApple, MinneApple, and Pascal VOC, respectively. Overall, our method is not only slightly better than some state-of-the-art methods but also has a good generalization performance.

1. Introduction

Fruit detection [1–3], a fundamental task for intelligence agriculture [4, 5], is the basis for the realization of yield prediction [6, 7], pesticide spraying [8], fruit harvesting [9, 10], and automatic monitoring of the whole process of fruit growth [11]. At present, the fruit detection algorithms are mainly based on the general object detection methods, which are optimized to use in the real orchard environment. However, the general object detection algorithms have the characteristic that the precision of small-scale objects is lower than that of large-scale objects. Strictly, the precision of small-scale objects is about half of that of large-scale objects [12], which greatly limits the improvement of fruit detection performance. Therefore, how to improve the precision of fruit detection, especially the precision of small

fruits, has attracted the attention of many scholars at home and abroad. Currently, the biggest problem lies with no strict and clear definition of small objects. One definition refers to fruits with smaller physical size in the real world, such as small fruits in the early stages of growth. Another definition is based on the evaluation index of MS COCO [13] and defines the object area with the number of pixels less than 32^2 as the small objects. More precisely, objects defined by the latter are divided into small objects with $0 < \text{area} < 32^2$, medium objects with $32^2 < \text{area} < 96^2$, and large objects with $\text{area} > 96^2$ according to the number of pixels in the object area. In this paper, the latter is adopted to define the small apples.

Traditional small object detection methods mainly based on the machine learning [14, 15] and digital image processing such as texture information [16] and geometric information [17] were employed to identify and detect the target

fruits. With the emergence of graphics processing units (GPUs), a series of small object detection models [18–22] have been proposed. The multiple-scale feature fusion architectures such as feature pyramid network (FPN) [23] and its variants are a typical trick to improve the performance of small object detection. FPN incorporates multiscale features through the lateral connections and a top-down structure. Since the small objects occupying a small number of pixels cannot meet the needs of target detection, the contextual information is integrated into all objects [12], especially small-scale objects. Contextual information about objects can expand the receptive field of objects to complement features of small objects. Although the performance of small targets can be improved to some extent via contextual information, there is still a large gap between the detection precision of small targets and large objects [12]. The small number of pixels occupied by small objects is an important factor hindering the improvement of small object detection performance. Therefore, how to increase the small number of pixels occupied by small objects is still the key to solving the problem. Based on this point of view, a series of small object detection algorithms based on generative adversarial networks (GANs) such as ESRGAN [24] and EESRGAN [25] are derived to generate high-resolution feature mapping and to increase the number of pixels of a small object. An extended feature map from the output of the ResNet/ResNeXt conv1 without the pooling operation was introduced to generate a high-resolution feature map and a foreground-background-balanced loss based on the distillation mechanism was also applied to balance the foreground and background [26].

In this paper, we take Faster R-CNN [27] as the baseline and optimize feature pyramid network to make it suitable for small apple detection. The Faster R-CNN consists of three components: ResNet50 [28], FPN, and a detector. Although FPN can incorporate multiple-scale features, there are unbalanced characteristics of different scale information in each feature layer of FPN. In other words, the fusion feature by element-wise addition of equal proportion will affect the precision of small apple detection. To present this problem more clearly, we quantified the prediction results of different scales of apples on the different resolution feature layer of the FPN structure.

Figure 1 presents the prediction results of each feature layer of FPN on the Faster R-CNN based on GreenApple, where the horizontal axis represents a certain FPN layer under the standard FPN and the vertical axis represents the mean average precision of the corresponding feature layer. In Figure 1, the precision of small apples falls dramatically with the increase of the number of layers compared with large-scale and medium-scale apples. Despite pyramid shape features being computed in an element-wise addition manner to incorporate multiple-resolution feature maps, the equal proportion addition way does not take into account the characteristics of the features of each layer. The equal proportion feature fusion way degrades the performance of small apple detection. Further, the precision of small apples is poor even on the lowest layer in the standard FPN. The main reason is that the pooling-based downsampling is to filter out the small apples feature information,

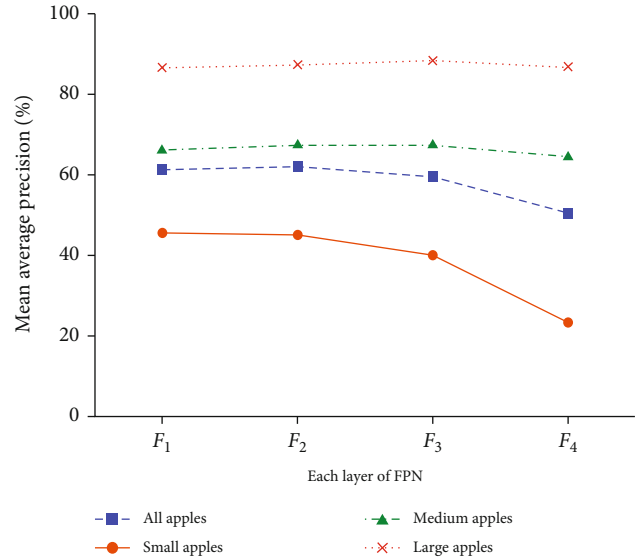


FIGURE 1: The mean average precision of all apples, small-scale, medium-scale, and large-scale apples on each layer of the standard FPN of Faster R-CNN on the GreenApple.

which leads to fewer features that can be mapped into small apples and makes it difficult to locate small apples.

To alleviate the information imbalance of multiscale objects between each layer of FPN and difficulty of locating small green apples, we propose a balanced feature pyramid network, namely, BFP Net, which can balance information from multiscale apples and the feature characteristic. Specifically, we design a weight-like feature pyramid fusion to balance the small apple information in each feature fusion structure of the vanilla FPN. Meanwhile, a feature of high resolution before pooling in the backbone is first introduced as an extended feature embedded into the lowest layer of FPN. Here, this new feature is different from the extended feature in EFPN [26]. The EFPN extended feature comes from the output of the second stage of ResNet/ResNeXt without pooling layer while the new extended feature is the input of pooling layer. Then, we devise a decoupled-aggregated module with a feature filter to alleviate the problem of difficulty in locating and apply to the new extended layer of FPN. The dual-path decoupled module decouples a feature into a spatial-aware feature and a content-aware feature to relieve the problem of difficulty in locating due to small pixels occupied by small apples and the closeness of color green apples between the fruit skin and the background. Note that the spatial-aware feature is utilized to complement the spatial information into a content-based feature map. An aggregated architecture is applied to incorporate spatial-aware and content-aware information in a point-wise addition manner. One significant advantage of the decoupled-aggregated module is that a feature with high resolution can contain both the semantic information and the location information of the small green apples. To make the consistent distribution of extended layer feature and the lowest feature of standard FPN, a Kullback-Leibler distillation loss is introduced to transfer advantage knowledge from the teacher model to the student model.

We evaluate our method and other state-of-the-art methods on the dataset benchmark of GreenApple and MinneApple, where the performance of each component is evaluated on the GreenApple. The experimental results illustrate that our method is slightly better than other state-of-the-art methods on the GreenApple and MinneApple. In addition, our method has a strong generalization performance on the generic object detection dataset of Pascal VOC.

For clarity, the main contributions of our work can be summarized as follows:

- (i) The balanced feature pyramid network is proposed to improve the performance of small apple detection
- (ii) A weight-like feature fusion structure is designed to balance feature information of small apples on different layers of standard FPN
- (iii) A decoupled-aggregated module with a feature filter is devised to decouple and aggregate features on the new extended layer differing from the EFPN, where the decoupled structure is applied to remedy difficulty of locating. In addition, a Kullback-Leibler distillation loss is applied to transfer advantage knowledge from the teacher model to the student model
- (iv) Experiments demonstrate that our approach is slightly better than other state-of-the-art detectors on the GreenApple and MinneApple and our method has a good generalization ability on the Pascal VOC

The organizational structure of this article is as follows: section 2 is related work that introduces the research status of small object detection, small fruit detection, and feature pyramid network and its variants. Section 3 elaborates on our method which mainly contains the overview architecture of our method and the details of each component. Experiments are conducted to validate our method in section 4, including evaluation metrics, experimental setting, ablation studies, experiments results, visualization, and qualitative analysis. The discussion and conclusion are presented in section 5.

2. Related Work

2.1. Small Object Detection. Due to the small objects with low shares in many datasets, the mean average precision of small objects had poor performance. Data enhancement, feature superresolution, and multiscale features are used to capture adequate small object information. To increase the small object’s resolution that is similar enough to real large objects, Li et al. [29] proposed a perceptual generative adversarial network by narrowing the representation differences between small objects and large objects to alleviate the above problems. Noh et al. [18] designed a new feature-level super-resolution method with a supervisor signal to reduce the loss of the feature generator. However, simulated features are generated with the help of GAN, and the differences between features from the generator and the original feature maps

affect the detection performance. Although the GAN-based methods can generate high-resolution images, the artifact features and fabricate fake textures lead to false positives. Moreover, data augmentation is also applied to optimize small object detection performance. Kisantal et al. [30] analyzed the reason for low performance and proposed an augmentation type by copying-pasting small objects multiple times to improve performance. Cui et al. [19] proposed a context-aware module to fuse multiple scale feature maps and the feature receptive field with the help of dilated convolution. Although some achievements have been made in the detection of small targets, the further research about the small object is still needed compared with the precision of medium and large objects.

2.2. Small Fruit Detection. Recently, small fruit detection algorithms have achieved remarkable achievements. A series of object detection methods based on deep learning containing single-stage detectors and two-stage detectors are transferred into fruit detection. Hussain et al. [31] proposed a deep convolutional neural network to automatically distinguish similar fruits and vegetables with difficult real-world scenarios. To detect and distinguish the graspable and ungraspable apples, an improved YOLOv5 was devised via BottleneckCSP, SE module, and bonding fusion mode of feature maps by Yan et al. [32]. These above methods are mainly aimed at the target fruits with a large color difference between the target fruits and the background such as orange or red. For the target fruits with similar colors between the fruit skin and the background, the detection precision is slightly lower. Therefore, the precision of small green fruits requires further improvement, which attracts the attention of scholars from home and abroad. Jia et al. [33] proposed an optimized Mask R-CNN to detect green overlapped apples. With the emergence of the attention mechanism and Transformer [34], the self-attention mechanism brings new inspirations to fruit detection. To detect immature/mature apples, the canopy-attention-YOLOv4 [35] was designed by introducing a convolutional block attention module to the generic YOLOv4 detector. A green pepper detection method [36] was proposed based on YOLOv4-tiny, which incorporates an attention mechanism and the idea of multi-scale prediction. However, Transformer-based methods require high memory and computing resources [37], which cannot meet the requirements of lightweight in fruit detection.

2.3. Feature Pyramid Network and Its Variants. Feature pyramid network was first proposed to incorporate shallow-level features and high-level semantic information. The shallow-level features contain rich texture information, edge information, etc. and the high-level features involve a lot of semantic information. To enhance the information flow, Liu et al. [38] proposed a path aggregation network by bottom-up feature fusion structure to enhance all feature hierarchy with accurate localization signals in lower layers. Tan et al. [39] designed a weighted bidirectional feature pyramid network to fast and ease incorporation of multiplescale feature maps and also devised a compound scaling method for backbone, feature network, and box/class prediction networks. Moreover, a recursive feature pyramid architecture is designed by Qiao et al. [40] to incorporate extra feedback

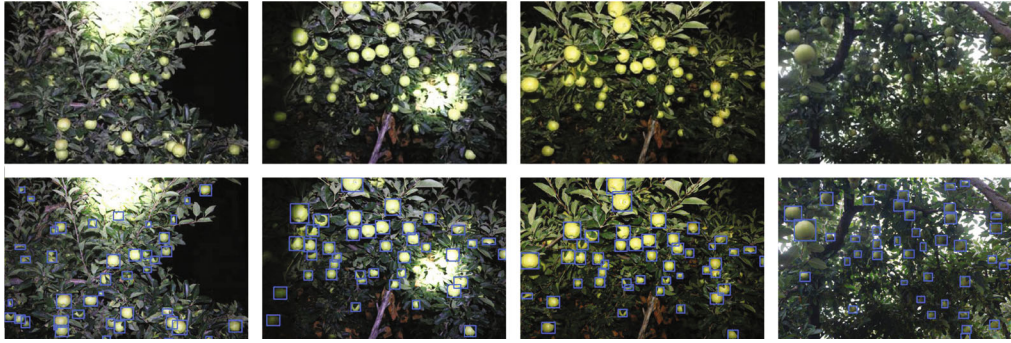


FIGURE 2: Some original images and notated results on the GreenApple dataset. The top is original images; the bottom is notated results.

connections from FPNs into the bottom-up backbone layers. These feature fusion structures are mainly fused by extracting feature maps of different layers in the backbone network. However, the feature fusion structure does not take into account the characteristics of its fused features. Specifically, taking small apple detection as an example, the information that can be mapped to small-scale targets gradually decreases or even disappears with the increase of the network layer. This is not conducive to the detection of small-scale apples. Therefore, considering the imbalance of feature information at different sizes, we will alleviate the information imbalance of small apples from two perspectives: balancing information of different scales in the multiresolution feature layers of FPN and constructing high-resolution feature mapping to locate apples.

3. Materials and Methods

In this section, the materials are introduced including three dataset benchmarks of GreenApple, MinneApple, and Pascal VOC first. Then, the overview of the balanced feature pyramid network is given. Next, the three components of our method are elaborated as follows.

3.1. Materials. Experiments are conducted on the GreenApple, MinneApple [41], and Pascal VOC [42]. Note that the GreenApple dataset is constructed for small green fruit detection and segmentation. MinneApple is a public dataset for small fruit detection and segmentation to validate the effectiveness of our proposed method. The GreenApple dataset is described in detail as follows. Pascal VOC is also a public dataset for general object classification and detection to validate the generalization.

GreenApple is a new green apple dataset, named GreenApple, containing 1361 images and 7137 apple instances. The image acquisition device is a Canon EOS (Electro-Optical System) 80D SLR (Single-Lens Reflex) camera from 8:00 to 22:00. The collected images are stored in JPEG format and saved with a resolution of 6000×4000 pixels. The location of the image collection is the Longwangshan apple production base of the Fushan District, Yantai City, Shandong Province. The latitude and longitude coordinates of the image collection site are $121^{\circ}3'$ east longitude and $37^{\circ}4'$ north latitude. The apple category is Gala. All images from multiple perspectives with varying illumination and weather

conditions are collected. These images cover multiple fruit patterns such as single apple without occlusions, branches or leaf occlusion, and overlapped fruits.

The boundaries of the fruit skin and the background are labeled using LabelMe [43]. These images are resized to 600×400 pixels and normalized by the mean and variance of all images in our dataset. The original images and notated results are presented in Figure 2. Statistically, according to the pixels occupied by each apple, the percentage of the small, medium, and large green apples of GreenApple dataset are 43.4%, 38.7%, and 17.8%, respectively. Lastly, the GreenApple Dataset is randomly divided into a training set and a test set based on the ratio of 7:3. The number and proportion of apple for small, medium, and large scales, as well as the total number of images and instances of apple, are introduced in Table 1.

MinneApple [41] is a public apple dataset for apple detection and segmentation. It contains a training set publicly and a test set privately. The training set contains 670 images and 28183 apple instances to train and test the method. We randomly divide training into a training set and a test set by the proportion of 9:1. According to the apple area occupied by pixel number, the division results are presented in Table 2. Note that this dataset does not contain large-scale apples.

Pascal VOC [42, 44] is a public dataset for classification and detection. It consists of Pascal VOC 2007 and Pascal VOC 2012, containing 20 classes. Pascal VOC 2007 and Pascal VOC 2012 trainval sets are considered as the training set, containing 16551 images. The test set of Pascal VOC 2007 is viewed as the test set, including 4952 images to assess our method. The 47223 boxes and 14976 boxes are included in the training set and test set, respectively. Pascal VOC is similar to the GreenApple dataset containing small-, medium-, and large-scale objects. The number of bounding boxes with small, medium, and large scales in Pascal VOC 2007 and 2012 are counted as follows in Table 3.

3.2. Overview of BFP Net. Typical FPN incorporates shallow-level feature maps and high-level semantic information in an element-by-element addition manner through the top-down architecture with lateral connections. For multiresolution features from different layers of FPN, the information characteristics between features are ignored. Specifically, shallow features contain the information mapped into all-scale

TABLE 1: The division results of GreenApple dataset. Note that small apples with area $< 32^2$, medium apples with $32^2 < \text{area} < 96^2$, and large apples with area $> 96^2$ are counted.

| Area | Small | 32^2 | Medium | 96^2 | Large | Total apples | Total images |
|----------|------------|--------|------------|--------|------------|--------------|--------------|
| Training | 2079/42.0% | 1 | 1973/39.9% | 0 | 890/18.0% | 4943 | 953 |
| Test | 1022/46.5% | 1 | 787/35.9% | 0 | 384/17.5% | 2194 | 408 |
| Total | 3101/43.4% | 2 | 2760/38.7% | 0 | 1274/17.8% | 7137 | 1361 |

TABLE 2: The division results of MinneApple dataset. Note that small apples with area $< 32^2$, medium apples with $32^2 < \text{area} < 96^2$, and large apples with area $> 96^2$ are counted.

| Area | Small | 32^2 | Medium | 96^2 | Large | Total apples | Total images |
|----------|-------------|--------|------------|--------|-------|--------------|--------------|
| Training | 22306/88.3% | 0 | 2943/11.7% | 0 | 0 | 25249 | 603 |
| Test | 2593/88.4% | 0 | 341/11.6% | 0 | 0 | 2934 | 67 |
| Total | 24899/88.3% | 0 | 3284/11.7% | 0 | 0 | 28183 | 670 |

TABLE 3: The division results of Pascal VOC. Note that small objects with area $< 32^2$, medium objects with $32^2 < \text{area} < 96^2$, and large objects with area $> 96^2$ are counted.

| Area | Small | 32^2 | Medium | 96^2 | Large | Total boxes | Total images |
|----------|------------|--------|-------------|--------|-------------|-------------|--------------|
| Training | 5300/11.2% | 5 | 14525/30.7% | 4 | 27389/57.9% | 47223 | 16551 |
| Test | 1661/11.1% | 3 | 5077/33.9% | 2 | 8233/54.9% | 14976 | 4952 |
| Total | 6961/11.1% | 8 | 19602/31.5% | 6 | 35622/57.2% | 62199 | 21503 |

apples, while high-level features mainly are mapped into apples of medium and large scales. In other words, there is an information imbalance in feature mapping for multiple-scale fruits. In addition, due to small physical sizes in the real world and small green apples with the closeness of color between the fruit skin and the background under the orchard environment, these greatly increase the locating and recognition difficulty of small apples.

To relieve the imbalance problem and difficulty of locating of small apple detection in an unstructured complex orchard environment, a balanced feature pyramid network is proposed with a weight-like feature fusion architecture and a decoupled-aggregated module with a Kullback-Leibler distillation loss for small apple detection, as shown in Figure 3.

Specifically, assuming that the feature maps of the original FPN structure are denoted by $C_i (i = 1, 2, 3, 4)$ in the bottom-up structure and $F_i (i = 1, 2, 3, 4)$ in the top-down structure, whose weight parameters are represented by k and t , respectively. Meanwhile, a new extended feature map C_0 with high resolution is introduced from the output of conv1 in the ResNet50, as shown in Table 4. This extended feature map differs from the extended feature of EFPN, and C_0 contains more spatial and content information due to its high resolution. To balance spatial information and content information and alleviate the problem of difficulty of locating small apple detection, a decoupled-aggregated module is designed on this extended layer. To reduce the difference between an extended feature F_0 and the feature maps F_1, F_1 and F_0 are viewed as a teacher model and a student model, respectively. With the help of

the distillation mechanism, a Kullback-Leibler distillation loss is introduced to transfer knowledge. The overview of our proposed method is shown in Figure 3.

3.3. Weight-Like Feature Fusion Architecture. The mean average precision of small object detection is about half of that of large objects with state-of-the-art models in recent years [12]. Small green apples also have similar problems. By investigating the characteristics of the different level feature maps, it can be discovered that shallow features have ample information about small apples. However, an element-wise addition feature fusion with the proportion of 1:1 at different levels ignores the imbalance problem of multiscale apple information on the different FPN feature layers.

Enlightened by the different sensitivities of feature maps on different layers for different scales of apples, a weight-like feature fusion architecture is devised in Figure 4. Two hyper-parameters are introduced, weight parameter k on the lateral connection and parameter t on the top-down pathway, respectively. For low-level feature maps, the weight-based feature maps on the bottom-up architecture can be applied to incorporate more features of all-scale apples, especially small apples. While the high-level feature maps, the weight-like feature maps on the top-down structure can extract more semantic information. As shown in Figure 4, $k \times$ denotes k times feature and $t \times$ represents t times feature.

The four features as input are fed to a detector in the original FPN. Assuming that the two features of the high layer are termed as high layer features on the standard FPN and the two features of the shallow layer are treated

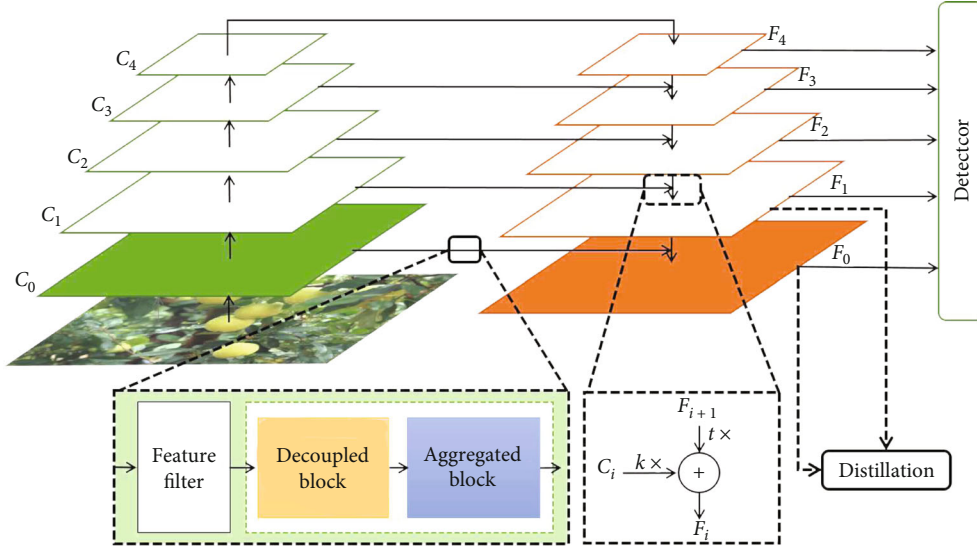


FIGURE 3: The overview of balanced feature pyramid network (BFP Net). The $C_i (i=0, 1, 2, 3, 4)$ refers to the input of FPN from the backbone of ResNet50 and the $F_i (i=0, 1, 2, 3, 4)$ denotes the output of FPN. The top 4 layers of FPN are vanilla FPN layers. To balance the multiscale information on different layers, an element-wise addition feature fusion is replaced by a weight-like element-wise addition feature fusion architecture in the vanilla FPN. A new extended layer before pooling is embedded into vanilla FPN to remain a high-resolution feature. Here C_0 and F_0 represent the input and an output of the extended layer. A decoupled-aggregated module with a feature filter is devised to extract content-aware and spatial-aware information. A Kullback-Leibler distillation mechanism is applied to transfer favorable knowledge from F_1 to F_0 . Finally, the pyramid-shaped features (F_0, F_1, F_2, F_3 , and F_4) are fed to the detector to detect.

TABLE 4: The layer structure of Faster R-CNN, EFPN, and BFP Net (ours). ResNet50 is applied as the backbone of these methods.

| Method | Faster R-CNN [27] | EFPN [26] | BFP Net (ours) |
|------------|-------------------------------------|-------------------------------------|-------------------------------------|
| Input size | (600, 400) | (600, 400) | (600, 400) |
| Stage 1 | $7 \times 7, 64, s=2, \text{pad}=3$ | $7 \times 7, 64, s=2, \text{pad}=3$ | $7 \times 7, 64, s=2, \text{pad}=3$ |
| Output | — | — | $C_0: (300, 200)$ |
| Stage 2 | 3×3 max pool, $s=2$ | — | 3×3 max pool, $s=2$ |
| | Residual block ($\times 3$) | Residual block ($\times 3$) | Residual block ($\times 3$) |
| Output | $C_2: (150, 100)$ | $C'_2: (300, 200)$ | $C_1: (150, 100)$ |

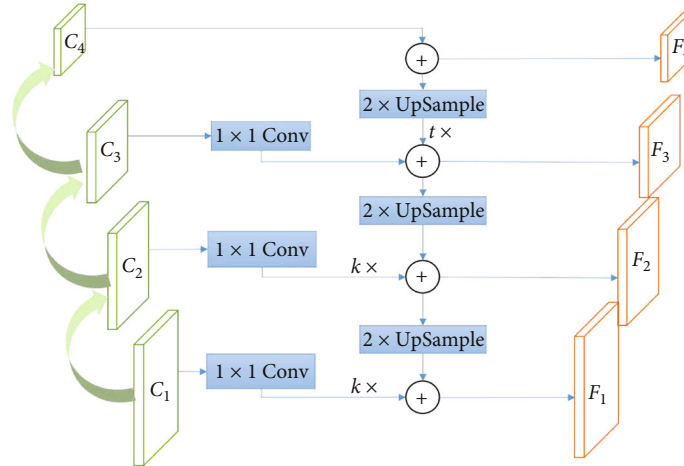


FIGURE 4: The diagram of weight-like.

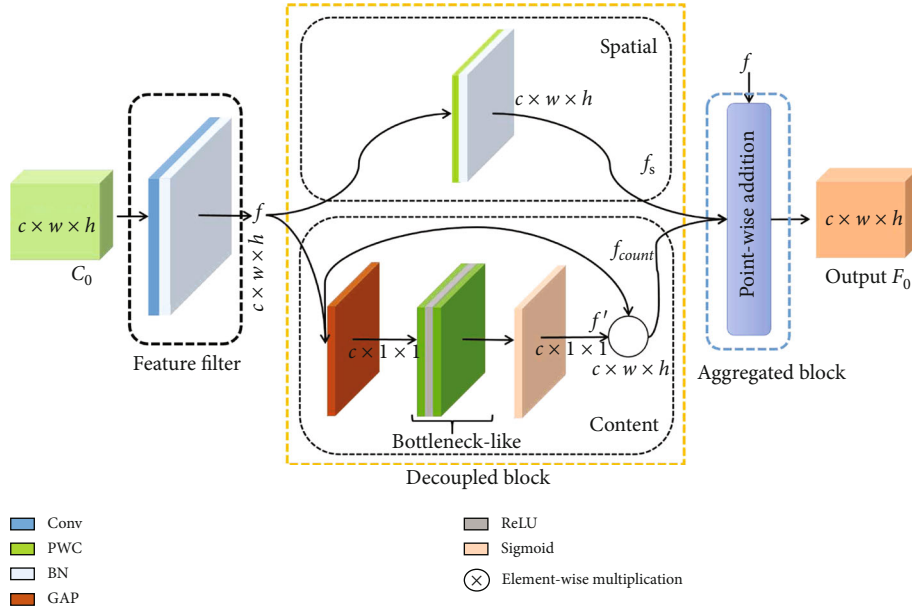


FIGURE 5: Decoupled-aggregated module.

as shallow layer features. To preserve more feature information, the standard weight of the feature is set to 1. Mathematically, a feature map F_i can be described as

$$F_i = \begin{cases} k \cdot C_i + \uparrow_{2 \times} F_{i+1} & 0 < i < 3, \\ C_i + t \cdot \uparrow_{2 \times} F_{i+1} & 3 \leq i < 4, \end{cases} \quad (1)$$

where k is the weight hyperparameter of the lateral connection in the shallow level, C_i represents the feature map of i -th lateral connection, $\uparrow_{2 \times}$ denotes double upsampling by nearest-neighbor interpolation, and t is a weight parameter of the top-down architecture on the high layer. We take the feature layer F_2 of the standard FPN as an example on the benchmark dataset of GreenApple. To facilitate the experiment, we set k to 2. The feature F_2 is obtained by adding double the feature layer C_2 and F_3 after double upsampling by nearest-neighbor interpolation in an element-wise addition manner.

3.4. Decoupled-Aggregated Module. The original feature pyramid network pays attention to multiple-resolution feature fusion via lateral connections with a top-down structure. However, different levels of features on the FPN architecture have different sensitivities for small, medium, and large apples. One reason is that only high-level features with low resolution are obtained and shallow information is ignored via a series of convolution and pooling operations. Another reason is that the small apples occupy fewer pixels, and small green apples in the growth stage have the characteristic of similar color between the fruit skin and the background. These greatly increase the locating difficulty of small green apples with area $< 32^2$. The FPN-based feature fusion architecture can incorporate shallow-level and high-level semantic information, ignoring spatial location information.

The decoupled-aggregated module with a feature filter is composed of a decoupled block and an aggregated block. The left is a feature filter to refine the feature. The middle is a decoupled block, where the spatial-aware branch of the top is applied to complement spatial location information and the content-aware branch of the bottom is adopted to capture global content information. The right is an aggregated block to incorporate features. Note that PWC refers to point-wise convolution.

To address the information imbalance problem between small apples and large apples, a new extended layer is also introduced to the FPN before the pooling operation, enlightened by the extended feature pyramid network [26]. The source of this extended feature is shown in Table 4. This extended feature layer is obtained after a convolution operation and before pooling in the ResNet50. However, this new extended feature not only contains abundant multiple-scale apple information but includes some noises. These noises would negatively impact the apple recognition and detection. Therefore, a feature filter is first introduced to discard disturbing noises and adjust the number of channels.

Inspired by the decoupling principle [45, 46], a decoupled-aggregated module is devised to complement spatial information. The extended feature map can be decoupled into a content-aware feature and a spatial-aware feature. Point-wise addition is applied to aggregate features. This architecture synthesizes spatial information and content information of images. The overview structure of the decoupled-aggregated feature block is shown in Figure 5. The decoupled-aggregated module is elaborated on as follows.

3.4.1. Feature Filter. For the feature filter, a convolution operation with a batch normalization on the extended layer is utilized to reduce noise interference. Specifically, assuming

that the input and output of the extended layer are C_0 and F_0 , a convolution operation with a kernel size of 3×3 , a stride of 2, and a padding of 1 are applied to filter interference noises. Next, a batch normalization is utilized to smooth features and obtain feature f .

3.4.2. Decoupled and Aggregated Block. For the decoupled-aggregated block, this module consists of a decoupled block and an aggregated block. The decoupled block can decompose a feature into two features, namely, a spatial-aware feature and a content-aware feature. Then, the aggregated block incorporates features in a point-wise addition manner. The decoupled module and aggregated module are described in detail as follows.

For the spatial-aware branch, a point-wise convolution (PWC) with a normalization layer is adopted to capture a feature map f_s with high resolution that can preserve spatial location information through two convolution operations with the normalization layer. Although the PWC ignores some spatial information [47], a high-resolution feature without pooling operation still preserves abundant spatial information [48].

For the content-aware branch, enlightened by the squeeze and excitation idea in SENet [49] and attention mechanism [34], a global average pooling is first used to transform the feature shape from $c \times h \times w$ to $c \times 1 \times 1$, which enlarges the receptive field of the feature map from local to global. In other words, this operation is equivalent to an attention mechanism that can capture abundant global contextual information, which breaks the limitation that convolution only extracts local information. Next, a bottleneck-like architecture is utilized to extract features from the global receptive field. The bottleneck-like structure consists of two point-wise convolutions and a ReLU. Specifically, a point-wise convolution with a dimension reduction factor γ is applied to obtain feature information. The dimension reduction factor is applied to squeeze the channel number. A ReLU is an activation function that prevents the gradient from disappearing. A point-wise convolution with channel c is utilized to stimulate more feature information, regulate channel number, and enhance feature representation capacity. A sigmoid function is adopted to rescale feature maps. Finally, a new feature map f' can be obtained. The f' is multiplied with the input feature f of the content-aware branch to obtain a content-based feature map f_{cont} .

For the aggregated block, a point-wise addition is applied to aggregate features with spatial-aware, content-aware, and output of a residual structure on the decoupled block. Finally, the output of aggregated block can be obtained to feed to the detector.

Mathematically, the decoupled architecture can be described as follows. The input of the decoupled block is defined as f . f_s and f_{cont} are viewed as outputs of the spatial-aware branch and content-aware branch. The output F_{DA} of decoupled-aggregated can be formulated as

$$F_{\text{DA}} = f + f_s + f_{\text{cont}}. \quad (2)$$

The feature map F_0 can be described as

$$F_0 = F_{\text{DA}} + F_1, \quad (3)$$

where F_1 is the feature map from the FPN; f_s is obtained through a point-wise convolution operation with a kernel size of 1×1 and a normalization operation. f_{cont} is captured by the squeezing operation $\text{Sq}(\cdot)$ based on a global average pooling and exciting operation $\text{Ex}(\cdot)$ based on the bottleneck-like structure and a sigmoid. Specifically, $\text{Sq}(\cdot)$ is represented as follows:

$$z_c = \text{Sq}(f) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W f_c(i, j), \quad (4)$$

where H , W , and C refer to height, width, and channel number, respectively. $f_c(i, j)$ denotes the feature value of the c -th channel, i -th row, and j -th column. z_c is the global average pooling operation of the c -th channel.

$\text{Ex}(\cdot)$ is formulated as

$$f' = \text{Ex}(z, w_1, w_2) = \sigma(w_2 \delta(w_1, z)), \quad (5)$$

where $\sigma(\cdot)$ denotes a sigmoid function, $\delta(\cdot)$ refers to activation function of ReLU, w_i refers to the weight of point-wise convolution operation of kernel 1×1 in the content-aware branch, respectively, and where $w_1 \in R^{(\gamma C \times C)}$ refers to the weight of the first convolution and $w_2 \in R^{(C \times \gamma C)}$ is the weight of the second convolution. z represents the output of $\text{Sq}(f)$. Finally, the output of the content-aware branch is calculated as follows:

$$f_{\text{cont}} = f \otimes f', \quad (6)$$

where \otimes denotes the element-wise multiplication operator.

3.5. Kullback-Leibler Distillation Loss. Although the feature map on the extended layer has been filtered and underwent decoupled and aggregated operations, there will still be a few noise information. To make the distribution more consistent between feature map F_1 and F_0 , a Kullback-Leibler (KL) divergence loss function is introduced to relieve the loss between features. The KL divergence D_{KL} can be described as

$$D_{\text{KL}}(p||q) = \sum_{i=0}^N p(x_i) \left(\log \frac{p(x_i)}{q(x_i)} \right), \quad (7)$$

where p and q are probability distribution of F_1 and F_0 . The N refers to the product of batch size, channel, width, and height of feature F_1 or F_0 . The x_i refers to the feature value of i -th position. D_{KL} calculates the difference between probability p and probability q .

The smaller the divergence value, the closer the relationship between probability q and probability p is. That is, the probability distribution of F_0 is closer to that of F_1 . p and q can be described as

$$\begin{aligned} p &= \text{softmax}(F_1), \\ q &= \log(\text{softmax}(F_0)). \end{aligned} \quad (8)$$

where $\text{softmax}(\cdot)$ can be formulated as follows:

$$\text{softmax}(x_i) = \frac{\exp(x_i)}{\sum_{c=1}^C \exp(x_c)}, \quad (9)$$

where x_i represents the i -th feature value of F_1 or F_0 ; the c refers to the c -th channel where x_i is located. C is the channel number. The $\exp(\cdot)$ function is the x power of e .

However, the KL loss can transfer knowledge between F_0 and F_1 via the back-propagation algorithm. Therefore, to prevent the adverse interference of the KL loss to our method, the distillation mechanism is introduced as a supervisor signal to supervise KL loss. Given that the distillation temperature T , the feature map F_1 , and an extended feature map F_0 are regarded as the teacher and student models, respectively. Mathematically, $\text{softmax}_{\text{KD}}(\cdot)$ can be optimized as follows:

$$\text{softmax}_{\text{KD}}(x_i, T) = \frac{\exp(x_i/T)}{\sum_{c=1}^C \exp(x_c/T)}, \quad (10)$$

where x_i represents the i -th feature value of F_1 or F_0 ; the c refers to the c -th channel. C is the channel number.

The final KL distillation loss can be formulated as follows:

$$L_{\text{KL-KD}} = D_{\text{KL-KD}} \cdot T^2, \quad (11)$$

where $D_{\text{KL-KD}}$ represents KL loss based on equation (7) and equation (10).

In this paper, we adopt the Faster R-CNN [27] as the baseline, as an example. The total loss Loss can be described as follows:

$$\begin{aligned} \text{Loss} &= L_{\text{frCNN}} + L_{\text{KL-KD}}, \\ L_{\text{frCNN}}(p_i, t_i) &= \frac{1}{N_{\text{cls}}} \sum_i L_{\text{cls}}(p_i, p_i^*) + \varepsilon \frac{1}{N_{\text{reg}}} \sum_i p_i^* L_{\text{reg}}(t_i, t_i^*), \end{aligned} \quad (12)$$

where L_{cls} represents the classification loss of cross entropy and L_{reg} refers to the regression loss of smooth L1 loss. p_i is the predicted probability of anchor i . N_{cls} is the minibatch size and N_{reg} refers to the number of anchor locations. ε represents a balancing weight. If the anchor is positive, the ground-truth label p_i^* is 1; otherwise, p_i^* is 0. t_i is a vector that represents the coordinate of the predicted bounding box. t_i^* is the vector representing the coordinate of the ground-truth bounding box.

4. Results

4.1. Evaluation Metrics. In this paper, we adopt the precision (P) and recall (R) [50] as the evaluation indexes. The P and R can be formulated on a certain intersection over union (IoU) [51] threshold as follows:

$$\begin{aligned} P &= \frac{\text{TP}}{\text{TP} + \text{FP}}, \\ R &= \frac{\text{TP}}{\text{TP} + \text{FN}}, \end{aligned} \quad (13)$$

where TP, FP, and FN refer to true positives, false positives, and false negatives, respectively. The mean average precision (AP) and mean average recall (AR) under the IoU = [0.5 : 0.05 : 0.95] are applied to evaluate the methods. To more clearly present the performance of objects at different scales, the average precision of small-scale, medium-scale, and large-scale objects are defined as AP_S , AP_M , and AP_L , respectively. The average recall of small-scale, medium-scale, and large-scale objects are set to AR_S , AR_M , and AR_L , respectively.

For the GreenApple, MinneApple, and Pascal VOC dataset, considering the proportion of small, medium, and large objects and artificial selection of a single IoU can affect the quality of the method assessment. The mean average precision (AP) and recall (AR) at IoU = [0.5 : 0.05 : 0.95] are reported. Moreover, AP_{50} and AP_{75} are also introduced to evaluate the performance of detection. In addition, we use #params to comprehensively measure the parameter quantity.

4.2. Implementation Details. Our approach and ablation experiments are implemented with a Faster R-CNN detector, where ResNet50 is applied as the backbone. All models are trained in a single NVIDIA 1080Ti GPU using PyTorch [52]. The original Faster R-CNN is first trained as the baseline. An SGD optimizer is utilized with the weight decay of $5e-4$ and a momentum of 0.9. Our method is trained based on 12 epochs. The step learning rate scheduler is applied to a learning rate adjustment strategy of our method, where the step size is set as 3. The aspect ratio size is set as [0.5, 1.0, and 2.0]. For the hyperparameters, the weight values k and t are both set to 2 to facilitate the experiment. The dimension reduction factor γ is set to 0.2. The distillation temperature T is set to 103. Further, to validate our method, experiments are conducted compared with state-of-the-art methods on the MinneApple and Pascal VOC.

4.3. Ablation Studies. Ablation experiments are conducted on the GreenApple to validate each component and analyze the contributions of each component. Due to the limitation of computer computing resources, all experiments are based on the Faster R-CNN with the backbone of ResNet50 and FPN. Models are trained and tested using the training set and test set of GreenApple. Experimental results are presented in Table 5. It can be observed that the AP_S and AR_S of baseline on small apples are 45.7% and 55.1% under the IoU = [0.5 : 0.05 : 0.95], respectively. Further, the precision and recall of our method on small apple detection are 47.0% and 57.2%, increased by 1.3% and 2.1%, respectively. When only a weight-like feature fusion architecture is adopted, the AP_S and AR_S are 46.1% and 55.7% increased by 0.4% and 0.6% for small apple detection, respectively. When a weight-like feature fusion architecture and a

TABLE 5: Performance of each component on the GreenApple test set. Note that D-A refers to the decoupled-aggregated module.

| Weight-like | D-A | Loss | AP _S | AP _M | AP _L | AR _S | AR _M | AR _L | AP ₅₀ | AP ₇₅ | AP |
|-------------|-----|------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|------------------|------|
| | | | 45.7 | 68.3 | 88.4 | 55.1 | 74.4 | 92.1 | 85.9 | 69.3 | 62.3 |
| ✓ | | | 46.1 | 68.5 | 87.9 | 55.7 | 74.4 | 91.4 | 85.8 | 70.2 | 62.2 |
| ✓ | ✓ | | 46.7 | 68.6 | 86.5 | 56.9 | 75.1 | 90.8 | 86.8 | 69.5 | 62.2 |
| ✓ | ✓ | ✓ | 47.0 | 68.1 | 87.5 | 57.2 | 75.4 | 91.2 | 86.7 | 69.1 | 62.3 |

TABLE 6: Performance of different weights on the GreenApple test set. To simplify the experiment, we take the same value for k and t . Meanwhile, values of k and t are integers.

| k | t | AP _S | AP _M | AP _L | AR _S | AR _M | AR _L | AP ₅₀ | AP ₇₅ | AP |
|-----|-----|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|------------------|------|
| 1 | 1 | 46.8 | 68.4 | 86.3 | 56.3 | 75.0 | 90.5 | 86.1 | 68.7 | 61.9 |
| 2 | 2 | 47.0 | 68.1 | 87.5 | 57.2 | 75.4 | 91.2 | 86.7 | 69.1 | 62.3 |
| 3 | 3 | 46.2 | 68.9 | 86.8 | 56.8 | 75.6 | 91.0 | 87.2 | 70.3 | 62.7 |
| 4 | 4 | 46.2 | 67.1 | 85.5 | 56.5 | 73.8 | 89.5 | 86.4 | 68.1 | 61.2 |

TABLE 7: The performance of our method based on different branches.

| Spatial | Content | AP _S | AP _M | AP _L | AR _S | AR _M | AR _L | AP ₅₀ | AP ₇₅ | AP |
|---------|---------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|------------------|------|
| ✓ | | 46.8 | 68.6 | 87.2 | 57.6 | 75.5 | 91.1 | 86.9 | 70.0 | 62.4 |
| | ✓ | 46.3 | 68.8 | 86.2 | 56.5 | 75.5 | 90.5 | 86.8 | 69.5 | 62.2 |
| ✓ | ✓ | 47.0 | 68.1 | 87.5 | 57.2 | 75.4 | 91.2 | 86.7 | 69.1 | 62.3 |

decoupled-aggregated module are adopted at the same time, the precision and recall are 46.7% and 56.9%, increasing by 1.0% and 1.8%, respectively. To improve the stability during training, a distillation loss improves the precision by 0.3% for small apples. It can be concluded that each block can improve the precision of small apple detection.

The bigger the weight, the better? The performance of our network is evaluated for different weight parameters with a decoupled-aggregated module and a KL distillation loss, as shown in Table 6. We assume that k and t are set from 1 to 4 to evaluate the precision and recall of different scales of apple under different parameters. It is noteworthy that k and t are two different hyperparameters. In this paper, we adopt same values and take integer values to simplify the experiment to verify the validity of our method. It can be observed that AP and AP_S do not increase with the magnified weight. When k and t take 2, AP_S and AR_S get the optimal value of 47.0% and 57.2%, increased by 1.3% and 2.1% compared with the baseline of Faster R-CNN. It can be seen that our method can improve small apple detection performance in Table 6. Note that when the weight is set as 1, our method is equivalent to the weight-like feature fusion architecture. While the precision for small apple detection is improved, the precision of medium and large apple detection is slightly lower with the weight of 2. Although AP_M and AP_L are slightly smaller. Finally, the weight is empirically assigned as 2 for optimal performance considering the mean average precision of the small, medium, and large scales comprehensively. Therefore, the weight-like feature fusion structure can effectively improve the precision of small apple detection.

Importance of decoupled block: the performance of the spatial branch and content branch are tried to assess, respectively. Experiments are conducted for every branch with the weight-like feature fusion structure, feature filter, and aggregated feature module. Experimental results are presented in Table 7. Only spatial-aware branches or content-aware branches can improve the performance of our proposed model by 1.1% and 0.6%. The precision can be improved by 1.3% using two branches of the spatial-aware and content-aware branches. It can be observed that the decoupled-aggregated feature block with the single spatial feature or single content feature can improve slightly our model. Compared with the content-aware branch, the spatial-aware branch can significantly improve the precision of small apple detection. Firstly, it benefits from the fact that the input of the extended layer feature is a high-resolution feature; secondly, the output of the spatial sensing branch only goes through the two convolution operations with the normalization layer and does not go through the downsampling of pooling, which can retain more feature information that can be mapped to small apples. Thus, it can be concluded that spatial architecture can play a greater role than the content-aware branch. The spatial branch is necessary for capturing object structure information via the point-wise convolution without downsampling. Content-aware branch obtains global contextual feature information by global average pooling, which breaks the limitation that convolution only extracts features locally. In contrast, the feature decoupled block and aggregated block can effectively improve the recognition precision of small apple detection. A feature map from the spatial-aware architecture is indispensable for supplying spatial information about the target.

TABLE 8: Our proposed method and comparison results on the GreenApple test set.

| Method | AP _S | AP _M | AP _L | AR _S | AR _M | AR _L | AP ₅₀ | AP ₇₅ | AP | AR |
|--------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|------------------|------|------|
| CARAFE [54] | 43.8 | 67.5 | 86.2 | 53.0 | 72.6 | 89.4 | 85.3 | 66.8 | 60.1 | 66.4 |
| SABL [55] | 46.7 | 69.6 | 87.7 | 55.0 | 75.3 | 90.8 | 86.6 | 68.3 | 62.4 | 68.6 |
| Dynamic R-CNN [56] | 45.6 | 68.5 | 86.5 | 53.2 | 73.6 | 89.7 | 85.5 | 67.8 | 61.3 | 67.0 |
| PISA [57] | 44.3 | 68.5 | 85.1 | 54.8 | 84.0 | 88.3 | 85.6 | 67.6 | 60.9 | 67.6 |
| Mask R-CNN [58] | 46.8 | 68.6 | 86.0 | 54.5 | 74.1 | 89.1 | 86.8 | 68.7 | 61.8 | 67.6 |
| Groie [59] | 45.3 | 68.3 | 86.7 | 52.8 | 73.6 | 89.6 | 86.7 | 68.0 | 61.3 | 66.7 |
| Faster R-CNN [27] | 45.7 | 68.3 | 88.4 | 55.1 | 74.4 | 92.1 | 85.9 | 69.3 | 62.3 | 68.5 |
| BFP Net | 47.0 | 68.1 | 87.5 | 57.2 | 75.4 | 91.2 | 86.7 | 69.1 | 62.3 | 69.7 |

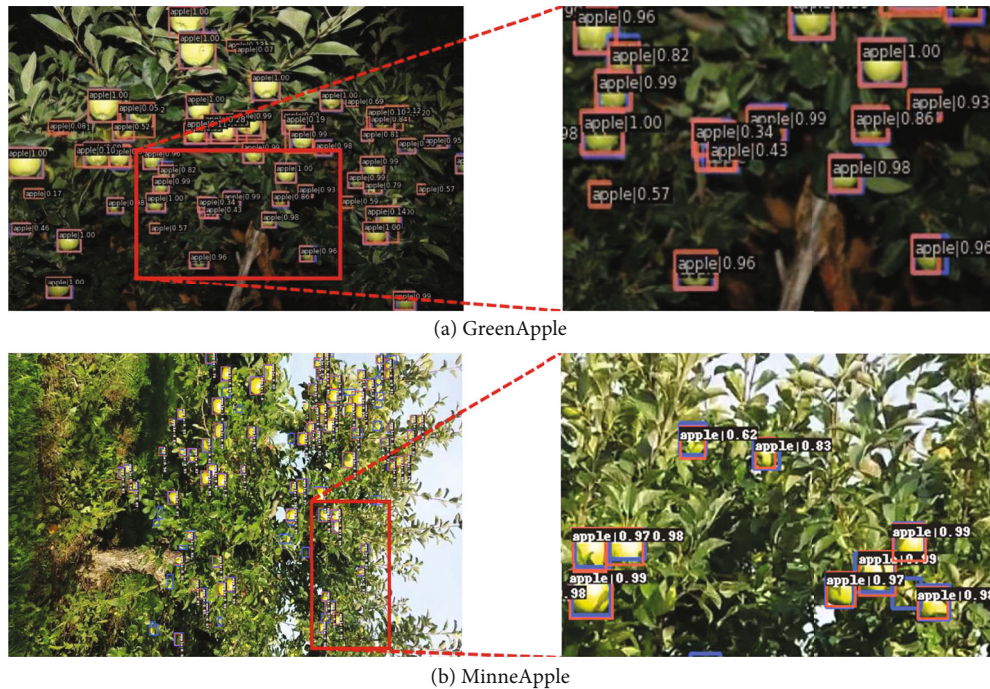


FIGURE 6: The visualization results of BFP Net on the GreenApple and MinneApple. The left images are prediction results and the right images show the detail of the red rectangular boxes. Note that we rotate the prediction results of MinneApple 90 clockwise. (a) GreenApple, (b) MinneApple.

4.4. Results on the GreenApple. Some state-of-the-art methods are compared with our method on the GreenApple and the results are shown in Table 8. All comparison models are trained and tested based on MMDetection [53] on a single NVIDIA 1080Ti GPU. The ResNet50 and FPN architecture are utilized as the backbone and neck. Precision and recall are applied as the evaluation indicators. For comparison methods, an SGD optimizer is adopted with an initial learning rate of 0.0025, a momentum of 0.9, and a weight decay of $5e-4$. $1 \times$ schedule and 12 epochs with 2 images per GPU are applied to reduce required computing resources. The step size and gamma are set to 3 and 0.33, respectively. To analyze the performance of all methods for small, medium, and large apples, the precision and recall are evaluated at different scales based on the evaluation index of $\text{IoU} = [0.5 : 0.05 : 0.95]$, $\text{IoU} = 0.5$, and $\text{IoU} = 0.75$. The dimension reduction factor γ is set to 0.2. The distillation temperature is set to 103°C .

In Table 8, experimental results are presented on the GreenApple. Compared with Faster R-CNN based on the ResNet50, AP_S and AR_S of our method can achieve 47.0% and 57.2% increased by 1.3% and 2.1%, respectively. It can be observed that BFP Net outperforms the Faster R-CNN and its variants for small apple detection. To better observe the detection performance of BFP Net, we present the prediction results of an image, as shown in Figure 6(a). The figure on the left shows the overall prediction results of the image and the figure on the right shows the local results corresponding to the red rectangle. Therefore, experimental results illuminate that BFP Net can improve the precision of small apple detection. The AP_S and AP on the GreenApple test set can reach 47.0% and 62.3%, respectively.

In addition, CARAFE, SABL, Dynamic R-CNN, PISA, Mask R-CNN, and Groie are applied to compare the effectiveness of BFP Net for small apple detection. The AP_S of our method is increased by 3.2%, 0.3%, 1.4%, 2.7%, 0.2%,

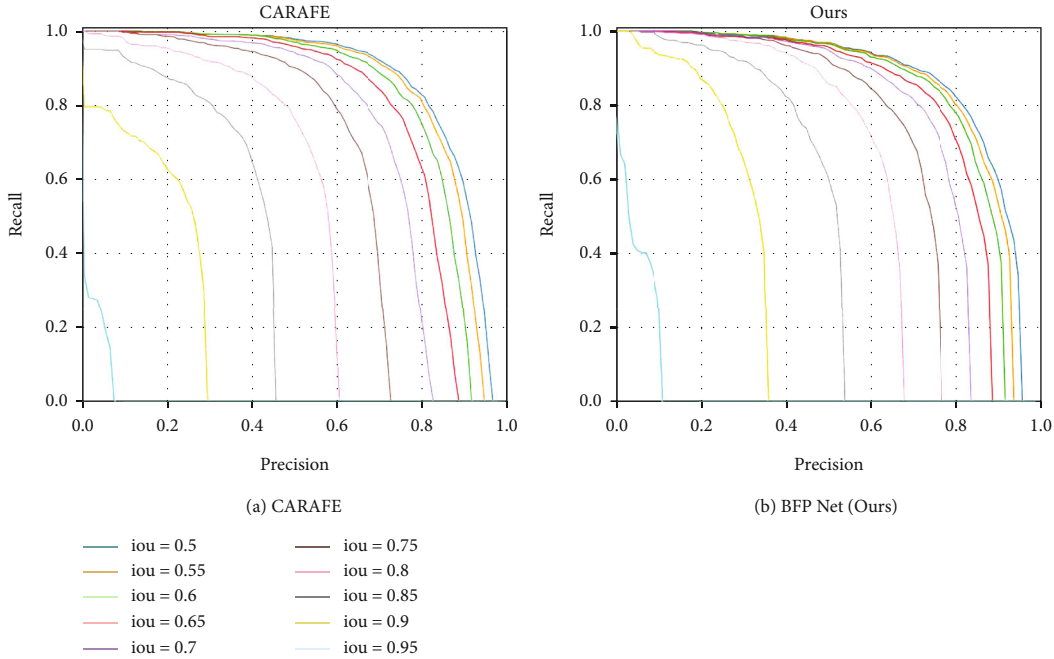


FIGURE 7: P-R curve of our model and CARAFE based on all thresholds of [0.5 and 0.95] on the GreenApple. (a) CARAFE, (b) ours.

and 1.7% compared with CARAFE [54], SABL [55], Dynamic R-CNN [56], PISA [57], Mask R-CNN [58], and Groie [59]. Meanwhile, the AR_S can be better than other state-of-the-art methods based on small apples, i.e., 57.2% vs 53.0% and 57.2% vs 54.8%, respectively. Therefore, our method is better than other state-of-the-art models for small apple detection.

In Figure 7, the precision-recall (P-R) curve of CARAFE and BFP Net are shown on the GreenApple test set. The area enclosed by the P-R curve and the two coordinate axes can be applied to evaluate the quality of a model. When the P-R curve of one model is surrounded by the precision-recall curve of another model, the latter model is better than the former model [50]. The P-R curves are drawn based on $IoU = [0.5 : 0.05 : 0.95]$ about CARAFE and our method firstly. It can be found that the area enclosed by P-R curves and two coordinate axes under the certain IoU threshold of our method is larger than that of CARAFE. It can be concluded that with the increase of IoU, the effect of the model is more significant. Compared with CARAFE, our performance can be better than other methods when the IoU threshold is greater than 0.6.

To more clearly show the effectiveness of our proposed model at a specific threshold, 0.75 is adopted for the IoU threshold as an example to draw the P-R curves of BFP Net and CARAFE for different scales of fruits, as shown in Figure 8. Blue and orange lines are the effects of BFP Net and CARAFE. From this picture, BFP Net has a significant effect on apple detection, especially on small and medium apple detection.

To evaluate the model more comprehensively, we also use the #params to evaluate the performance of our method. The evaluation results of parameters are shown in Table 9. In Table 9, it can be observed that the #params of BFP Net

is 42.55 M. Compared with the baseline of Faster R-CNN, the #params of our method increased by 1.43 M. Therefore, we can conclude that the BFP Net can improve the detection performance of small apple detection without increasing a large number of parameters.

4.5. Results on the MinneApple. To validate our method, MinneApple is applied to assess BFP Net and other state-of-the-art methods. The results are presented in Table 10. It can be found that our method can outperform other state-of-the-art methods. The AP_S and AR_S of our method are 42.2% and 48.2%, increased by 1.0% and 1.3%, respectively, compared with the baseline of Faster R-CNN. Note that AP_L and AR_L are -1 because there is no large-scale apples. A prediction image is presented in Figure 6(b). The orange and blue bounding boxes represent the prediction and ground-truth bounding boxes, respectively. By comprehensively analyzing the prediction results and visualization results, we can conclude that our method is applicable to small apple detection in a real orchard environment.

We also compared BFP Net with many state-of-the-art models, and the results are presented in Table 10. It can be observed that the AP_S is increased by 1.3%, 1.7%, 3.6%, 0.3%, 1.3%, and 1.2%, compared with CARAFE [54], SABL [55], Dynamic R-CNN [56], PISA [57], Mask R-CNN [58], and Groie [59]. Meanwhile, AP_M , AP_L , and AP of BFP Net outperform some state-of-the-art methods in Table 10. Similarly, AR_S , AR_M , AR_L , and AR are better than the comparison models. Overall, BFP Net can outperform some state-of-the-art methods on the MinneApple.

4.6. Results on the Pascal VOC. We also verified the generalization ability by the generic object detection dataset of

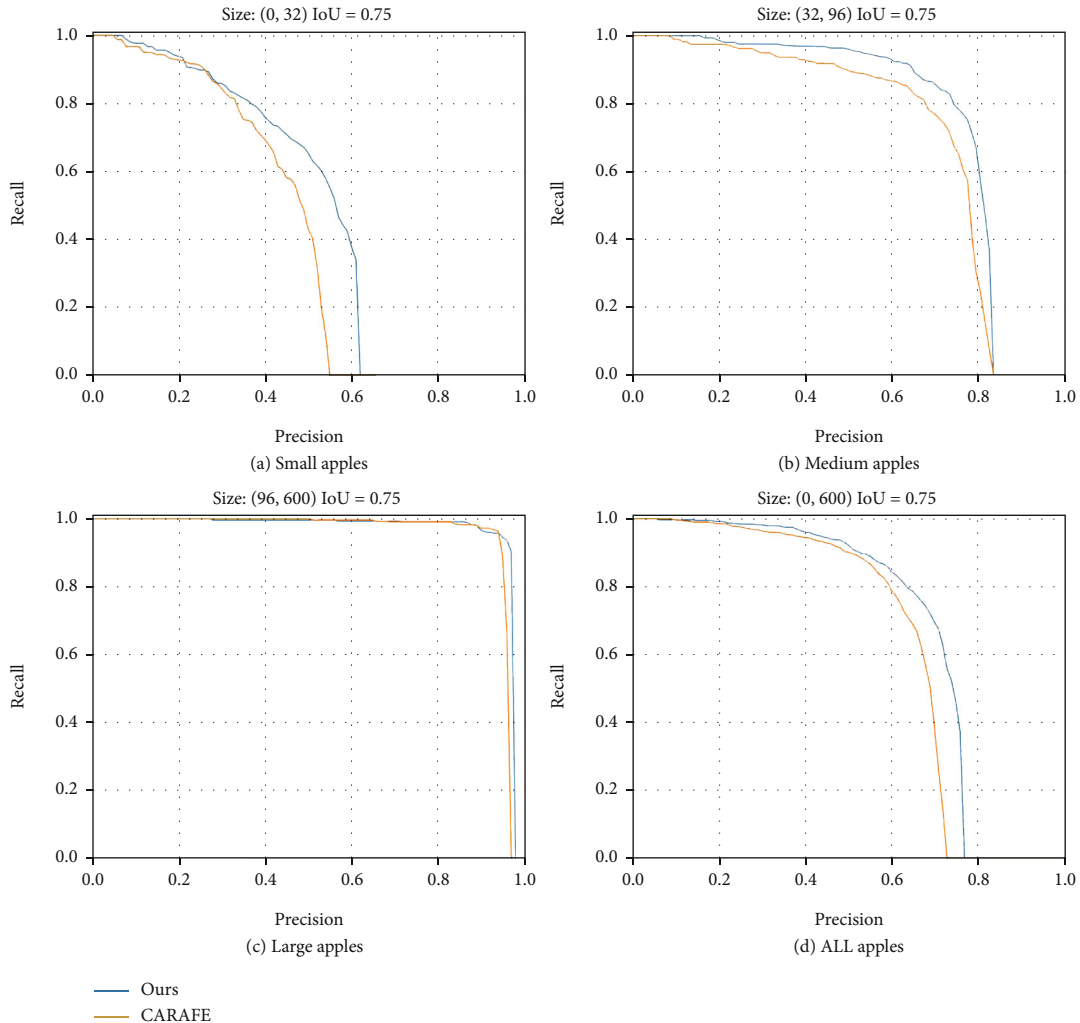


FIGURE 8: Comparison precision and recall of BFP Net and CARAFE with IoU = 0.75 on the GreenApple test set. (a) Small apples, (b) medium apples, (c) large apples, and (d) all apples.

TABLE 9: Analysis of model parameters.

| Method | Input size | #params |
|--------------------|---------------|---------|
| CARAFE [54] | (3, 600, 400) | 46.73 M |
| SABL [55] | (3, 600, 400) | 41.91 M |
| Dynamic R-CNN [56] | (3, 600, 400) | 41.12 M |
| PISA [57] | (3, 600, 400) | 41.12 M |
| Mask R-CNN [58] | (3, 600, 400) | 43.75 M |
| Groeie [59] | (3, 600, 400) | 43.02 M |
| Faster R-CNN [27] | (3, 600, 400) | 41.12 M |
| BFP Net | (3, 600, 400) | 42.55 M |

Pascal VOC. The experimental results are presented in Table 11. In Table 11, experimental results of our proposed method and baseline show that the AP_S and AR_S of our method can achieve 35.6% and 48.9%, increased by 8.5% and 10.4%, respectively, based on $IoU = [0.5 : 0.05 : 0.95]$, compared with Faster R-CNN. The AP_M and AR_M can reach

42.9% and 57.1%, respectively. As a whole, BFP Net outperforms the baseline of Faster R-CNN on the Pascal VOC. Meanwhile, we also validated and compared our method with some state-of-the-art models in Table 11. It can be observed that the AP_S , AP_M , AR_S , and AR_M of BFP Net are better than some state-of-the-art methods, which illustrates that our method can be better than some state-of-the-art models on the small-scale objects and medium-scale objects. However, the AR_L and AR_L are 54.3% and 82.1%, respectively, which is better than that of CARAFE [54], Dynamic R-CNN [55], PISA [56], and baseline, but is slightly lower than SABL [57], Mask R-CNN [58], and Groeie [59].

For this phenomenon, compared with the apple fruits with a class of apple and shape of spherical in the real orchard environment, we found that the number of categories in Pascal VOC is larger than that of MinneApple/GreenApple and the shapes of samples in each category are complex and diverse, which greatly increase the difficulty of target detection. Although BFP Net is not completely

TABLE 10: Comparison results on the MinneApple.

| Method | AP _S | AP _M | AP _L | AR _S | AR _M | AR _L | AP ₅₀ | AP ₇₅ | AP | AR |
|--------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|------------------|------|------|
| CARAFE [54] | 40.9 | 55.4 | -1 | 46.5 | 63.9 | -1 | 82.6 | 37.5 | 42.0 | 48.5 |
| SABL [55] | 40.5 | 57.4 | -1 | 46.4 | 66.3 | -1 | 81.5 | 37.7 | 42.0 | 48.7 |
| Dynamic R-CNN [56] | 38.6 | 53.4 | -1 | 44.7 | 62.6 | -1 | 81.7 | 33.9 | 39.9 | 46.8 |
| PISA [57] | 41.9 | 57.5 | -1 | 47.9 | 65.6 | -1 | 84.5 | 39.3 | 43.4 | 49.9 |
| Mask R-CNN [58] | 40.9 | 55.1 | -1 | 46.6 | 63.8 | -1 | 82.9 | 37.7 | 42.0 | 48.6 |
| Groie [59] | 41.0 | 55.3 | -1 | 46.7 | 63.9 | -1 | 82.6 | 38.0 | 42.3 | 48.7 |
| Faster R-CNN [27] | 41.2 | 56.3 | -1 | 46.9 | 65.0 | -1 | 83.1 | 38.8 | 42.6 | 49.0 |
| BFP Net | 42.2 | 58.3 | -1 | 48.2 | 66.8 | -1 | 84.6 | 40.6 | 43.5 | 50.4 |

TABLE 11: Our proposed method and comparison results on the Pascal VOC 2007 test set based on the IoU threshold of [0.5:0.05:0.95].

| Method | AP _S | AP _M | AP _L | AR _S | AR _M | AR _L | AP ₅₀ | AP ₇₅ | AP | AR |
|--------------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|------------------|------|------|
| CARAFE [54] | 26.7 | 40.2 | 54.0 | 38.1 | 49.2 | 63.5 | 81.0 | 54.4 | 50.2 | 59.9 |
| SABL [55] | 28.5 | 42.7 | 57.0 | 39.8 | 52.4 | 67.6 | 79.3 | 56.9 | 52.9 | 63.7 |
| Dynamic R-CNN [56] | 29.8 | 41.7 | 53.8 | 36.2 | 50.2 | 63.2 | 79.8 | 54.9 | 50.7 | 60.0 |
| PISA [57] | 27.5 | 40.4 | 52.6 | 38.1 | 48.8 | 62.1 | 80.0 | 53.3 | 49.3 | 58.9 |
| Mask R-CNN [58] | 32.0 | 41.3 | 57.0 | 43.3 | 53.2 | 67.0 | 83.6 | 57.8 | 53.0 | 63.7 |
| Groie [59] | 29.6 | 40.7 | 56.7 | 39.9 | 50.2 | 66.7 | 83.7 | 57.3 | 52.6 | 62.5 |
| Faster R-CNN [27] | 27.1 | 40.0 | 51.6 | 38.5 | 48.8 | 62.0 | 80.1 | 51.2 | 48.5 | 58.9 |
| BFP Net | 35.6 | 42.9 | 54.3 | 48.9 | 57.1 | 65.9 | 82.1 | 56.2 | 51.4 | 63.8 |

superior to the comparison models, as a whole, the BFP Net has a strong generalization ability for detecting the general object, especially for small target detection.

4.7. Visualization. To present clear detection results, BFP Net and some state-of-the-art models are visualized based on the GreenApple. The weight file obtained from the training set is applied to validate the model on the GreenApple test set. The visualization results are presented on the image through the orange rectangular box annotated with category and precision. Note that the category and precision of the target are located at the top left of the target rectangle. In addition, the ground-truth bounding box are presented with blue color.

Some visualization results about state-of-the-art models, baseline (Faster R-CNN) and our method are shown on the GreenApple, as shown in Figure 9. Specifically, visualization results of CARAFE, Dynamic R-CNN, PISA, SABL, Faster R-CNN (baseline), and BFP Net are presented using 4 cases. From this figure, it can be concluded that many methods can detect small apples with closeness of color between the fruit skin and the background in the real complex orchard environment. Compared with other state-of-the-art methods, our method can effectively improve the precision of small apple detection. It can be concluded that our method is suitable for for small apple detection to handle real-world complexities with varying light conditions, as well as occlusion or overlapped fruits.

4.8. Qualitative Results. To further validate our method, we qualitatively analyze our method and CARAFE on the benchmark dataset of GreenApple, MinneApple, and Pascal

VOC. Four prediction results from GreenApple and MinneApple are selected to analyze the effectiveness of the BFP Net and CARAFE, as shown in Figure 10.

Detection results of GreenApple and MinneApple are shown in Figure 10. It can be observed that precision of the BFP Net is higher than that of CARAFE. Although the apples is small and the color of the fruit skin is similar to the background, small apples can still be recognized in GreenApple. Compared with CARAFE based on ResNet50 with the FPN, BFP Net is better on the GreenApple and MinneApple, especially in small and dense fruit scenes. Overall, our method has strong competitiveness on large public datasets.

In addition, we also analyze the prediction results on the Pascal VOC. The results illustrated that our method has good generalization ability. To make the visualization results of qualitative analysis more consistent with this journal, we only show the visualization results based on the GreenApple and MinneApple in Figure 10.

5. Discussion and Conclusion

5.1. Discussion. The effectiveness of variants of general object detection and small object detection for small apple detection has been widely used in a series of studies [31–36]. In this article, a balanced feature pyramid network achieved the precision improvement of small apple detection. The balanced feature pyramid network could be directly embedded into existing object detectors to detect small apples, which had great potential for an unstructured orchard environment.

In previous studies, fruit detection methods based on conventional machine learning and deep learning had been

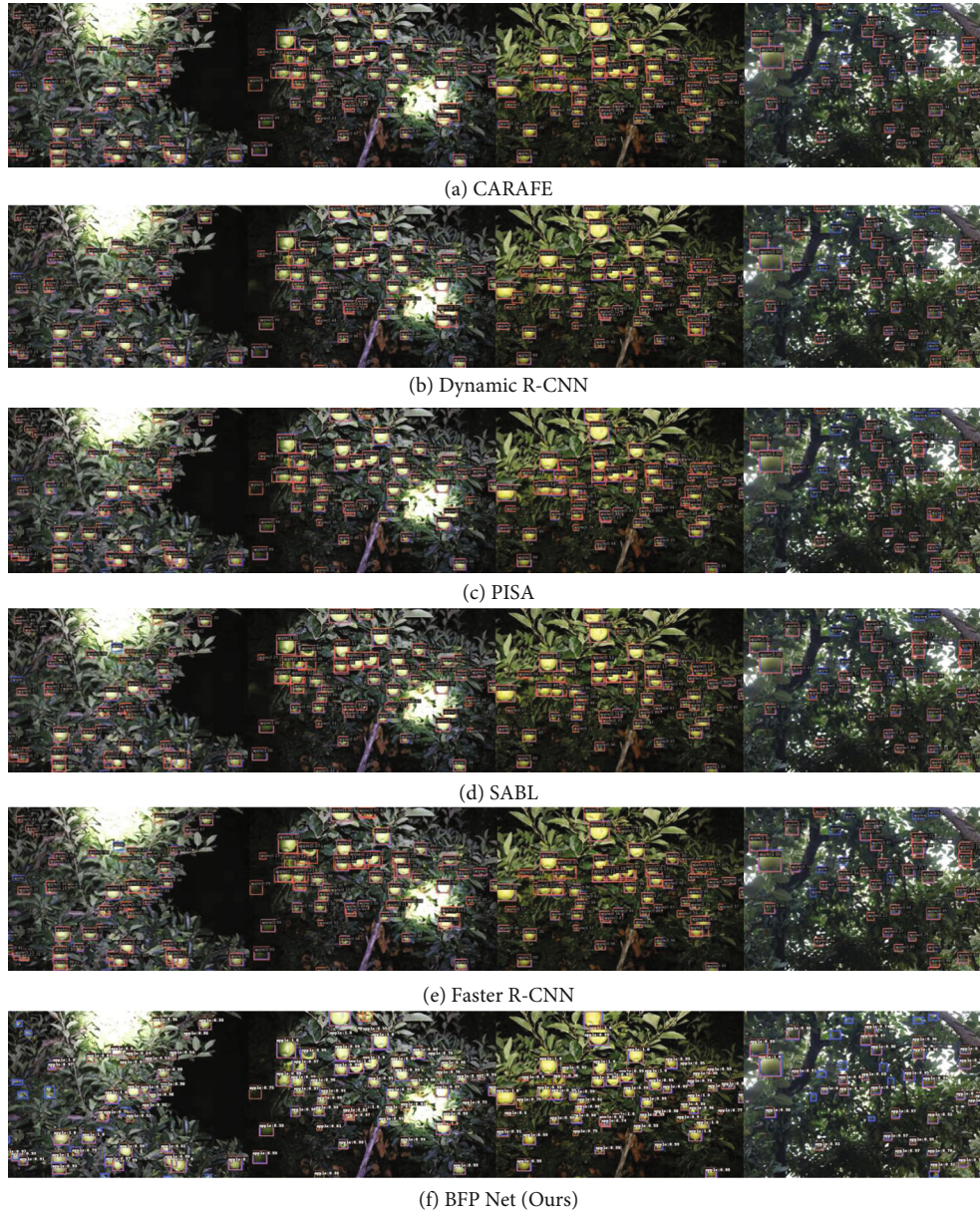


FIGURE 9: Visualization results of BFP Net and some state-of-the-art methods on the GreenApple. The four image cases in each line represent the four visualization results of corresponding method. The visualization methods contain CARAFE, Dynamic R-CNN, PISA, SABL, Faster R-CNN, and BFP Net (ours) from top to bottom. Note that the blue bounding box refers to the ground-truth bounding box and the orange bounding box represents the prediction bounding box on the GreenApple test set.

widely used. The results of the deep learning method always outperformed machine learning methods [60]. Manually extracted features from a series of small apple detection methods based on machine learning could not be comprehensively acquired [61], which leads to low efficiency and high labor cost. On the other hand, some generic object detection models based on deep learning are not completely suitable for fruit detection, due to small objects with fewer pixels under the complex natural orchard environment and the closeness of color between the fruit skin and the background for small fruits. The precision/accuracy of small objects is half of the large objects [12]. The unbalanced traits

between multiple-scale fruits limit the improvement of small target performance. The potential of multiscale information imbalance could not be fully revealed.

Balanced multiple-resolution feature fusion and spatial information presented promising results for small apple detection. Incorporating multiple fine-grained scale apple features and multiple-resolution pyramid features can acquire a better precision for small apple detection. Some studies have studied multiple scale objects [19, 62]. Overall, these methods showed that aggregating multiple scale object features and pyramid features can balance different scale features. The good performances illustrated that balanced

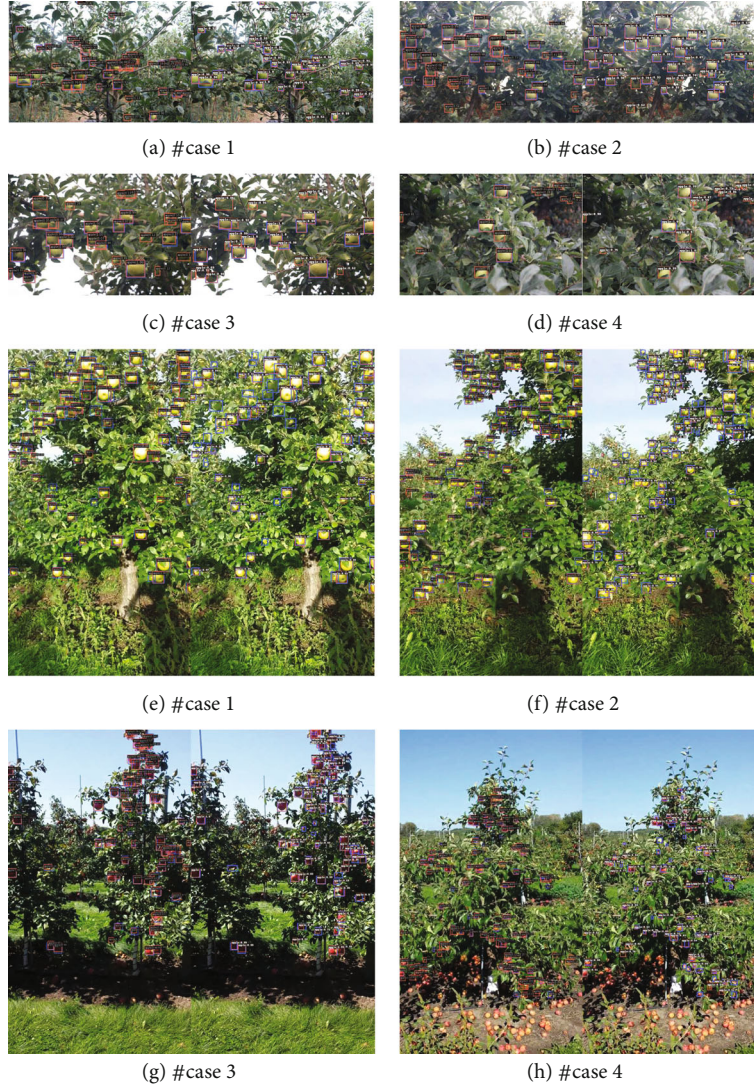


FIGURE 10: Qualitative results about the CARAFE and BFP Net (ours) on the GreenApple and MinneApple. Each dataset shows four samples for qualitative analysis. Note that the blue bounding boxes are the ground-truth bounding boxes; the orange bounding boxes are the prediction bounding boxes. The (a)–(d) are the results of GreenApple and (e)–(h) are results of MinneApple.

different scale features and different characteristic features on the extended layer have the great potential to aggregate more information for small apple detection.

Extended feature extracted abundant traits of multiple-scale apples. The fusion of extended features and pyramid-shaped multiple-scale features aggregated different scale features. The results proved that using the extended feature with high resolution had the potential to improve the precision of small apple identification and detection. In many studies, the authors also extended features on the feature fusion architecture to aggregate multiple-scale features and enhance the strong representation ability of apple features [26].

Moreover, the variations of performance gain between the weight-like feature fusion structure and decoupled-aggregated block with a KL distillation loss can be roughly evaluated. Compared with a limited feature from the backbone on the weight-like feature fusion architecture, an

extended feature through decoupled-aggregated block with a KL distillation loss can generate more feature gain to improve the performance of small apple detection. Compared with a lightweight content-aware upsample operator in CARAFE [54], the weight-like feature fusion structure by balancing reliable features from the backbone could acquire more fine-grained fruit traits. Similarly, some artifact features and fabricated fake textures constructed by small object detection methods based on the GAN lead to false positives [18]. It can be concluded that the reliable feature from original images could generate more gains compared with artifact features. For the decoupled-aggregated block with a loss function, an abundant feature from before pooling remains more features of small-scale apple, which can obtain more gain compared with a limited feature. In ablation studies, these views had proved that an unlimited feature was more helpful to optimize small fruit detection performance. The results presented the potential of the

balanced reliable feature with abundant information, and more efforts should be made to balance reliable features. In addition, proposal region or sample assign strategy etc. [55–57, 59] for detector were also designed to optimize the precision in many studies. Given the poor performance of small targets compared to large-scale targets, the optimization of target detectors has limitations for small-scale target performance improvement, and the results have been validated.

With the emergence of feature extraction and multiple scale feature fusion techniques, phenotyping features with strong representation ability can be obtained from different aspects. How to balance information features of different scales to recognize and detect multiple-scale apples during the whole process of fruit growth to maturity is of importance. Information balance of multiple aspects and scales provided an effective alternative to balance multiple scale features and spatial/semantic information on the extended layer. The performance of information balance for small apple detection proved that the information balance of different traits was promising. Therefore, different views could be analyzed to assess the effectiveness of small apple detection. The information balance of multiple perspectives will provide more fine-grained information and alleviate the problem of blurred boundaries on small apples, and how to effectively balance the phenotyping traits from multiple views remain a challenging issue. This article provided an efficient choice for information balance of different scale phenotyping features for small apple detection in real complex orchard environments, which could aggregate information from different perspectives to apply to the problem of the closeness of color between the fruit skin and the background of branches and leaves. Last but not least, we directly chose an existing detector to detect small apples but this detector cannot completely apply to the detection of different fruit scales in the whole process of fruit growth. How to optimize the fruit detector and making it suitable for small fruit/vegetable detection is also a potential alternative, which could provide some technical support for yield prediction, picking of mature fruits, and automatic monitoring of agriculture.

5.2. Conclusion. In conclusion, balanced apple information of different scales from two perspectives of multiple-scale fruits on the different layers of FPN and a characteristic of an extended feature has a certain effect on reducing the gap of apple information of different scales. The BFP Net presents strong potential to improve the performance of small apple detection. The experimental results also illustrated that a weight-like feature fusion structure can refine multiple-scale feature information. A decoupled-aggregated module can balance spatial-aware information and content-aware information to complement the spatial information of small apples based on an extended layer. The Kullback-Leibler distillation loss make the feature distribution consistent between a high-level feature and extended feature map. The results demonstrate that each component of the balanced feature pyramid network facilitates the improvement of small apple detection precision,

in particular, the precision of small apple localization. In addition, this method has the ability to improve the precision without adding a large number of parameters. This method can not only be applied to other small apple detection based on feature pyramid feature fusion architecture but also provide a reference for other fruit/vegetable detection.

Data Availability

The data used in this study are freely available. Anyone who wants to use the data can contact the corresponding author Weikuan Jia. The author is with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (jwk_1982@163.com).

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this article.

Authors' Contributions

M. Sun, Y. Zheng, and W. Jia conceived the idea for the paper; L. Xu, and X. Chen with contributions for data curation; M. Sun and Z. Ji wrote the code and designed and conducted the experiments; M. Sun, L. Xu, and X. Chen with contributions for visualization and validation; M. Sun and W. Jia with contributions for writing—original draft preparation; Z. Ji, Y. Zheng, and W. Jia with contributions for editing and reviewing of the manuscript.

Acknowledgments

This work is supported by the Natural Science Foundation of Shandong Province in China (No.: ZR2020MF076); the National Nature Science Foundation of China (No.: 62072289); the Focus on Research and Development Plan in Shandong Province (No.: 2019GNC106115); the Taishan Scholar Program of Shandong Province of China; and the New Twentieth Items of Universities in Jinan (2021GXRC049).

References

- [1] W. Zhang, K. Chen, J. Wang, Y. Shi, and W. Guo, "Easy domain adaptation method for filling the species gap in deep learning-based fruit detection," *Horticulture Research*, vol. 8, no. 1, p. 119, 2021.
- [2] C. C. Ukwuoma, Q. Zhiguang, M. B. Bin Heyat, L. Ali, Z. Almaspoor, and H. N. Monday, "Recent advancements in fruit detection and classification using deep learning techniques," *Mathematical Problems in Engineering*, vol. 2022, Article ID 9210947, 29 pages, 2022.
- [3] N. Mamdouh, M. Wael, and A. Khattab, "Artificial intelligence-based detection and counting of olive fruit flies: a comprehensive survey," *Deep Learning for Sustainable Agriculture, Chapter 14*, pp. 357–380, 2022.
- [4] P. Maheswari, P. Raja, O. E. Apolo-Apolo, and M. P'erez-Ruiz, "Intelligent fruit yield estimation for orchards using deep learning based semantic segmentation techniques—a review," *Frontiers in Plant Science*, vol. 12, p. 1247, 2021.

- [5] P. Singh and A. Kaur, "A systematic review of artificial intelligence in agriculture," *Deep Learning for Sustainable Agriculture, Chapter 2*, pp. 57–80, 2022.
- [6] A. Koirala, K. B. Walsh, Z. Wang, and C. McCarthy, "Deep learning – method overview and review of use for fruit detection and yield estimation," *Computers and Electronics in Agriculture*, vol. 162, pp. 219–234, 2019.
- [7] X. Ni, C. Li, H. Jiang, and F. Takeda, "Deep learning image segmentation and extraction of blueberry fruit traits associated with harvestability and yield," *Horticulture Research*, vol. 7, no. 1, p. 692, 2020.
- [8] B. S. Faiçal, H. Freitas, P. H. Gomes et al., "An adaptive approach for uav-based pesticide spraying in dynamic environments," *Computers and Electronics in Agriculture*, vol. 138, pp. 210–223, 2017.
- [9] Y. Tang, M. Chen, C. Wang et al., "Recognition and localization methods for vision-based fruit picking robots: a review," *Frontiers in Plant Science*, vol. 11, p. 510, 2020.
- [10] W. Jia, Y. Zhang, J. Lian, Y. Zheng, D. Zhao, and C. Li, "Apple harvesting robot under information technology: a review," *International Journal of Advanced Robotic Systems*, vol. 17, no. 3, article 172988142092531, 2020.
- [11] S. K. Behera, P. K. Sethy, S. K. Sahoo, S. Panigrahi, and S. C. Rajpoot, "On-tree fruit monitoring system using iot and image analysis," *Concurrent Engineering*, vol. 29, no. 1, pp. 6–15, 2021.
- [12] K. Tong, Y. Wu, and F. Zhou, "Recent advances in small object detection based on deep learning: a review," *Image and Vision Computing*, vol. 97, article 103910, 2020.
- [13] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.
- [14] K. G. Liakos, P. Busato, D. Moshou, S. Pearson, and D. Bochtis, "Machine learning in agriculture: a review," *Sensors*, vol. 18, no. 8, p. 2674, 2018.
- [15] S. K. Behera, A. K. Rath, A. Mahapatra, and P. K. Sethy, "Identification, classification & grading of fruits using machine learning & computer intelligence: a review," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–11, 2020.
- [16] S. Chaivivatrakul and M. N. Dailey, "Texture-based fruit detection," *Precision Agriculture*, vol. 15, no. 6, pp. 662–683, 2014.
- [17] J. Gené-Mola, E. Gregorio, F. A. Cheein et al., "Fruit detection, yield prediction and canopy geometric characterization using lidar with forced air flow," *Computers and Electronics in Agriculture*, vol. 168, article 105121, 2020.
- [18] J. Noh, W. Bae, W. Lee, J. Seo, and G. Kim, "Better to follow, follow to be better: towards precise supervision of feature super-resolution for small object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9725–9734, Korea, 2019.
- [19] L. Cui, P. Lv, X. Jiang et al., "Context-aware block net for small object detection," *IEEE Transactions on Cybernetics*, vol. 52, no. 4, 2020.
- [20] H. Zheng, J. Chen, L. Chen, Y. Li, and Z. Yan, "Feature enhancement for multi-scale object detection," *Neural Processing Letters*, vol. 51, no. 2, pp. 1907–1919, 2020.
- [21] Y. Bai, Y. Zhang, M. Ding, and B. Ghanem, "Sod-mtgan: small object detection via multi task generative adversarial network," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 206–221, Germany, 2018.
- [22] Z. Liang, J. Shao, D. Zhang, and L. Gao, "Small object detection using deep feature pyramid networks," in *Pacific Rim Conference on Multimedia*, pp. 554–564, Springer, Cham, 2018.
- [23] T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2117–2125, Honolulu, 2017.
- [24] X. Wang, K. Yu, S. Wu et al., "Esrgan: enhanced super-resolution generative adversarial networks," in *Proceedings of the European conference on computer vision (ECCV) workshops*, Germany, 2018.
- [25] J. Rabbi, N. Ray, M. Schubert, S. Chowdhury, and D. Chao, "Small-Object detection in remote sensing images with end-to-end edge-enhanced Gan and object detector network," *Remote Sensing*, vol. 12, no. 9, p. 1432, 2020.
- [26] C. Deng, M. Wang, L. Liu, Y. Liu, and Y. Jiang, "Extended feature pyramid network for small object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1968–1979, 2022.
- [27] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.
- [28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 743–770, Las Vegas, 2016.
- [29] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1222–1230, Honolulu, 2017.
- [30] M. Kisantal, J. M. Wojna, J. Naruniec, and K. Cho, "Augmentation for small object detection," 2019, <http://arxiv.org/abs/1902.07296>.
- [31] D. Hussain, I. Hussain, M. Ismail, A. Alabrah, S. S. Ullah, and H. M. Alaghbari, "A simple and efficient deep learning-based framework for automatic fruit recognition," *Computational Intelligence and Neuroscience*, vol. 2022, Article ID 6538117, 8 pages, 2022.
- [32] B. Yan, P. Fan, X. Lei, Z. Liu, and F. Yang, "A real-time apple targets detection method for picking robot based on improved yolov5," *Remote Sensing*, vol. 13, no. 9, p. 1619, 2021.
- [33] W. Jia, Y. Tian, R. Luo, Z. Zhang, J. Lian, and Y. Zheng, "Detection and segmentation of overlapped fruits based on optimized mask r-cnn application in apple harvesting robot," *Computers and Electronics in Agriculture*, vol. 172, article 105380, 2020.
- [34] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [35] S. Lu, W. Chen, X. Zhang, and M. Karkee, "Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense- foliage tree architectures for early crop load estimation," *Computers and Electronics in Agriculture*, vol. 193, article 106696, 2022.
- [36] X. Li, J. Pan, F. Xie et al., "Fast and accurate green pepper detection in complex backgrounds via an improved yolov4-tiny model," *Computers and Electronics in Agriculture*, vol. 191, article 106503, 2021.
- [37] K. Han, Y. Wang, H. Chen et al., "A survey on vision transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

- [38] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8759–8768, Salt Lake City, 2018.
- [39] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10781–10790, Seattle, 2020.
- [40] S. Qiao, L.-C. Chen, and A. Yuille, "Detectors: detecting objects with recursive feature pyramid and switchable atrous convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10213–10224, Kuala Lumpur, 2021.
- [41] N. Häni, P. Roy, and V. Isler, "MinneApple: a benchmark dataset for apple detection and segmentation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 852–858, 2020.
- [42] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [43] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International Journal of Computer Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.
- [44] A. Borji, "Empirical upper bound, error diagnosis and invariance analysis of modern object detectors," 2020, <http://arxiv.org/abs/2004.02877>.
- [45] J. Zhou, V. Jampani, Z. Pi, Q. Liu, and M.-H. Yang, "Decoupled dynamic filter networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6647–6656, Kuala Lumpur, 2021.
- [46] J. Guo, K. Han, Y. Wang et al., "Distilling object detectors via decoupled features," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2154–2164, Kuala Lumpur, 2021.
- [47] P. Mazumder, P. Singh, and V. Namboodiri, "Cpwc: contextual point wise convolution for object recognition," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4152–4156, IEEE, Barcelona, Spain, 2020.
- [48] M. Sun, Z. Song, X. Jiang, J. Pan, and Y. Pang, "Learning pooling for convolutional neural network," *Neurocomputing*, vol. 224, pp. 96–104, 2017.
- [49] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, Salt Lake City, 2018.
- [50] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240, New York, NY, USA, 2006.
- [51] H. Rezatofghi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 658–666, California, 2019.
- [52] A. Paszke, S. Gross, F. Massa et al., "Pytorch: an imperative style, high-performance deep learning library," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [53] K. Chen, J. Wang, J. Pang et al., "Mmdetection: open mmlab detection toolbox and benchmark," 2019, <http://arxiv.org/abs/1906.07155>.
- [54] J. Wang, K. Chen, R. Xu, Z. Liu, C. C. Loy, and D. Lin, "Carafe: content-aware reassembly of features," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3007–3016, Korea, 2019.
- [55] J. Wang, W. Zhang, Y. Cao et al., "Side-aware boundary localization for more precise object detection," in *European Conference on Computer Vision*, pp. 403–419, Springer, Cham, 2020.
- [56] H. Zhang, H. Chang, B. Ma, N. Wang, and X. Chen, "Dynamic r-cnn: towards high quality object detection via dynamic training," in *European conference on computer vision*, pp. 260–275, Springer, Cham, 2020.
- [57] Y. Cao, K. Chen, C. C. Loy, and D. Lin, "Prime sample attention in object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11583–11591, Seattle, 2020.
- [58] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, Venice, 2017.
- [59] L. Rossi, A. Karimi, and A. Prati, "A novel region of interest extraction layer for instance segmentation," in *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 2203–2209, IEEE, Milan, Italy, 2021.
- [60] H. Kuang, C. Liu, L. L. H. Chan, and H. Yan, "Multi-class fruit detection based on image region selection and improved object proposals," *Neurocomputing*, vol. 283, pp. 241–255, 2018.
- [61] 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.
- [62] L. Cui, R. Ma, P. Lv et al., "Mdssd: multi-scale deconvolutional single shot detector for small objects," 2018, <http://arxiv.org/abs/1805.07009>.