Multi-Class on-Tree Peach Detection Using Improved YOLOv5s and Multi-Modal Images postprint

Authors: LUO Qing, RAO Yuan, JIN Xiu1, JIANG Zhaohui, WANG Tan, Fengyi Wang, ZHANG Wu, RAO Yuan

Date: 2023-02-17T00:00:00+00:00

Abstract

Accurate peach detection is a prerequisite for automated agronomic management, e.g., peach mechanical harvesting. However, due to uneven illumination and ubiquitous occlusion, it is challenging to detect the peaches, especially when the peaches are bagged in orchards. To this end, an accurate multi-class peach detection method was proposed by means of improving YOLOv5s and using multi-modal visual data for mechanical harvesting in this paper. RGB-D dataset with multi-class annotations of naked and bagging peach was proposed, including 4127 multimodal images of corresponding pixel-aligned color, depth, and infrared images acquired with consumer-level RGBD camera. Subsequently, an improved lightweight YOLOv5s (small depth) model was put forward by introducing a direction-aware and position-sensitive attention mechanism, which could capture long-range dependencies along one spatial direction and preserve precise positional information along the other spatial direction, helping the networks accurately detect peach targets. Meanwhile, the depthwise separable convolution was employed to reduce the model computation by decomposing the convolution operation into convolution in the depth direction and convolution in the width and height directions, which helped to speed up the training and inference of the network while maintaining accuracy. The comparison experimental results demonstrated that the improved YOLOv5s using multimodal visual data recorded the detection mAP of 98.6% and 88.9% on the naked and bagging peach with 5.05 M model parameters in complex illumination and severe occlusion environment, increasing by 5.3% and 16.5% than only using RGB images, as well as by 2.8% and 6.2% when compared to YOLOv5s. As compared with other networks in detecting bagging peaches, the improved YOLOv5s performed best in terms of mAP, which was 16.3%, 8.1% and 4.5% higher than YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0, respectively. In addition,

the proposed improved YOLOv5s model offered better results in different degrees than other methods in detecting Fuji apple and Hayward kiwifruit, verified the effectiveness on different fruit detection tasks. Further investigation revealed the contribution of each imaging modality, as well as the proposed improvement in YOLOv5s, to favorable detection results of both naked and bagging peaches in natural orchards. Additionally, on the popular mobile hardware plat-form, it was found out that the improved YOLOv5s model could implement 19 times detection per second with the considered five-channel multi-modal images, offering real-time peach detection. These promising results demonstrated the potential of the improved YOLOv5s and multi-modal visual data with multi-class annotations to achieve visual intelligence of automated fruit harvesting systems.

Full Text

Preamble

Multi-Class On-Tree Peach Detection Using Improved YOLOv5s and Multi-Modal Images

Qing Luo[1,2,3], Yuan Rao[1,2,3]*, Xiu Jin[1,2,3], Zhaohui Jiang[1,2,3], Tan Wang[1,2,3], Fengyi Wang[1,2,3], Wu Zhang[1,2,3]

[1] College of Information and Computer Science, Anhui Agricultural University, Hefei 230036, China
[2] Key Laboratory of Agricultural Sensors, Ministry of Agriculture and Rural Affairs, Hefei 230036, China
[3] Anhui Provincial Key Laboratory of Smart Agricultural Technology and Equipment, Hefei 230036, China

Abstract: Accurate peach detection is a prerequisite for automated agronomic management, such as mechanical harvesting. However, due to uneven illumination and ubiquitous occlusion, detecting peaches—especially bagged peaches in orchards—remains challenging. To address this, we propose an accurate multi-class peach detection method by improving YOLOv5s and utilizing multi-modal visual data for mechanical harvesting. Specifically, we present an RGB-D dataset with multi-class annotations for both naked and bagged peaches, comprising 4,127 multi-modal image sets of pixel-aligned color, depth, and infrared images acquired with a consumer-level RGB-D camera. Subsequently, we propose an improved lightweight YOLOv5s (small depth) model by introducing a direction-aware and position-sensitive attention mechanism that captures long-range dependencies along one spatial direction while preserving precise positional information along the other, helping the network accurately detect peach targets. Meanwhile, depthwise separable convolution is employed to reduce model computation by decomposing convolution operations into depthwise and width/height-wise convolutions, accelerating network training and inference while maintaining accuracy. Experimental results demonstrate that the improved YOLOv5s using multi-modal visual data achieves detection mAPs

of 98.6% and 88.9% for naked and bagged peaches, respectively, with 5.05 M model parameters in complex illumination and severe occlusion environments —representing improvements of 5.3% and 16.5% over using RGB images alone, and 2.8% and 6.2% compared to YOLOv5s. For bagged peach detection, the improved YOLOv5s outperforms other networks in mAP, achieving gains of 16.3%, 8.1%, and 4.5% over YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0, respectively. The proposed model also delivers superior results for Fuji apple and Hayward kiwifruit detection, verifying its effectiveness across different fruit detection tasks. Further investigation reveals the contribution of each imaging modality and the proposed YOLOv5s improvements to favorable detection results for both naked and bagged peaches in natural orchards. Additionally, on popular mobile hardware platforms, the improved YOLOv5s model achieves 19 detections per second with five-channel multi-modal images, enabling real-time peach detection. These promising results demonstrate the potential of improved YOLOv5s and multi-modal visual data with multi-class annotations for achieving visual intelligence in automated fruit harvesting systems.

Keywords: multi-class detection; YOLOv5s; multi-modal visual data; mechanical harvesting; deep learning

CLC number: S662.1; S126
Document code: A
Article ID: SA202210004

Citation: LUO Qing, RAO Yuan, JIN Xiu, JIANG Zhaohui, WANG Tan, WANG Fengyi, ZHANG Wu. Multi-class on-tree peach detection using improved YOLOv5s and multi-modal images[J]. Smart Agriculture, 2022, 4(4): 84-104. (in English with Chinese abstract)

1 Introduction

Peach is the third most productive temperate tree species after apple and pear and is an excellent source of vitamin C[1]. Peach harvesting is a time-consuming and labor-intensive task. Efficient harvesting methods are required to meet the growing global demand for fruit and improve orchard productivity[2]. Research and improvements in automated technologies like mechanical harvesting provide farmers with practical approaches to increase production. Traditional methods typically use segmentation algorithms or shape features such as color, shape, or texture to detect fruit[3-5]. However, peach detection in orchards is challenging due to occlusion from branches and leaves, as well as ever-changing illumination, making accurate detection difficult using traditional methods[6]. The first task in automated peach harvesting is accurate detection. While in-field fruit detection has been widely applied to various fruits, most images acquired using traditional methods were captured under controlled illumination, making them vulnerable to complex orchard environments[7]. Additionally, other environmental factors such as changing fruit appearance and morphology can critically

affect detection accuracy.

Compared to traditional methods, deep learning exhibits strong adaptability to variations within working scenes and has become one of the most promising techniques for learning image features. Consequently, deep learning algorithms have been widely used for fruit detection in agricultural robots operating in unstructured environments[8-10]. However, in real fruit orchards, one of the greatest challenges in fruit detection stems from complex orchard environments, such as changing backgrounds[11]. Meanwhile, varying scales of fruit targets also cause substantial difficulties, especially when fruits are bagged in orchards. Therefore, with the vigorous development of deep learning derived from the imitation of human vision, attention mechanisms have been applied to enhance model perception ability under complex conditions. In recent years, many attention mechanism strategies have been widely adopted for various fruit detection tasks. Li et al.[12] introduced a deep learning target detection algorithm based on improved YOLOv4-tiny that combined an attention mechanism with multi-scale prediction to improve recognition of occluded and small-target green peppers. Jiang et al.[13] efficiently detected young apples by adding a non-local attention module and convolutional block attention model to YOLOv4. Huang et al.[14] extended the target detection algorithm by adding a convolutional block attention module (CBAM) to improve citrus detection performance. Overall, these studies demonstrated that attention mechanisms could enhance detection performance and adapt to natural environments with complex backgrounds.

Nevertheless, many challenges remain for effective fruit detection in practical scenes. Under actual agricultural production conditions, using RGB images as the sole information source for fruit detection is undesirable when interference factors such as occlusion, fruit overlap, and changing illumination are present[15]. Fortunately, with the development of consumer-level RGB-D cameras such as Microsoft Kinect and Intel RealSense, additional information like depth and infrared data provides cues to address these problems. Sa et al.[16] input RGB and infrared images to Faster R-CNN for sweet pepper identification. Fu et al.[15] developed an outdoor machine vision system with an RGB-D camera to improve apple identification by using depth features to filter out background objects. Arad et al.[17] presented a sweet pepper harvesting robot equipped with an RGB-D camera that acquired color and depth information for detecting and locating each fruit. These studies claimed that introducing modalities beyond RGB could improve fruit detection performance. However, they mainly focused on detecting naked fruits without severe occlusion and overlap due to standard planting or fruiting-wall architectures. In fact, bagging late-ripening and high-quality fruits is a popular method to prevent diseases and extend storage duration[4]. This agronomic practice increases the difficulty of in-field fruit detection by introducing more severe occlusion and irregular target shapes. Therefore, in addition to detecting naked fruits, investigating effective detection of bagged peaches is also meaningful.

For these reasons, we propose an efficient detection model that utilizes three-

dimensional spatial geometry and backscatter signal intensity information from multi-modal images to detect in-field naked and bagged peaches for guiding mechanical harvesting. Specifically, we present an RGB-D dataset of naked and bagged peaches comprising 4,127 corresponding color, depth, and infrared images obtained by an RGB-D camera. According to fruit picking strategies and field occlusion status, peaches are classified into four classes: un-occluded, occluded by leaves, occluded by fruits, and occluded by branches. Remarkably, we propose an optimized detector for peach detection by introducing coordinate attention mechanism and depthwise separable convolution into YOLOv5s. To evaluate the performance of the improved YOLOv5s using multi-modal images and explore the contribution of each imaging modality to environmental adaptation, we conduct extensive experiments from various aspects. Further investigation reveals the contribution of each imaging modality and the improved YOLOv5s in alleviating the negative influence of complex illumination and severe occlusion. Additionally, the computational time of the proposed detection model meets real-time detection requirements through successful optimization and deployment on NVIDIA Jetson Nano. This study provides the possibility and foundation for performing visual intelligence in mechanical harvesting by utilizing improved YOLOv5s and multi-modal visual data with multi-class annotations.

---

2 Materials and Methods

2.1 Data Acquisition

Images were acquired using a Microsoft Azure Kinect RGB-D camera (key parameters listed in), which incorporates an RGB (Red-Green-Blue) sensor and a depth sensor based on the ToF (Time of Flight) principle. Data were collected in a peach orchard located in Dawei Town, Hefei City, Anhui Province, China. Two agronomic measures were present in the orchards: naked and bagged peaches. According to planting methods and ripening periods, high-quality and late-ripening peaches were typically bagged with red paper to prevent extreme climate and disease damage, while early-ripening peaches tended to be naked to facilitate harvesting.

[Figure 1: see original paper] shows the data acquisition scenario, with naked peach trees on the left side and bagged peach trees on the right. The RGB-D camera provides three data types: RGB image, IR backscattered intensity (IR), and depth image (Depth) for peach localization. Image data were collected in peach orchards during sunny and cloudy weather from 7 a.m. to 9 p.m. during August and September. During acquisition, the camera was aimed perpendicular to the sunlight direction to capture multi-modal images under normal illumination conditions. The camera' s viewing direction was set parallel to the sunlight direction to capture multi-modal images under strong illumination. Multi-modal images were also gathered under artificial illumination dur-

ing nighttime. Considering that occlusion affects detection performance, some images were collected with different degrees of occluded targets from multiple viewing angles.

Based on the proportion of target area occluded by branches and leaves, occlusion levels were classified as: Slight occlusion (0-30% occluded), General occlusion (30%-60% occluded), and Severe occlusion (60%-100% occluded). To simulate varying camera distances during mechanical harvesting, the camera was placed 0.1 m to 1.5 m from the tree trunk. Distance within 0.3 m was considered close distance to simulate the end-effector approaching the target. Distance from 0.3 m to 1 m was considered average distance to simulate the camera position detecting most target fruits. Distance greater than 1 m was considered far distance to simulate the camera position relatively far from target fruits.

Specific software written in C++ was developed to automatically collect and save data. The software drove the RGB-D camera to record data in situ at 5 Hz. Each recording contained pixel-aligned RGB, infrared, and depth images. In total, 4,127 pairs of multi-modal images were acquired, with examples shown in [Figure 2: see original paper].

2.2 Multi-Class Peach RGB-D Dataset

Manual annotation was applied after image collection. Considering ubiquitous occlusions among leaves, branches, and fruits in natural orchards, peaches were manually labeled with bounding boxes tangent to peach outlines according to robotic picking strategy and in-field occlusion status. In cases of occlusion, peaches with occlusion area greater than 85% and targets at image edges with less than 15% area were not labeled[19]. After labeling, TXT format annotation files were generated containing peach class names and bounding box pixel coordinates.

Peaches were classified into four classes based on picking strategy and occlusion status to achieve selective picking and prevent damage to end-effectors or robots[18]: 1. **NO** (Not occluded): Peaches without occlusion 2. **OL** (Occluded by Leaves): Peaches occluded only by leaves, not by other peaches or branches 3. **OF** (Occluded by Fruits): Peaches occluded by other peaches 4. **OB** (Occluded by Branches): Peaches occluded by branches

When OB and OF appeared simultaneously for the same peach, OB was prioritized. When OF and OL appeared simultaneously, OF was considered. For the four annotated classes, peaches inside white, green, cyan, and brown boxes represented NO, OL, OF, and OB, respectively.

As shown in [Figure 3: see original paper], all peaches were manually labeled with bounding boxes. The dataset contained 4,127 peach images divided into two types: 2,077 naked peach images and 2,050 bagged peach images. This dataset has been made publicly available at https://github.com/tsing-luo/Multi-

class-peach-RGB-D-dataset.

2.3 Improved YOLOv5s Network

The YOLO series[20-24] has become one of the most popular deep learning frameworks among one-stage detectors and is widely used in target detection tasks. In practical agricultural management, real-time detection under limited computational and storage resources is required, while there are limitations on the size and inference time of fruit detection algorithms. The newly proposed YOLOv5s performs well in pursuing a trade-off between accuracy and speed, offering inference speeds up to 140 FPS (frames per second). Additionally, the YOLOv5 model weight file is only 7.2 MB, nearly 90% smaller than YOLOv4.

As depicted in [Figure 4: see original paper], YOLOv5s was employed as the basis for the fruit detection model. The model mainly includes three parts: Backbone, Neck, and Prediction head. The original YOLOv5s uses CSPDarknet53 as the backbone network. However, due to complex backgrounds in orchards, target features extracted from images are easily disturbed, particularly when weeds and soil have colors similar to peaches, leading to incorrect detection results. Meanwhile, shallow feature maps extracted from the backbone have small receptive fields suitable for detecting small targets like fruits[25]. Nevertheless, using low-dimensional feature maps to increase feature information for small targets might introduce significant background noise, particularly when using multi-modal images, potentially decreasing detection accuracy.

To solve these problems, as shown in [Figure 4: see original paper] and [Figure 5: see original paper], we modified the CSP module design in the backbone and improved the neck by introducing an efficient attention mechanism called coordinate attention (CA), which inherits the benefits of channel attention methods while simultaneously capturing long-range dependencies with precise positional information, suppressing unimportant features and promoting useful ones[26]. We also introduced depthwise separable convolution (DSC) to substitute part of regular convolutions in the backbone and neck network to reduce model parameters and speed up detection inference time without penalizing accuracy[31-33].

Coordinate Attention Mechanism Previous studies have proven that adding coordinate attention to the feature extraction part of the model can enhance attention region representation, while adding attention mechanisms to the neck can improve position sensitivity in the detection head, preserving relative positions between features and achieving more accurate detection results[27-29]. Specifically, given shallow feature map input, a pair of direction-aware feature maps are yielded using two spatial extents of pooling kernels (H$×1)*and*(1×$W) to encode each channel along horizontal and vertical coordinates, respectively. These transformations allow the attention block to capture long-range dependencies along one spatial direction and preserve precise positional information along the other, helping the network accurately locate peaches.

The feature maps produced by the coordinate information embedding block are concatenated and sent to a shared 1$×$1 convolutional transformation, converting them to dimensions C/r × 1 × (H+W), where r is the reduction ratio controlling block size as in the SE block[30]. After Batch Normalization and non-linear activation, the feature maps are split into separate tensors along the spatial dimension. Another 1$×$1 convolution transforms horizontal dimension tensors f_h and vertical dimension tensors f_w to tensors with the same channel number as the input C × H × W. The outputs can be formulated as:

$$g_h = \sigma(C_h(f_h)) \quad (1)$$
$$g_w = \sigma(C_w(f_w)) \quad (2)$$

where C denotes convolutional transformation and σ is the sigmoid function. Finally, the output Y is written as:

$$Y_{i,j} = X_{i,j} \times g_h(i) \times g_w(j) \quad (3)$$

Depthwise Separable Convolution Depthwise separable convolution (DSC) is a combination of depthwise convolution and pointwise convolution. Depthwise convolution contains c_1 convolution kernels of size h × w × 1, performing filtering on each channel. Pointwise convolution contains c_2 kernels of size 1 × 1 × c_1, converting channels by acting on the depthwise convolution output. The parameters for DSC and traditional convolution are:

$$P_{DSC} = h \times w \times c_1 + 1 \times 1 \times c_1 \times c_2 \quad (4)$$
$$P_{conv} = h \times w \times c_1 \times c_2 \quad (5)$$

Comparing P_{DSC} and P_{conv} reveals that DSC effectively decomposes traditional convolution by separating spatial filtering from feature generation mechanisms.

For backbone parameters, we chose multi-modal images with 640$×$640$resolutionasmodelinput.Shallowfeatur$
$layerDSC, aCSP_{CA}1, andaCSP_{CA}3module, convertingfeaturedimensionsto$128×128×$64.
Additional features were extracted through a two-layer CSP_{CA}3 module, two-layer DSC, and an SPP module[23]. Three adequate feature levels were obtained: the first two focused on small-scale and medium-scale features, and the last on large-scale features, which were then transferred to the neck.

In the prediction head, the k-means clustering algorithm was used to find anchor boxes, and Complete IoU (CIoU)[34] was used for bounding box regression loss, considering three geometric properties: overlap area, central point distance, and aspect ratio, leading to faster convergence and better performance. The formulas are:

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \quad (6)$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \quad (7)$$

$$\alpha = \frac{v}{(1 - IoU) + v} \quad (8)$$

where c represents the diagonal distance of the smallest closure area containing both predicted and ground truth bounding boxes; $\rho^2(b, b^{gt})$ represents Euclidean distance between predicted and ground truth center points; IoU represents overlap between predicted and ground truth bounding boxes; and L_{CIoU} is the loss function.

2.4 Model Deployment

The PyTorch framework was used to train the network, generating a model in PTH format. After training, the model was deployed on NVIDIA Jetson Nano to evaluate real-time detection potential. Jetson Nano supports TensorRT for model acceleration, which improves neural network processing speed by optimizing algorithm architecture. First, the PTH format model was converted to ONNX format as an intermediate framework bridging PyTorch and TensorRT models. Then, the ONNX model was converted to TensorRT format and tested on Jetson Nano. After deployment, time consumption was verified when using multi-modal images to detect naked and bagged peaches.

---

3 Experimental Results and Analysis

To thoroughly evaluate the performance of improved YOLOv5s using multi-modal images and explore each imaging modality' s contribution when detecting multi-class naked and bagged peaches in natural orchards, different multi-modal image combinations were input into the improved YOLOv5s. Model performance was evaluated in terms of precision (P), recall (R), mean average precision (mAP), and detection speed. First, to evaluate multi-modal image performance in model generalization, the improved YOLOv5s was trained and validated based on different imaging modality combinations, with quantitative analysis performed on test results for naked and bagged peach detection. Second, to explore each imaging modality' s contribution in different orchard environments, detection results were compared and analyzed across several typical scenarios (different illumination conditions, fruit occlusion levels, and camera distances). Finally, ablation studies were conducted to verify the effectiveness of the coordinate attention mechanism and depthwise separable convolution.

3.1 Training Platform and Parameters

The deep learning framework used was PyTorch 1.11.0. The training and testing platform included a server with an Intel Xeon Gold 5118 @ 2.30 GHz 12-core CPU, one NVIDIA RTX2080Ti (1620 MHz) GPU with 4,352 CUDA cores and 11 GB memory, running CentOS 7.9. Software tools included CUDA 11.2, CUDNN 7.6.5, and Python 3.7. shows network initialization parameters. All input images were adjusted to 640$×$640 pixels to meet network requirements. Considering server memory constraints, batch size was set to eight. One hundred fifty epochs were used to better analyze training. Parameters like momentum, learning rate, and weight decay followed the original YOLOv5s model.

3.2 Evaluation Indicators

Model performance was evaluated using average precision (AP), mAP, and detection speed. AP was estimated using precision (P) and recall (R), reflecting network sensitivity to target detection and model performance[35]. P and R are defined as:

$$P = \frac{TP}{TP + FP} \quad (9)$$

$$R = \frac{TP}{TP + FN} \quad (10)$$

where TP (True Positive) is correctly detected targets; FP (False Positive) is incorrectly classified detections; FN (False Negative) is missed targets. AP is defined in Equation (11) as the area under the P-R curve. mAP is defined in Equation (12) as the average AP value:

$$AP = \int_0^1 P(R)dR \quad (11)$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \quad (12)$$

3.3 Performance of Improved YOLOv5s Using Different Multi-Modal Images

To evaluate improved YOLOv5s performance using multi-modal images for detecting multi-class naked and bagged peaches, the model was trained, validated, and tested using different imaging modality combinations. The dataset included the multi-class naked peach dataset (2,077 multi-modal images randomly divided into 70% training, 10% validation, and 20% testing) and the bagged peach dataset (1,454 training, 208 validation, and 415 testing pairs). A multi-modal image means a set of RGB, Depth, and IR images that are channel-fused to obtain a four- or five-channel image. For example, RGB+Depth and RGB+IR

stack RGB with corresponding Depth or IR images to create four-channel images. RGB+Depth+IR denotes fusing all three modalities into a five-channel image. The image count was identical across modalities; only the number of input channels differed.

3.3.1 Training Assessment From validation curves in Figure 7: see original paper, the model did not overfit during training. For training curves, the loss function reached lower values when using RGB+Depth+IR combination (plotted in brown), while fastest convergence occurred with RGB-only modality (plotted in red). When using additional modalities like infrared (plotted in cyan) or depth (plotted in purple), models showed lower loss values than RGB-only. Validation loss values closely matched training loss values after convergence, proving the model learned accurate naked peach target features.

For bagged peach detection (Figure 7: see original paper), the model started overfitting earlier with RGB-only images compared to adding infrared, depth, or both modalities. The model converged at approximately 100 epochs then began overfitting. RGB-only showed fastest convergence but suffered severe overfitting. For training curves, loss reached lower values with RGB-only, but validation losses showed opposite results, indicating overfitting. Introducing infrared and depth modalities enabled stronger overfitting avoidance at the expense of slightly more iterations. Compared to naked peaches, infrared and depth modalities made greater contributions to improving model generalization and overfitting alleviation, with best results achieved using all imaging modalities.

3.3.2 Quantitative Analysis of Test Results Regarding the test set, Figure 8: see original paper and (b) present mAP for naked and bagged peaches with different modality combinations versus epochs. details detection results using four modality combinations. Comparing RGB images and four-channel images (Table 3, rows 1-3), RGB with additional infrared modality offered best performance with mAPs of 94.7% and 78.2% for naked and bagged peaches, followed by RGB-only with 93.3% and 72.4%, respectively. The least valuable combination was RGB+Depth, which was even less effective than RGB-only. Best results were obtained when combining all modalities, achieving mAPs of 98.6% and 88.9% for naked and bagged peaches. The most important benefit of introducing infrared and depth modalities was the precision metric for bagged peach detection, increasing by 19.1% from 69.2% (RGB) to 88.3% (RGB+Depth+IR). This is because extra geometric information from infrared and depth modalities helps reduce false positives. The recall metric also increased but less significantly. Using RGB+Depth+IR improved mAP by 5.3% and 16.5% over RGB-only for naked and bagged peaches, respectively.

Between infrared and depth modalities, infrared contributed more to mAP improvement. Regarding inference speed, the increase in image channels only affected the first convolutional layer, making the additional computational cost negligible for the whole network. These results demonstrate that three-

dimensional spatial geometry and backscattered signal intensity information from infrared and depth modalities can effectively improve fruit detection accuracy, especially for bagged peaches. It is always more challenging to detect bagged peaches than naked peaches in orchards.

The improved YOLOv5s model was optimized by TensorRT to increase inference speed on Jetson Nano. The model supports three precision types: floating-point 32 (FP32), floating-point 16 (FP16), and integer 8 (INT8). Since Jetson Nano does not support INT8 optimization, the model was converted to FP32 and FP16 operations, achieving detection speeds of 14 and 19 FPS, respectively —implementing 14 and 19 detections per second on five-channel multi-modal images. Therefore, the improved YOLOv5s model optimized by TensorRT-FP16 precision was selected for deployment on the Jetson Nano development board, which is adequate for computer-vision-based peach detection and harvesting.

3.4 Contribution of Different Imaging Modalities in Typical Scenarios

To explore each imaging modality' s contribution to peach detection under different orchard environments, we analyzed test set visualization results under typical scenarios: different illumination conditions, fruit occlusion levels, and camera distances. Note that when concurrent fruit occlusion occurs, model output follows labeling rules from Section 2.2, meaning the priority sequence is OB > OF > OL > NO.

3.4.1 Comparison Under Different Illumination Conditions [Figure 9: see original paper] shows detection results for multi-class naked and bagged peaches using different modality combinations under three typical illumination conditions. Peach trees were approximately 1 m from the camera. For each condition, four detection results are presented based on input data type: RGB (first row), RGB+Depth (second row), RGB+IR (third row), and all modalities simultaneously (fourth row), with odd columns showing naked peaches and even columns showing bagged peaches.

Under normal and strong illumination, the model using RGB+Depth performed even worse than RGB-only, suffering more missing detections. This is because ToF-based depth cameras in outdoor environments are prone to noise interference from sunlight exposure, with accuracy decreasing as measurement distance increases. Consequently, fusion with RGB images may cause misjudgment. Although depth images are unsuitable for peach detection in direct sunlight, they contribute to better results under artificial illumination. At night, the depth camera is not interfered with by sunlight noise and helps accurately reconstruct peach shape. Especially for some OB and OF peaches, RGB images show non-colored, invisible peach edges with similar colors between peaches and leaves, while depth images show highly distinctive geometric features.

When comparing results before and after using infrared modality in daytime natural environments, a reduction in false positives for NO and OL peaches was

observed, especially for bagged peaches. This is because significant differences exist in infrared intensity between fruits and leaves during daytime, enabling infrared images to effectively distinguish fruits from backgrounds under bright illumination. Therefore, depth images help reduce false and missing detections under artificial illumination, while infrared images improve detection accuracy in bright illumination compared to RGB-only. Nevertheless, using all imaging modalities simultaneously yields the best results in any illumination environment.

3.4.2 Comparison Under Different Occlusion Levels As shown in [Figure 10: see original paper], further experiments analyzed different imaging modalities' contributions to multi-class naked and bagged peach detection across occlusion levels. Based on the proportion of peaches occluded by branches and leaves, occlusion levels were classified as Slight occlusion, General occlusion, and Severe occlusion.

Under Slight occlusion (first and second columns), some peaches were missed when using only RGB images, whereas all naked and bagged peaches were accurately detected after fusing Depth or Infrared images. As bagged peaches were wrapped in similarly colored bags (e.g., OF peaches), several overlapping peaches were not correctly detected by the RGB detector. Meanwhile, naked peaches that were occluded and overlapping could be correctly detected after fusing Depth or Infrared images.

Under General occlusion (third and fourth columns), fusing infrared and depth channels provided deeper peach target features, and five-channel fusion showed more significant improvement in accuracy and recall compared to other combinations. Under Severe occlusion (fifth and sixth columns), although missing detections still occurred even with five-channel images, results were significantly better than other multi-modal combinations. Multi-modal images were more effective in improving bagged peach detection accuracy than naked peaches, substantially reducing missing and false detection rates.

Thus, introducing infrared and depth modalities provides the model with more valuable information (e.g., geometric features), improving fruit detection accuracy and recall even for severely occluded peaches.

3.4.3 Comparison Under Different Camera Distances [Figure 11: see original paper] shows the effect of different imaging modalities on multi-class naked and bagged peach detection at various camera distances. As mentioned in Section 2.1, close distance was within 0.3 m from the tree trunk, average distance was 0.3-1 m, and far distance was greater than 1 m.

In close distance scenes (first and second columns), where some peaches were less than 0.2 m from the camera and others within 0.3 m, the RGB-only model achieved best detection results for both naked and bagged peaches. In contrast, models fusing depth and infrared images suffered many missing and false de-

tections. What' s worse, peaches within 0.2 m failed to be detected accurately. This is because depth information was obtained based on the ToF mechanism, which has operational distance requirements for the Azure Kinect DK camera. When peaches were not within the camera' s operational distance, depth and infrared information was lost, negatively affecting detection after fusion with RGB images.

Similar results occurred in far distance detection, where some distant and severely occluded peaches failed to be detected even with five-channel images. At average distances (0.3-1 m from tree trunk), best detection results were achieved using five-channel fusion.

Therefore, appropriate camera operating distance is required to improve peach detection accuracy through infrared and depth modalities. In conclusion, introducing infrared and depth channels can improve occluded fruit detection accuracy, but only when the camera operates within an appropriate distance.

3.5 Ablation Experiments of Improved YOLOv5s

To demonstrate the effectiveness of YOLOv5s improvements, we conducted ablation studies using all imaging modalities for multi-class peach detection. The comparison includes a baseline and three cases. The baseline model was original YOLOv5s without attention mechanism or depthwise separable convolution. Then, coordinate attention (CA) and depthwise separable convolution (DSC) were integrated into YOLOv5s separately to enhance learning of important information and reduce model parameters. The network with DSC is denoted YOLOv5s-DSC, while the network with only CA is YOLOv5s-CA. Results were compared using the same RGB training dataset as Section 3.3 plus corresponding infrared and depth images.

As summarized in , four comparison experiments investigated CA and DSC module performance. When embedding attention mechanism into YOLOv5s, YOLOv5s-CA achieved 98.8% mAP for naked peaches, increasing 3% over YOLOv5s and outperforming all other models. After substituting regular convolution with DSC, YOLOv5s-DSC achieved 95.0% mAP, slightly lower than YOLOv5s. Notably, YOLOv5s-CA and YOLOv5s-DSC achieved 89.4% and 80.0% mAP for bagged peaches, respectively. YOLOv5s-DSC had only 5.03 M parameters—39.9% fewer than original YOLOv5s and the smallest among all models. In terms of detection speed, YOLOv5s-DSC was 30.5% faster than YOLOv5s, achieving 77.5 FPS, which was 14.6% faster than YOLOv5s-CA.

After fusing both DSC and CA modules, the improved YOLOv5s model achieved better results than original YOLOv5s and YOLOv5s-DSC, increasing mAP by 2.8% and 6.2% for naked and bagged peach detection, respectively. The improved YOLOv5s mAP decreased slightly compared to YOLOv5s-CA but achieved 77.5 FPS detection speed—14.6% faster than YOLOv5s-CA. Overall, these results demonstrate that introduced CA and DSC effectively improve detection accuracy and reduce computational cost, enabling the model to detect

in-field peaches with faster speed and higher accuracy while requiring fewer parameters.

3.6 Comparison and Discussion

3.6.1 Comparison with Other Object Detection Networks To further analyze improved YOLOv5s performance, we compared it with three lightweight object detection networks: YOLOX-Nano[36], PP-YOLO-Tiny[37], and EfficientDet-D0[38]. The same training, validation, and test sets with five-channel images (RGB+Depth+IR) were used. Detection results on the test set are shown in .

Improved YOLOv5s achieved best results in precision, recall, and mAP compared to the other three networks. The mAP of improved YOLOv5s was 98.6%—22.9%, 18.1%, and 10.9% higher than YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0 for naked peaches, respectively. For bagged peaches, improved YOLOv5s also achieved best mAP, which was 16.3%, 8.1%, and 4.5% higher than YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0, respectively. Although improved YOLOv5s average detection speed was 77.5 FPS—slower than the other three networks—detection accuracy was effectively improved. These results demonstrate that the peach detection network based on improved YOLOv5s can detect peaches more effectively and accurately than other lightweight networks.

3.6.2 Comparison with Other Fruit Detection Studies Additionally, we assessed improved YOLOv5s effectiveness on two open-source fruit datasets: apple and kiwifruit. Gené-Mola et al.[2] presented a Fuji apple dataset acquired at night using a depth camera and used Faster R-CNN for detection. Suo et al.[18] classified a kiwifruit dataset into five classes based on occlusion status and used YOLOv4. Since these datasets had different label classifications and image resolutions, we standardized them to match our experimental conditions. Parameters like momentum, learning rate, and weight decay followed the original YOLOv5s model. Both datasets were split into training and test sets for improved YOLOv5s training and testing.

Experimental results in reveal that the proposed improved YOLOv5s offered better results to different degrees than other methods for fruit detection, verifying effectiveness across different detection tasks.

---

4 Conclusions

Developing effective methods to detect fruits with different agronomic measurements is crucial for improving mechanical harvesting popularity. This paper presents a publicly available multi-class RGB-D dataset of natural naked and bagged peaches—the first multi-class peach detection dataset. We propose an

improved multi-class peach detector based on YOLOv5s by fusing multi-modal images as input and introducing coordinate attention mechanism and depthwise separable convolution.

Experimental results show that improved YOLOv5s using multi-modal visual data achieves detection mAPs of 98.6% and 88.9% for naked and bagged peaches in complex illumination and severe occlusion environments—increasing by 5.3% and 16.5% over RGB-only images, and by 2.8% and 6.2% compared to YOLOv5s. For bagged peach detection, improved YOLOv5s achieves best mAP, which is 16.3%, 8.1%, and 4.5% higher than YOLOX-Nano, PP-YOLO-Tiny, and EfficientDet-D0, respectively. The improved YOLOv5s with multi-modal visual data enhances model perception ability for detecting both naked and bagged peaches under severe occlusion and various illumination conditions.

Specifically, depth imaging modality can reduce false and missing detections under artificial illumination, while infrared imaging modality can improve detection accuracy under strong illumination. Additionally, the proposed detection model achieves 19 detections per second with five-channel multi-modal images on popular embedded platforms, meeting real-time fruit harvesting system requirements.

The main limitation of using five-channel multi-modal images is underutilization of spatial geometric information in depth and infrared images. Future work includes exploring stronger fruit detection networks and multi-modal image fusion methods to further improve in-field bagged fruit detection, including various bag types.

References

[1] YADAV S, SENGAR N, SINGH A, et al. Identification of disease using deep learning and evaluation of bacteriosis in peach leaf[J]. Ecological Informatics, 2021, 61: ID 101247.

[2] GENE-MOLA J, VILAPLANA V, ROSELL-POLO J R, et al. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities[J]. Computers and Electronics in Agriculture, 2019, 162: 689-698.

[3] NGUYEN T T, VANDEVOORDE K, WOUTERS N, et al. Detection of red and bicoloured apples on tree with an RGB-D camera[J]. Biosystems Engineering, 2016, 146: 33-44.

[4] LIU X, JIA W, RUAN C, et al. The recognition of apple fruits in plastic bags based on block classification[J]. Precision Agriculture, 2018, 19(4): 735-749.

[5] LIU T, EHSANI R, TOUDESHKI A, et al. Identifying immature and mature pomelo fruits in trees by elliptical model fitting in the Cr-Cb color space[J]. Precision Agriculture, 2019, 20(1): 138-156.

[6] LIU Y, CHEN B, QIAO J. Development of a machine vision algorithm for recognition of peach fruit in a natural scene[J]. Transactions of the ASABE, 2011, 54(2): 613-626.

[7] WILLIAMS H A M, JONES M H, NEJATI M, et al. Robotic kiwifruit harvesting using machine vision, convolutional neural networks, and robotic arms[J]. Biosystems Engineering, 2019, 181: 140-156.

[8] NAVAS E, FERNANDEZ R, SEPULVEDA D, et al. Soft grippers for automatic crop harvesting: A review[J]. Sensors, 2021, 21(8): ID 2689.

[9] TU S, PANG J, LIU H, et al. Passion fruit detection and counting based on multiple scale faster R-CNN using RGB-D images[J]. Precision Agriculture, 2020, 21(5): 1072-1091.

[10] HÄNI N, ROY P, ISLER V. A comparative study of fruit detection and counting methods for yield mapping in apple orchards[J]. Journal of Field Robotics, 2020, 37(2): 263-282.

[11] LU S, CHEN W, ZHANG X, et al. Canopy-attention-YOLOv4-based immature/mature apple fruit detection on dense-foliage tree architectures for early crop load estimation[J]. Computers and Electronics in Agriculture, 2022, 193: ID 106696.

[12] LI X, PAN J, XIE F, et al. Fast and accurate green pepper detection in complex backgrounds via an improved YOLOv4-tiny model[J]. Computers and Electronics in Agriculture, 2021, 191: ID 106503.

[13] JIANG M, SONG L, WANG Y, et al. Fusion of the YOLOv4 network model and visual attention mechanism to detect low-quality young apples in a complex environment[J]. Precision Agriculture, 2022, 23(2): 445-461.

[14] HUANG H, HUANG T, LI Z, et al. Design of citrus fruit detection system based on mobile platform and edge computer device[J]. Sensors, 2021, 22(1): ID 59.

[15] FU L, GAO F, WU J, et al. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review[J]. Computers and Electronics in Agriculture, 2020, 177: ID 105687.

[16] SA I, GE Z, DAYOUB F, et al. Deepfruits: A fruit detection system using deep neural networks[J]. Sensors, 2016, 16(8): ID 1222.

[17] ARAD B, BALENDONCK J, BARTH R, et al. Development of a sweet pepper harvesting robot[J]. Journal of Field Robotics, 2020, 37(6): 1027-1039.

[18] SUO R, GAO F, ZHOU Z, et al. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking[J]. Computers and Electronics in Agriculture, 2021, 182: ID 106052.

[19] TIAN Y, YANG G, WANG Z, et al. Apple detection during different growth

stages in orchards using the improved YOLO-v3 model[J]. Computers and Electronics in Agriculture, 2019, 157: 417-426.

[20] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: Unified, real-time object detection[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2016: 779-788.

[21] REDMON J, FARHADI A. YOLOv3: An incremental improvement[J/OL]. arXiv:1804.02767[cs.CV], 2018.

[22] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2017: 7263-7271.

[23] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. arXiv:2004.10934[cs.CV], 2020.

[24] YAN B, FAN P, LEI X, et al. A real-time apple targets detection method for picking robot based on improved YOLOv5[J]. Remote Sensing, 2021, 13(9): ID 1619.

[25] LIU S, QI L, QIN H, et al. Path aggregation network for instance segmentation[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2018: 8759-8768.

[26] HOU Q, ZHOU D, FENG J. Coordinate attention for efficient mobile network design[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2021: 13713-13722.

[27] FANG L, WU Y, LI Y, et al. Ginger seeding detection and shoot orientation discrimination using an improved YOLOv4-LITE network[J]. Agronomy, 2021, 11(11): ID 2328.

[28] SHI C, LIN L, SUN J, et al. A lightweight YOLOv5 transmission line defect detection method based on coordinate attention[C]// 2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC). Piscataway, New York, USA: IEEE, 2022, 6: 638-642.

[29] ZHA M, QIAN W, YI W, et al. A lightweight YOLOv4-based forestry pest detection method using coordinate attention and feature fusion[J]. Entropy, 2021, 23(12): ID 1624.

[30] HU J, SHEN L, SUN G. Squeeze-and-excitation networks[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2018: 7132-7141.

[31] ZHANG Y, YU J, CHEN Y, et al. Real-time strawberry detection using deep neural networks on embedded system (RTSD-net): An edge AI application[J]. Computers and Electronics in Agriculture, 2022, 192: ID 106586.

[32] CHOLLET F. Xception: Deep learning with depthwise separable convolutions[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2017: 1251-1258.

[33] BOCHKOVSKIY A, WANG C Y, LIAO H Y M. YOLOv4: Optimal speed and accuracy of object detection[J/OL]. arXiv:2004.10934[cs.CV], 2020.

[34] ZHENG Z, WANG P, LIU W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Piscataway, New York, USA: IEEE, 2020, 34(7): 12993-13000.

[35] POWERS D M W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation[J/OL]. arXiv:2010.16061[cs.LG], 2020.

[36] GE Z, LIU S, WANG F, et al. YOLOx: Exceeding yolo series in 2021[J/OL]. arXiv:2107.08430[cs.CV], 2021.

[37] LONG X, DENG K, WANG G, et al. PP-YOLO: An effective and efficient implementation of object detector[J/OL]. arXiv:2007.12099[cs.CV], 2020.

[38] TAN M, PANG R, LE Q V. EfficientDet: Scalable and efficient object detection[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway, New York, USA: IEEE, 2020: 10781-10790.

Note: Figure translations are in progress. See original paper for figures.