Detecting and Mapping Individual Fruit Trees in Complex Natural Environments via UAV Remote Sensing and Optimized YOLOv5

Yongzhu Xiong, *Member, IEEE*, Xiaofeng Zeng, Weiqian Lai, Jiawen Liao, Yankui Chen, Mingyong Zhu, and Kekun Huang, *Member, IEEE*

***Abstract*—The location and number of individual fruit trees (IFTs) are critical for investigations on planting areas, fruit yield predictions, and smart orchard planning and management. These data are conventionally obtained through manual and statistical investigations that require long, laborious, and costly efforts. Object detection models of deep learning could provide an opportunity to detect IFTs accurately, which is essential for rapidly obtaining these data and reducing human operation errors. This study proposed an approach for detecting IFTs and mapping their spatial distributions by integrating deep learning with unmanned aerial vehicle (UAV) remote sensing. UAV remote sensing was used to collect high-resolution images of fruit trees in pomelo orchards in Meizhou, South China. Based on these images, a new individual pomelo tree image sample dataset was created through manual interpretation and field investigation. The evaluation results revealed that YOLOv5s was the best model among the five YOLOv5 models (i.e., YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, whose layers, parameters, and floating-point operations all increased with the depth and width of layers) of different scales considered for optimization. Moreover, the coordinate attention (CA) optimized YOLOv5 model (YOLOv5s-CA) is the best model (named FruitNet) with the best overall accuracy for detecting IPTs among all seven attention-optimized YOLOv5 models and other state-of-the-art object detection models, such as faster R-CNN and YOLOv8s. The IPTs in the study areas were detected using FruitNet, their number and planting area were counted, and their spatial distributions were mapped based on the predicted results of the IPTs. This study suggested that our proposed approach could provide key data and technical support for smart orchard management.**

***Index Terms*—Deep learning, individual tree detection, remote sensing, spatial distribution, unmanned aerial vehicle (UAV).**

I. Introduction

SPATIAL and attribute data on individual fruit trees (IFTs) in orchards play an important role in the accurate surveys of planting areas, rapid pest and disease control, precise fruit yield prediction, and smart orchard management [1]. Field investigations and statistics, including spatial distribution, location, and number of IFTs, are traditionally used to collect these data orchards. These investigations are time-consuming, labor-intensive, and costly [2], and the data obtained from traditional methods cannot meet the needs of smart orchards due to the insufficiently accurate IFT location information they provide. Therefore, developing a fast, inexpensive, and accurate method for investigating and mapping IFTs is necessary to obtain these data to advance the development of precision agriculture.

Remote sensing images of fruit trees in relatively large orchards can be captured via satellite or aerial imaging. In the case of satellite remote sensing, cloudy weather is a major challenge that makes it difficult to detect fruit trees due to the possibly poor quality of the images captured [3], [4]. The limitation of the spatial resolution of satellite images is another major challenge for the accurate detection of IFTs. Aerial imaging includes photography using a manned or unmanned aircraft [5]. Human-crewed aircraft are unsuitable for detecting IFTs due to high costs and inconvenient operations. Unmanned aerial vehicle (UAV) remote sensing is the best alternative for capturing images of fruit trees in relatively large orchards. UAV remote sensing has the advantages of automation, intelligence, and specialization and can be used to quickly obtain spatial remote sensing information about land, resources, environments, events, etc., and to conduct real-time processing, modeling, and analysis of advanced emerging aerial remote sensing technology solutions [6]. IFT detection is a fundamental task for UAV-based

Manuscript received 7 September 2023; revised 1 January 2024 and 3 March 2024; accepted 13 March 2024. Date of publication 22 March 2024; date of current version 5 April 2024. This work was supported in part by the Guangdong Provincial Key-Field Special Project for Universities (New Generation Information Technology) under Grant 2020ZDZX3044, in part by Guangdong Provincial Science and Technology Innovation Strategy Special Fund (Climbing Plan) Project under Grant PDJH2020b0552, in part by the Guangdong Provincial Key Laboratory of Conservation and Precision Utilization of Characteristic Agricultural Resources in Mountainous Areas under Grant 2020B121201013, in part by Guangdong Pomelo Engineering Technology Development Center under Grant 2019GCZX007, in part by the National Natural Science Foundation of China under Grant 61976104, and in part by Guangdong Provincial Natural Science Foundation under Grant 2024A1515011986. *(Corresponding author: Yongzhu Xiong.)*

Yongzhu Xiong, Yankui Chen, and Mingyong Zhu are with the School of Geography and Tourism, Jiaying University, Meizhou 514015, China, and also with the Guangdong Provincial Key Laboratory of Conservation and Precision Utilization of Characteristic Agricultural Resources in Mountainous Areas, Meizhou 514015, China (e-mail: xiongyz@jyu.edu.cn; 519958394@qq.com; 201701027@jyu.edu.cn).

Xiaofeng Zeng and Weiqian Lai were with the School of Geography and Tourism, Jiaying University, Meizhou 514015, China. They are now with the Guangdong Airace Technology Company Ltd., Huizhou 516001, China, and also with the Satxspace Technology Company Ltd., Dongguan 523830, China (e-mail: zengxf12123@163.com; 1299136522@qq.com).

Jiawen Liao is with the School of Geography and Tourism, Jiaying University, Meizhou 514015, China (e-mail: ljw1257871614@gmail.com).

Kekun Huang is with the School of Mathematics, Jiaying University, Meizhou 514015, China (e-mail: kkcocoon@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3379522

site-specific management in orchards [7]. UAV remote sensing has great potential for acquiring the image data from IFTs in orchards quickly and economically for integration with deep learning to carry out IFT detection.

Over the past decade, with the rapid progress of computer hardware and the rapid development of artificial intelligence technology, convolutional neural networks (CNNs) used in deep learning have been used to pioneer new ways to detect objects and extract features in remote sensing images [8], [9], [10], [11], [12]. Many CNN architectures have been proposed for object detection in computer vision and image analysis, and they can be divided into two categories: two-stage and one-stage models [13]. Two-stage models divide the detection process into region proposal and classification stages, while one-stage detectors contain a single feedforward fully convolutional network that directly provides bounding boxes and object classification. Girshick et al. [14] proposed a two-stage object detection region-based CNN (R-CNN) model based on classification problems. Based on R-CNN, fast R-CNN and faster R-CNN were subsequently proposed to improve the efficiency and accuracy of object detection. Two years later, Redmon et al. [15] proposed a single-stage-based object detection model called You Only Look Once (YOLO). The YOLO model not only simplifies the size of the neural network but also improves the detection speed and accuracy. Etten [16] proposed a You Only Look Twice (YOLT) pipeline based on YOLOv2 to achieve rapid multiscale object detection in large-scale satellite imagery. Yan et al. [17] recognized *Rosa roxbunghii* in the natural environment based on an improved faster R-CNN, for which the average recognition accuracy was 92.01% for 11 classes of *Rosa roxbunghii* fruit. Xiong et al. [18] demonstrated that deep transfer learning based on YOLOv2 could be used as a new archaeological remote sensing method to detect historical buildings accurately and rapidly in aerial photographs. Liu and Wang [19] proposed a tomato disease and pest detection algorithm based on YOLO, whose results showed that YOLO outperforms faster R-CNN. Xiong et al. [20] proposed a visual detection method that used UAV images and YOLOv2 to rapidly detect green mangoes on the surface of tree crowns and estimate the number of mango fruits in orchards. YOLO-v5m can be a useful component of an automated plantation management system and helps forecast date production and monitor the condition of date palm trees [21]. Recently, Liu et al. [22] proposed a method for detecting and localizing pineapples in natural environments based on binocular stereo-vision and an improved YOLOv3 model. The above research indicates that deep learning object detection via YOLO models in remotely sensed images has drawn great interest from researchers from multidisciplinary communities, including remote sensing, computer vision, and precision agriculture. However, there is still not in consensus on which YOLO model can best be used to detect objects, especially individual trees.

More recently, several deep learning object detection and segmentation models have been adopted for the detection and segmentation of individual trees, such as olive, palm, and coconut trees, based on high-resolution visible and light detection and ranging (LiDAR) images acquired from satellites and UAVs [5], [21], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43]. For instance, Santos et al. [44] proposed and evaluated the use of CNN-based methods combined with UAV high spatial-resolution red–green–blue (RGB) imagery for the detection of law-protected tree species and reported that RetinaNet achieved more accurate results than did faster R-CNN and YOLOv3. An analysis of satellite images by Brandt et al. [34] pinpointed individual tree canopies over a large area of West Africa. Their results suggest that, with certain limitations, it will soon be possible to map the location and size of every tree worldwide using deep learning [34], [35]. Safonova et al. [45] used mask R-CNN and UAV images for olive tree crown and shadow segmentation to further estimate the biovolume of individual trees. Jintasuttisak et al. [21] carried out the automatic detection of crowded date palm trees in drone imagery using YOLOv5 and a comparison of one-stage object detection methods [YOLOv3, YOLOv4, and the single-shot multibox detector (SSD300)] for date palm tree detection. The authors found that the YOLOv5m model had the highest accuracy with a mean average precision of 92.34%. Hu et al. [43] presented a pipeline for monitoring and clustering of 259 peach tree crowns based on the UAV images of a peach orchard in Southeast China and designed conditional generative adversarial networks to extract the crown area. Yu et al. [46] showed that the mask R-CNN model achieved the highest accuracy in individual tree detection compared with the local maxima algorithm and marker-controlled watershed segmentation. Most of the detection objects (trees) mentioned above are sparsely and evenly distributed in orchards or areas under simple conditions but not complex natural conditions (e.g., the natural environment of pomelo trees planted in plains and hilly areas and mixed with other roadside and forest trees, such as loquat trees, lychee trees, and eucalyptus trees).

Several recent studies have demonstrated that the integration of deep learning CNNs with UAV images can be used to realize the accurate detection of individual trees, including fruit trees, such as coconut [47], citrus [48], and peach [43] trees. However, most of those previous works needed to improve the architecture of CNN networks to increase the performance of deep learning models. Yuan et al. [49] proposed an improved YOLOx-nano algorithm to detect pomelo trees and compared it with several state-of-the-art object detection algorithms, such as faster R-CNN, SSD, YOLOv3, and YOLOv4-tiny. Their method showed better suitability for pomelo tree detection in UAV images with better performance and fewer parameters. However, accurately carrying out IPT detection and mapping in a large orchard area in a complex natural environment while interfering with other coexisting objects for smart orchard management is still a challenge. One reason for this is the need for open-source, high-resolution, and high-quality IPT samples for training and validation because preparing these samples is extremely time-consuming and labor-intensive. Most of those previous studies on individual trees focused on other fruit trees or forest trees, such as coconut, palm, olive, or pine trees, yet did not involve pomelo trees, which are widely planted in South China (e.g., in the provinces of Guangdong, Guangxi, Fujian, Jiangxi, and Hunan), Southeast Asia, and South America. There is a lack of a

dataset of high-quality IPT samples. The second and key reason is that a specific fruit tree detection model based on deep learning requires training, validation, and testing on a large dataset of high quality with massive computational resources since deep learning is a data-driven science and application. At present, there is no CNN model trained for IPT detection on the IPT dataset. Although there has been substantial work in individual tree detection using deep learning and UAV images, previous models cannot work well for pomelo orchards for the above two reasons. This topic merits in-depth study on optimizing IPT detection models, especially regarding mapping IPTs with an optimal deep learning model for use in complex natural environments.

Inspired by the great progress in deep learning and UAV remote sensing, an approach to establishing the accurate IFT detection model by integrating an optimized YOLOv5 model with UAV remote sensing images of IFTs is proposed to address the gap mentioned above. With this approach, the spatial distribution of IFTs can be rapidly and accurately mapped, and the number and planting area of IFTs can also be quickly determined for precision agriculture and smart orchard management. It is hypothesized that IFTs can be detected in high-resolution true-color images at low cost, with high accuracy and with high efficiency and that a thematic map of IFTs can be made using our proposed approach. Pomelo trees were selected as the fruit trees in this study, and YOLOv5 was empirically selected as the baseline deep learning model to test our hypotheses. Based on the previous studies, multiple state-of-the-art channels and/or spatial attention mechanisms were applied to optimize the YOLOv5 models.

This study aimed to integrate deep learning methods with UAV-based images for detecting and mapping individual pomelo trees (IPTs) and to provide a set of reliable and timely basic data and technical support for smart orchard management and precision agriculture development.

The main contributions of our study are given as follows.

1) A novel approach for IPT detection and mapping was proposed.
2) An IPT image sample (IPTIS) dataset was created with high-resolution images captured by UAV remote sensing.
3) YOLOv5s was verified to be the best among all YOLOv5 family models and was optimized by adding seven popular attention mechanisms.
4) The performance and robustness of the seven attention-optimized YOLOv5 models were compared and the best model (called FruitNet) was used to detect IPTs.
5) IPTs in two large-scale pomelo orchards were detected and their locations and spatial distributions were mapped using our proposed approach.

II. Materials and Methods

A. Study Areas and Pomelo Trees

Two large pomelo orchards (site A and site B), which cover the areas of approximately 36.51 ha and 48.15 ha, respectively, were selected as the experimental study areas. These sites are located in the towns of Shishan and Sanxiang, Meixian district in Meizhou, a city (23°23′–24°56′ N and 115°18′–116°56′ E) situated in northeastern Guangdong Province, South China [see Fig. 1(a)]. Meizhou city has jurisdiction over two districts, Meijiang and Meixian; five counties, Pingyuan, Jiaoling, Dabu, Fengshun, and Wuhua; and one city, Xingning [see Fig. 1(b)], with a total land area of 15 876 km^2. Hilly and mountainous Meizhou is a very suitable location for growing fruit trees, such as pomelo trees. The city of Meizhou, which is the hometown of golden pomelo fruit, is the largest pomelo fruit-producing area in Guangdong Province. Its yield accounted for 90% of the total pomelo output of Guangdong Province, 20% of that of China, and 10% of that of the world in 2018. Meixian is one of the regions with the most pomelo trees planted in the city of Meizhou. The towns of Shishan and Sanxiang in Meixian are abundant in pomelo trees and convenient for field investigation and UAV remote sensing monitoring. For this reason, two pomelo orchards in these two towns were selected as the areas of our experimental study. The two study sites lie in a hilly area [see Fig. 1(c)] and a plain surrounded by low mountains [see Fig. 1(d)]. Fig. 1(e) shows a typical pomelo orchard landscape at site A. Fig. 1(g) shows a field scene of IPTs at site A. In addition, another orchard near site A in Shishan (i.e., site B in Sanxiang) was selected to validate the robustness of our optimized model.

Pomelo (*Citrus maxima*) is a large citrus fruit [see Fig. 1(f)] with thick yellow skin that tastes similar to grapefruit but is sweeter than it is. It is a delicious and popular fruit due to its rich nutritional and medicinal value. The pomelo tree is a medium- and large-scale evergreen broad-leaved tree of the family Rutaceae that is more commonly known as Citrus. Most of the pomelo trees in the study areas are 3–10 years old, their heights go to 3–6 m on average, and their crown diameters generally reach 2–8 m. They can be easily recognized in UAV-based high-resolution images by visual interpretation with fieldwork assistance. Therefore, we hypothesize that their crowns can be individually detected via deep learning from UAV-based high-resolution images. However, most pomelo trees grow in hilly areas mixed with other roadside and forest trees, and their crown forms and shapes are also similar to those of surrounding trees; consequently, pomelo trees are not easily distinguished from some other trees in UAV images only from human vision without fieldwork. It is challenging to detect and map IPTs in large orchard areas in complex natural environments.

B. Our Proposed Approach for Detecting and Mapping IFTs

We propose a new systematic approach for detecting IFTs, mapping their spatial distribution, and counting their planting area and number by integrating an attention-optimized YOLOv5 model with high-resolution and low-altitude UAV remote sensing images. The workflow of our proposed approach is illustrated in Fig. 2 and is composed of the following six phases.

1) Capturing and preprocessing UAV images.
2) Creating a new dataset of individual fruit tree image samples (IFTIS).
3) Training and validating the five standard YOLOv5 models on the IFTIS to evaluate their performance.

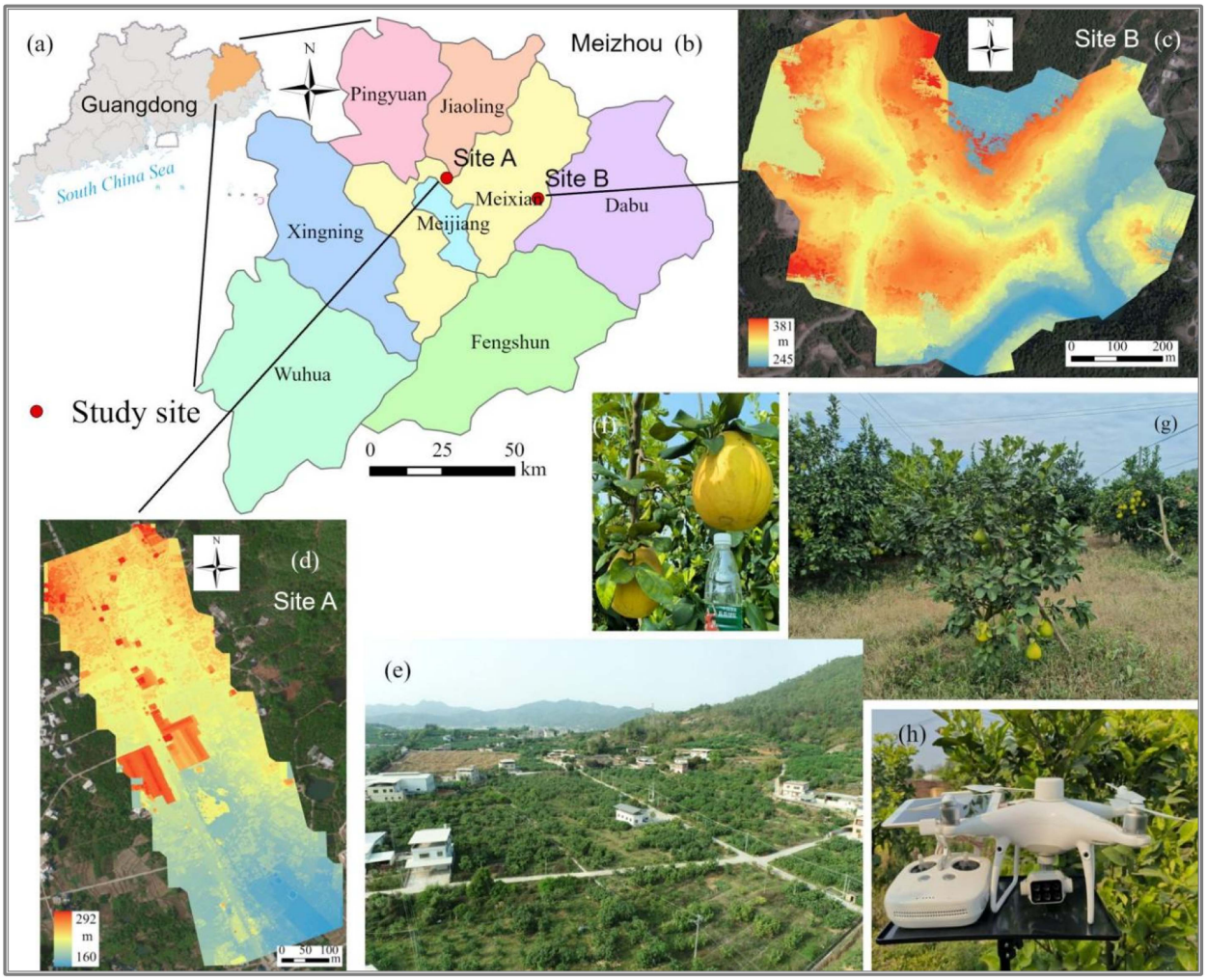

Fig. 1. Overview of the experimental study areas. (a) Location of Meizhou in Guangdong. (b) Administrative map of Meizhou showing the study areas (sites A and B) in red circles. (c) DSM map of site B. (d) DSM map of site A. (e) Typical pomelo orchard landscape at site A. (f) Matured pomelos on the trees. (g) Field scene showing the IPTs at site A. (h) DJI Phantom 4 multispectral drone with RTK photographed near a pomelo tree at site A.

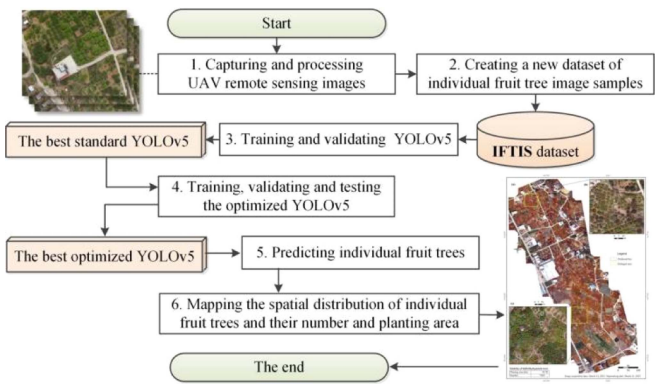

Fig. 2. Workflow chart of our proposed approach for detecting and mapping IFTs via YOLOv5 integrated with UAV remote sensing. IFTIS denotes the individual fruit tree image samples.

4) Selecting the YOLOv5 model with the best performance for further optimization by using multiple attention mechanisms and training them on the IFTIS.
5) Predicting IFTs by the attention-optimized YOLOv5 model with the best performance (named FruitNet).
6) Making thematic maps of the IFTs of the study areas based on the predicted results of FruitNet.

To test and validate the feasibility of our proposed approach, we carried out an experimental study on IPT detection and mapping in Meizhou. The key procedures of our proposed approach are described in detail in the following sections.

1) UAV Remote Sensing Image Capture and Preprocessing: The DJI[1] Phantom 4 multispectral drone [see Fig. 1(h)] was used as a UAV system to capture low-altitude high-resolution remote sensing images. It is equipped with six 1/2.9″ complementary metal–oxide–semiconductors, including one RGB sensor for visible light imaging and five monochrome sensors (blue, green, red, red-edge, and near-infrared bands) for multispectral imaging. It uses a real-time kinematic (RTK) enabled global navigation satellite system (GNSS), including the global positioning system, BeiDou, and Galileo. It can provide efficient

[1][Online]. Available: www.dji.com

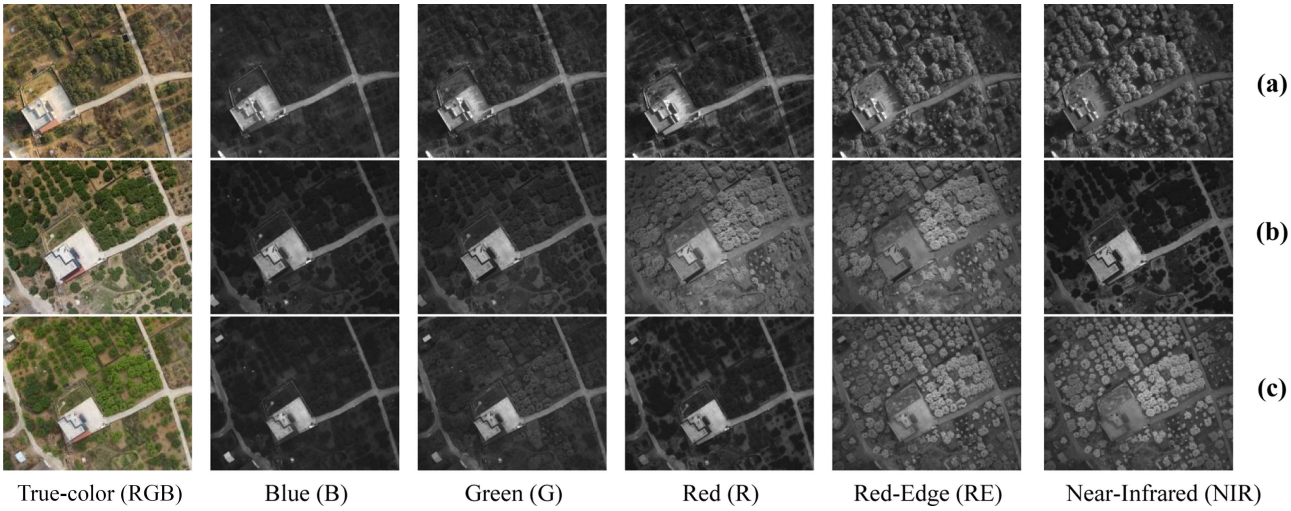

Fig. 3. Examples of the raw true-color (RGB) and multispectral (five bands) images captured by UAV remote sensing. (a) 3 February 2021. (b) 12 March 2021. (c) 12 April 2021.

tools for farmers and researchers in precision agriculture, greatly improving the efficiency of environmental data acquisition. In this study, high-quality true-color and multispectral remote sensing images were captured using this drone without the ground control points required in traditional aerial surveys.

Because pomelo trees exhibit different morphological and spectral characteristics during different growth and phenological periods, it is particularly necessary to construct a dataset of pomelo samples consisting of UAV remote sensing images from different phenological periods. In the present study, three flight plans for aerial photography tasks were set to capture images from site A on three days in late winter and spring 2021 (i.e., 3 February, 12 March, and 12 April 2021, during which pomelo trees develop spring sprouts and grow quickly), and one flight plan was used for capturing images from site B on 16 January 2022. A flight altitude of 120 m with both 60% heading and lateral overlaps was set to capture UAV raw images, resulting in a spatial resolution of approximately 0.06 m for sites A and B. To obtain high-quality UAV raw data, we used the same flight parameters but different extents for the three aerial photography tasks before takeoff. Fig. 3 shows the examples of the raw true-color (RGB) and multispectral (five bands) images collected via UAV aerial photography. Only the true-color images (column 1 in Fig. 3) were used to construct a dataset of true-color image samples of IPTs, while multispectral images were not used in the present study. The reason is that, currently, only RGB images can be used to train standard YOLOv5 object detection models.

The original images obtained by UAV remote sensing on each date were preprocessed to generate a digital orthographic mosaic image model using DJI Terra.[2] The preprocessing steps include the following aspects.

1) Confirming the integrity and quality of original image data, including camera parameters, image clarity, and GNSS information.
2) Establishing project files and importing original image data, performing engineering, adding image data, setting image attributes, and camera model parameters in DJI Terra.
3) Processing the UAV images automatically, including initialization, point cloud generation, 3-D reconstruction, and digital orthographic model (DOM) and digital surface model (DSM) generation in DJI Terra.

The four mosaic DOMs of the study areas generated through these processes are shown in Fig. 4, which reveals the different characteristics of the color tones. Because the pomelo trees grew quickly from February to April, many trees in the images became greener with time. Notably, the three DOMs at site A, including boundaries and shapes, have different actual extents. UAV aerial photography took approximately 40–50 min and three or four flights, depending on the weather conditions and flight point limitations of the DJI P4 drone, to complete the remote sensing image acquisition tasks at each site. The DOM of site B looks different in color because its images were acquired in a different lighting environment compared with those from site A.

2) Creating the Dataset of IPTISs: The processed mosaic orthographic images of pomelo trees were imported into ArcGIS 10.8,[3] where the deep learning module was used to annotate the IPT samples. After labeling, cropping, and exporting, an IPT dataset based on UAV remote sensing image chips was generated and named the IPTISs dataset. The steps are shown in detail as follows.

First, a polygon feature-class shapefile for each month's image was created in ArcGIS 10.8. Professionals manually drew circle features for the pomelo sample annotations according to the records of field investigations. Two fields, ClassName and ClassType, were added to the property tables of these shapefiles, and their values for the IPT sample category were identified

[2][Online]. Available: www.dji.com

[3][Online]. Available: www.esri.com

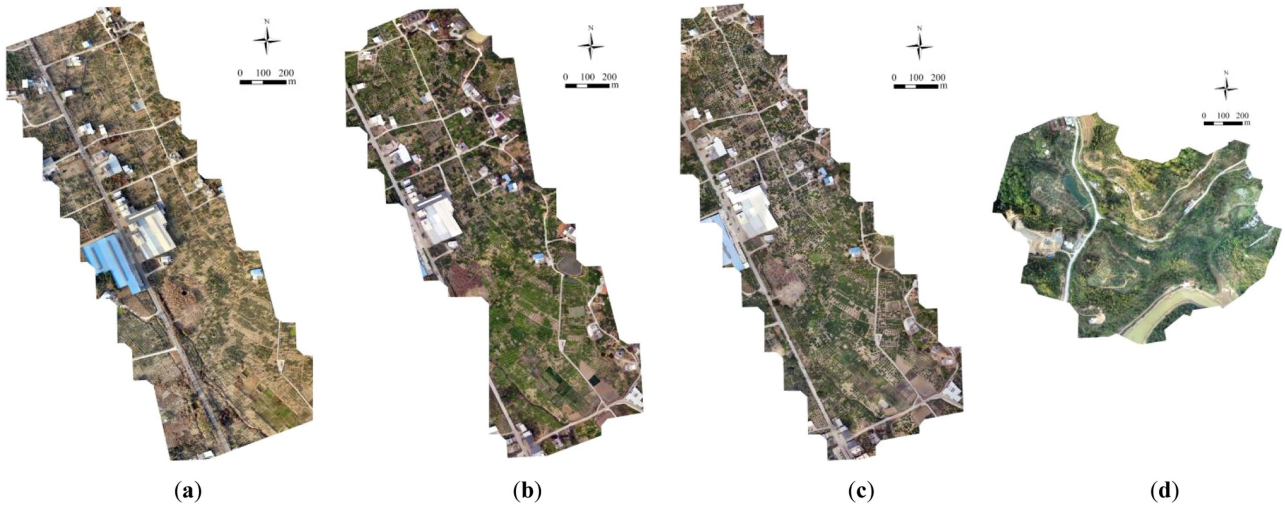

Fig. 4. Digital mosaic orthographic images in the study areas for the four dates; site A with (a) 10 279 × 16 272 pixels, (b) 9453 × 15 186 pixels, and (c) 10 289 × 17 692 pixels and site B with (d) 17 959 × 15 774 pixels. (a) 3 February 2021. (b) 12 March 2021. (c) 12 April 2021. (d) 16 January 2022.

as "P" and "0," respectively. Three annotated clipped image examples are shown in Fig. 5(a1)–(c1).

Second, polygon feature-class shapefiles were used to export the images and their corresponding annotations, which were suitable for subsequent research. The digital orthographic image of the study area taken each month was cropped into clipped images with a size of 640 × 640 pixels and zero overlaps using the deep learning module of ArcGIS. Images without pomelo tree annotations were excluded when exported from ArcGIS software.

Finally, the IPTIS dataset was created in PASCAL VOC data format [50] by combining all the exported clipped images for the three months, for a total of 438 images. Three labeled examples of the clipped images of the dataset obtained after cropping and exporting are shown in Fig. 5(a2)–(c2). The actual label of an IPT is the minimum bounding rectangle of the drawn circle, which is the ground truth for model training, validation, and testing for the task of IPT detection. In addition, 42 chip image samples acquired on 16 January 2022, from another site B in Meixian, were used to supplement the IPTIS dataset, for a total of 480 images. This dataset is small but suitable for proportional division for deep learning training, validation, and testing.

3) Attention-Optimized YOLOv5 Models: YOLO is a state-of-the-art, real-time, one-stage object detection algorithm created by the Ultralytics team in 2015. It uses a single neural network to process an entire image. The image is divided into regions, and the algorithm predicts each region's classes, probabilities, and bounding boxes [51]. Since 2015, the Ultralytics team has been working on improving this model, and many versions of the model have been released. The latest version is YOLOv8, released by Ultralytics in January 2023. It is a cutting-edge, state-of-the-art model that can perform a wide range of object detection, image segmentation, and image classification tasks [52]. The fifth version of this algorithm, YOLOv5, was officially released in June 2020 (YOLOv5-v1.0). Since then, YOLOv5 has been updated several times for a series of releases. The newest one (YOLOv5-7.0) was released in November 2022.

YOLOv5 has been broadly used in the detection of many objects, such as faces, vehicles, and ships, due to its high speed and accuracy. In the present study, YOLOv5-6.0 (henceforth YOLOv5), released in October 2021, was empirically selected for the task of IPT detection due to its ease of use, high performance, and flexibility.

Inherited from its predecessors in the YOLO family [51], the architecture of YOLOv5 consists of two components: backbone and head, and the head network is composed of the neck and detect parts (see Fig. 6).

1) *Backbone:* The backbone network extracts rich feature representations from images. It helps reduce the spatial resolution of the image and increase its feature (channel) resolution.
2) *Neck:* The neck network is used to extract feature pyramids. It helps the model to generalize to the objects of different sizes and scales.
3) *Detect:* The detection network is used to perform the final stage operations. It applies anchor boxes on feature maps and renders the final output: classes, confidence scores, and bounding boxes.

YOLOv5 was released at five different scales (see Table I): YOLOv5n (nano, extra small), YOLOv5s (small), YOLOv5m (medium), YOLOv5l (large), and YOLOv5x (extra large). Their layers, parameters, and floating-point operations (FLOPs) increase with the depth and width of layers, boosting the complexity of neural-network models [51]. Furthermore, the complexity of the model increases accordingly, resulting in more weights and biases being trained and accuracy improvements being pretrained on the COCO dataset [51]. Fig. 6 shows the architecture of the standard YOLOv5s model. The other scales of YOLOv5 have similar architectures but different layers and parameters (see Table I).

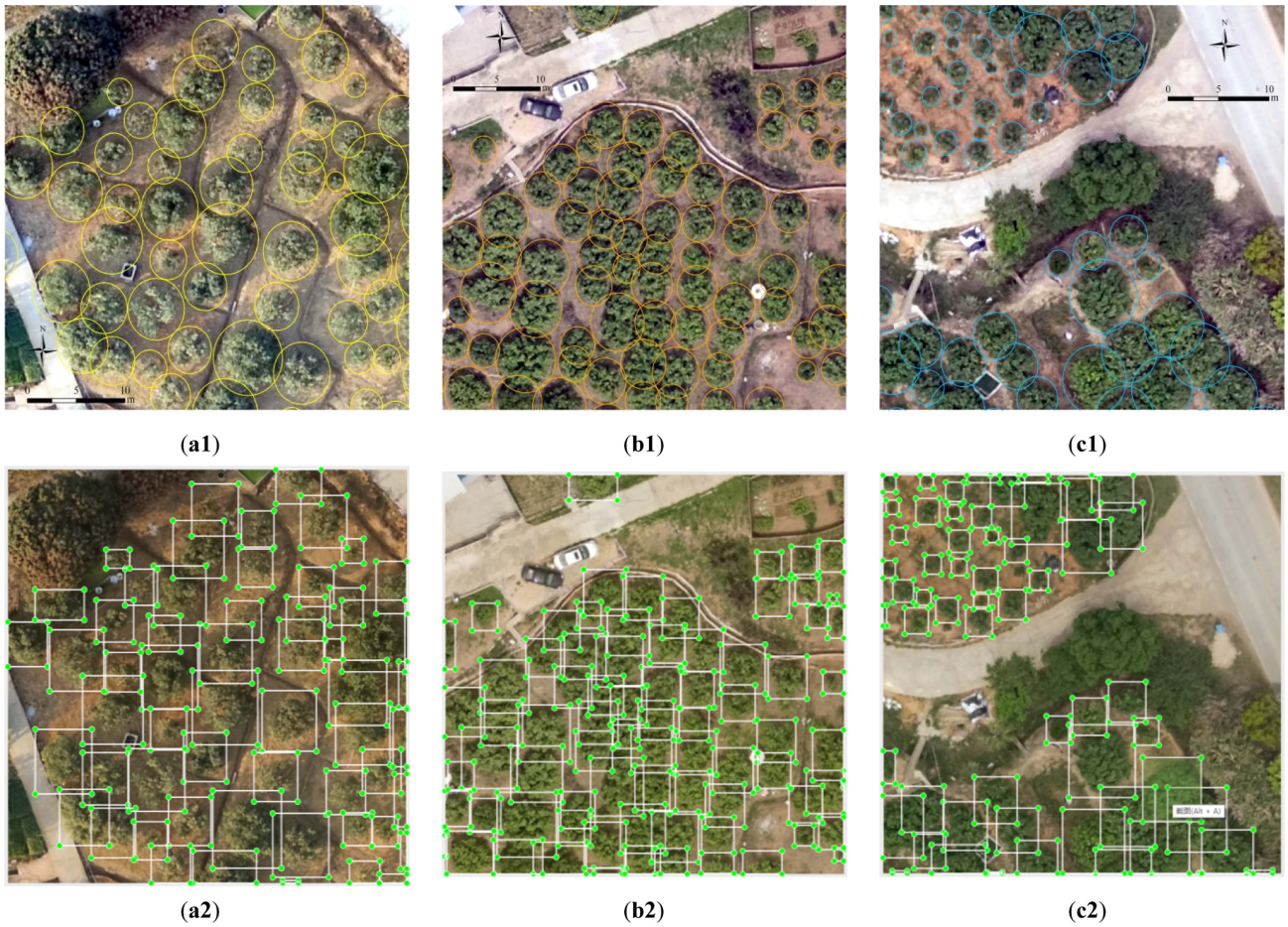

Fig. 5. Annotated samples (640 × 640 pixels). Images a1, b1, and c1 are clipped in ArcGIS, and trees are indicated by the labeled circles in yellow, brown, and cyan on the clipped images. Images a2, b2, and c2 were clipped in LabelImg, and the counterparts of the labeled trees are all shown in white. Some other trees and crops lie in the same clipped images, and IPTs of different sizes exist sparsely or densely, which negatively affects the detection of IPTs. (a1) 3 February 2021. (b1) 12 March 2021. (c1) 12 April 2021. (a2) 3 February 2021. (b2) 12 March 2021. (c2) 12 April 2021.

TABLE I
CONFIGURATION OF VARIOUS SCALES OF YOLOV5 MODELS

Model	Depth	Width	Number of layers	Number of parameters/M	FLOPs*/G
YOLOv5n	0.33	0.25	213	1.9	4.5
YOLOv5s	0.33	0.50	283	7.2	16.5
YOLOv5m	0.67	0.75	391	21.2	49.0
YOLOv5l	1.00	1.00	499	46.5	109.1
YOLOv5x	1.33	1.25	607	86.7	205.7

*FLOPs denote the floating-point operations, [51].

Based on the standard YOLOv5s model, seven optimized YOLOv5s models were created and trained to improve the performance of IPT detection by integrating attention modules. An attention mechanism is a powerful tool developed to enhance the performance of the encoder–decoder architecture on neural-network-based tasks, such as natural language processing and computer vision [53]. A great number of attention mechanisms have been proposed to improve the performance of deep learning models over the past decade. In the present study, seven state-of-the-art attention mechanism modules were inserted into the end of the neck network (i.e., the C3_1_F module in Fig. 6) before the Conv layer of the detection part of YOLOv5s, as shown in Fig. 7. Instead of replacing some C3 modules in the backbone, as in other researchers [54], our aim is to directly optimize the detection structure and enhance the channel and/or spatial feature extraction capacity and output performance. These attention modules include seven attention mechanisms, namely, coordinate attention (CA), split attention (SA), squeeze-and-excitation (SE), convolutional block attention module (CBAM), Ghost, simple attention mechanism (SimAM), and the transformer encoder, which are explained as follows.

a) Coordinate attention: The CA mechanism [see Fig. 8(a)] captures not only cross channel but also direction-aware and position-sensitive information, which helps models locate and recognize the objects of interest more accurately [55]. It can fuse both channel and spatial information to enhance the localization accuracy of object detection. Recently, the CA-based YOLOv5s model was used for mummy berry disease detection, which showed that the overall performance of the improved YOLOv5s-CA network

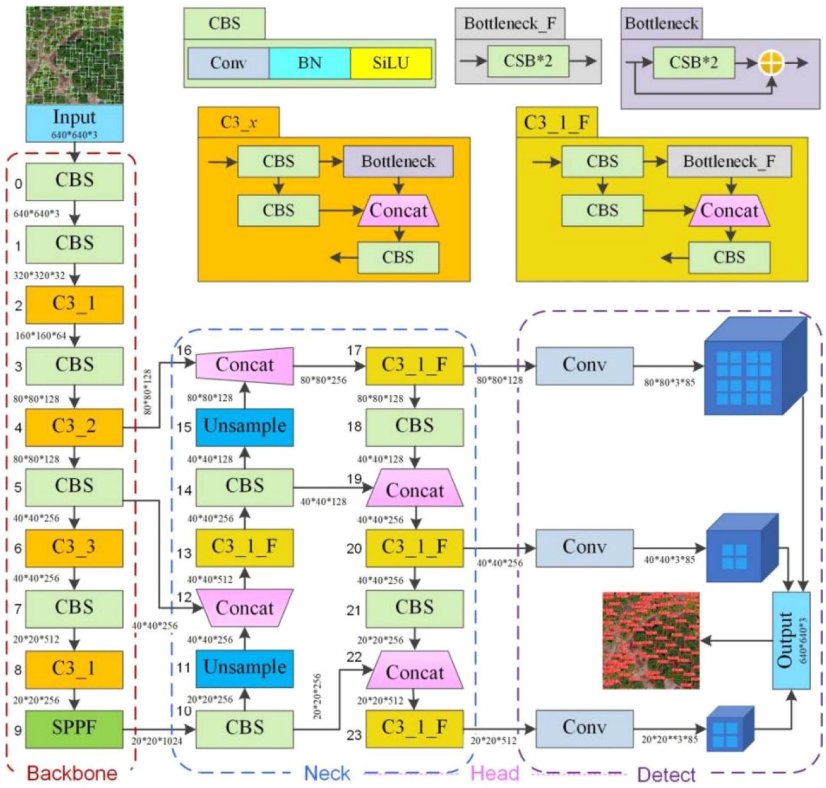

Fig. 6. Architecture of the standard YOLOv5s model.

Fig. 7. Attention-based optimization of YOLOv5s.

model was superior to that of the original YOLOv5s model [56].

b) Split attention: SA, as a simple and unified computation block [see Fig. 8(b)], applies channelwise attention to different network branches to leverage their success in capturing cross-feature interactions and learning diverse representations; ResNet outperforms EfficientNet in accuracy and latency tradeoff on image classification and has achieved superior transfer learning results on several public benchmarks serving as the backbone [57].

c) Squeeze-and-excitation: The SE module [see Fig. 8(c)] adaptively recalibrates channelwise feature responses by explicitly modeling interdependencies between channels, which significantly improves the performance of the existing state-of-the-art CNNs at a slight additional computational cost [58].

d) Convolutional block attention module: CBAM [see Fig. 8(d)], a simple, lightweight, yet effective attention module for feedforward CNNs, sequentially infers attention maps along two separate dimensions, channel and spatial. The attention maps are multiplied by the input feature map for adaptive feature refinement, which can be seamlessly integrated into any CNN architecture with negligible overhead, and CBAM is end-to-end trainable along with base CNNs [59]. The CBAM module has two sequential submodules: channel and spatial. The intermediate feature map is adaptively refined through CBAM at every convolutional block of the deep networks [see Fig. 8(d)].

e) YOLOv5s-ghost: The ghost module [see Fig. 8(e)], as a lightweight and plug-and-play component for existing CNNs, generates more feature maps that apply a series of linear transformations at low cost to generate many ghost feature maps that can fully reveal the information underlying intrinsic features based on a set of intrinsic feature maps. Ghost provides an impressive alternative to convolution layers in baseline models and has a higher recognition performance than MobileNetV3 [60].

f) Simple attention mechanism: SimAM is a conceptually simple, parameter free but highly effective attention module [see Fig. 8(f)] for CNNs that infer 3-D attention weights for the feature map in a layer without adding parameters to the original networks; moreover, SimAM is flexible and effective in improving the representation ability of many CNNs evaluated on various visual tasks [61].

g) Transformer encoder: The transformer is a new simple network architecture based on self-attention [see Fig. 8(g)] that dispenses recurrent information and convolutions entirely; this architecture performs very well on tasks, such as machine translation [53], ChatGPT [62], image generation [63], and classification [54]. Based on YOLOv5, Zhu et al. [54] replaced the original prediction heads with transformer prediction heads (TPHs), demonstrating that the improved TPH-YOLOv5 has

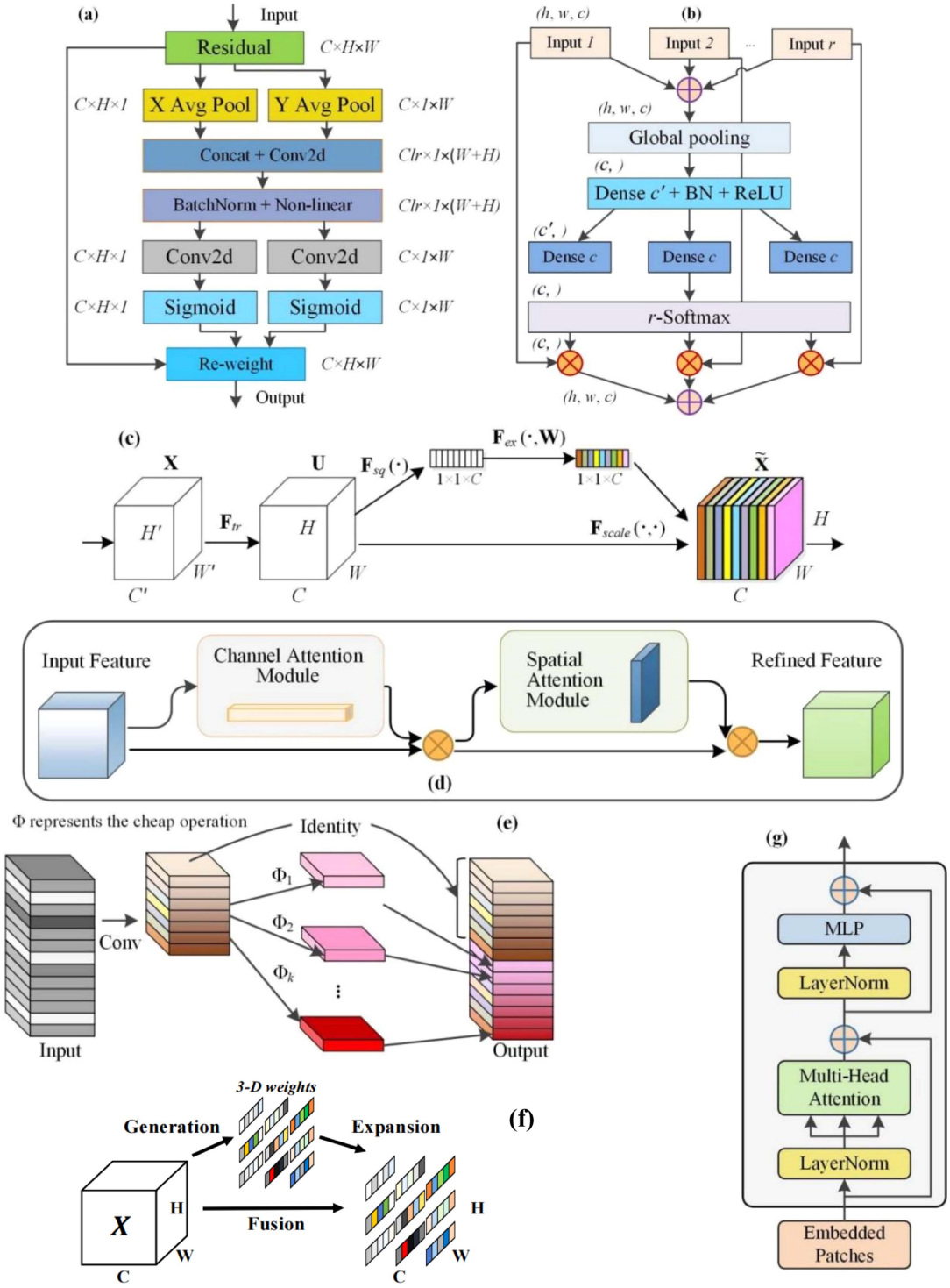

Fig. 8. Seven attention mechanism modules used for the optimization of YOLOv5s. (a) CA block [55], (b) SA block [57], (c) SE block [58], (d) CBAM [59], (e) ghost module for outputting the same number of feature maps [60], (f) SimAM [61], and (g) transformer encoder module [64].

good performance with impressive interpretability in scenarios with drone-captured data.

In this work, the attention mechanisms mentioned above were integrated into YOLOv5s to enhance its channel and spatial feature extraction ability and increase its performance in IPT detection, resulting in seven optimized YOLOv5s models, i.e., YOLOv5s-CA, YOLOv5s-SA, YOLOv5s-SE, YOLOv5s-CBAM, YOLOv5s-Ghost, YOLOv5s-SimAM, and YOLOv5s-Transformer. It is desirable to compare the accuracy of these methods and select the optimal method for IPT detection. Five steps were implemented to obtain the optimal model based on YOLOv5 as follows.

First, the five scales of the YOLOv5 models were trained and validated on the IPTIS dataset.

Second, the YOLOv5 model, which had the highest accuracy metric (YOLOv5s), was selected as the best choice for optimization.

Third, all the optimized YOLOv5 models were trained from scratch on the IPTIS dataset.

Fourth, the optimized YOLOv5s model, which we call FruitNet, had the highest accuracy metrics and was determined to be the preferred IPT detection model.

Fifth, the FruitNet model was additionally tested on images from other areas to confirm its robustness and generalizability.

4) Mapping Spatial Distributions and Counting the Number and Planting Area of IPTs: Due to the limited memory of the graphics card and the input image size of the model, a large-scale UAV high-resolution image of the whole area cannot be directly detected and predicted using the default YOLOv5 detection code. Etten [16] proposed the YOLT method based on YOLOv2 to perform object detection in large-scale remote sensing images. In accordance with the postprocessing idea of YOLT, the whole DOM images of the two study areas were cropped in memory by slicing and input to the model for prediction based on a specific image size (i.e., 640 × 640 pixels) and a given overlap degree (i.e., 50%). The cropped images were not stored on a hard drive but were stored in memory for rapid detection. The 50% overlap ensured that all regions were detected and avoided dividing an IPT at the edge of an image into more than one part. However, this overlap results in the necessary elimination of many redundant detection boxes. Nonmaximum suppression (NMS) is a postprocessing technique used in object detection to eliminate duplicate detections and select the most relevant detected objects. This helps reduce false positives and the computational complexity of a detection algorithm. Originally, the criteria used to arrive at the desired results were most commonly some form of probability number and some form of overlap measure (e.g., intersection over union). To better remove redundant detection boxes, we applied an improved NMS method to obtain the final prediction results of the bounding boxes. Because an IPT has a round crown, the width-to-height and height-to-width ratios (both set to 0.80) of a predicted bounding box were added to the eliminated conditions of redundantly predicted boxes in the original NMS algorithm. This approach helped delete the redundant prediction boxes of the split IPTs at the boundaries of the sliced images as much as possible and, thus, increased the accuracy of the IPT detection. The final bounding boxes were stored in a shapefile with actual geographic coordinates stemming from the original DOM image, which made it easy to overlay the DOM image.

In general, the following four steps were adopted to construct spatial distribution maps of IPTs.

First, the optimized YOLOv5 model (FruitNet) was used to detect IPTs and obtain the coordinates of the detected trees in a whole DOM image by slicing the image.

Second, based on the overlap degree of 50% of neighboring slices, the repeatedly detected boxes in the overlapping slice area were deleted by using our improved NMS algorithm for bounding boxes.

TABLE II
CONFUSION MATRIX FOR IPT DETECTION

Confusion matrix		Predicted label	
		true	**false**
Actual label	positive	TP*	FP
	negative	TN	FN

Third, the two shapefiles of the final detection results of the IPTs and the original mosaic DOM images were imported into ArcGIS to map the IPTs. The numbers of detected IPTs in the study areas were recorded from the attribute tables of the shapefiles. The planting areas of the detected IPTs in the study areas were summed with the attribute tables of the shapefile layers in ArcGIS.

Finally, two thematic maps of the detection results of the IPTs and their spatial distributions were constructed using ArcGIS software. The numbers and planting areas of the detected IPTs were added to the thematic maps as important statistical annotation information.

C. Evaluation Metrics

In deep learning, it is highly important to evaluate models [65]. The present study used a confusion matrix (see Table II) to help evaluate the performance of the selected object detection models. Table II presents the confusion matrix for IPT detection, where each column represents the predicted value and each row represents the actual pomelo category. The precision, recall, F_1 score, and average precision metrics were used as evaluation metrics based on the confusion matrix of each model.

1) Precision and Recall: According to Table II, the precision (P) and recall (R) metrics are defined in (1) and (2), respectively. The precision denotes the proportion of actual positive IPT samples among all the results predicted as positive IPT samples. The recall, also known as the sensitivity, denotes the proportion of the IPT samples predicted as positive IPT examples by the classifier to the actual number of positive IPT examples and describes the classifier's sensitivity to the category of positive IPT examples

$$P = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP}) \tag{1}$$

$$R = \mathrm{TP}/(\mathrm{TP} + \mathrm{FN}) \tag{2}$$

where P and R denote the precision and recall, respectively; TP, FP, and FN have the same meanings as in Table II.

2) F_1 Score: The F_1 score is the harmonic mean of precision and recall, and is calculated as follows:

$$F_1 = 2 \times P \times R/(P + R) \tag{3}$$

where F_1 denotes the F_1 score and P and R denote the precision and recall, respectively.

3) Average Precision: The average precision (AP) was calculated from the precision–recall (PR) curve drawn after training. The area below the PR curve of a specific category is referred

Fig. 9. Intersection over union of the ground truth and prediction bounding box.

to as the AP of that category, as defined in the following equation:

$$\mathrm{AP} = \int_0^1 f(c)\, d(c) \tag{4}$$

where AP represents the average precision and $f(c)$ represents the precision–recall curve of category c. In the PR plot, the closer the curve is to the upper right corner, the higher the model's accuracy.

4) Intersection Over Union: The ratio of intersection over union (IoU) is used to evaluate the degree of matching of the predicted IPT bounding box and the ground truth IPT box (see Fig. 9), calculated as follows:

$$\mathrm{IoU} = \frac{\mathrm{Area}(P) \cap \mathrm{Area}(G)}{\mathrm{Area}(P) \cup \mathrm{Area}(G)} \tag{5}$$

where Area(P) represents the area of the predicted IPT box, and Area(G) represents the area of the ground truth IPT bounding box. The higher the ratio is, the better the degree of matching. The predicted box and ground truth box overlap completely in an ideal result, and the IoU is 1 in this case.

The IoU is generally negatively correlated with the AP of a model. The present study's threshold for positive cases was IoU > 0.5; otherwise, the case was negative. The AP@*0.5* used below denotes the average precision when the IoU > 0.5, and the AP@*0.5:0.95* used below represents the average precision when the IoU lies between 0.5 and 0.95.

In addition, the inference time is an important indicator of a model's ability to detect objects. Frames per second (FPS) were used to measure the model's inference speed in the study.

D. Implementation Details

1) Hardware and Software Environment: A high-performance computing environment is necessary to conduct deep learning object detection tasks. A suitable experimental environment was set up to fulfill the task of the present study and was configured as follows.

1) CPU: Intel (R) Core (TM) i7-7700K @ 4.20 GHz/i3-10100F CPU @ 3.60 GHz (virtual machine).
2) Memory: Kingston DDR4 3200 MHz 2×16 GB.
3) GPU: NVIDIA GeForce RTX 2080 Ti, 11 GB.
4) OS: Windows 10/Linux-5.4.0 Debian 64-bit.
5) Language: Python-3.7.11+torch-1.9.0+cu111 CUDA.

TABLE III
CONFIGURATION OF THE IPT DATASET FOR CROSS VALIDATION AND TESTING

Area	Date	Item	Training	Validation	Testing	Total
Shishan (Site A)	February 2021 (D_F)	Clip image	117	15	15	147
		Label	4 923	716	732	6 371
	March 2021 (D_M)	Clip image	100	13	13	126
		Label	5 631	598	538	6 767
	April 2021 (D_A)	Clip image	131	17	17	165
		Label	6 543	756	648	7 947
Sanxiang (Site B)	January 2022 (D_J)	Clip image	36	3	3	42
		Label	797	59	66	922
Total		Clip image	384	48	48	480
		Label	17 894	2 129	1 984	22 007

TABLE IV
CONFIGURATION OF THE IPT DATASET FOR CROSS VALIDATION AND TESTING

Hyperparameter	Value	Hyperparameter	Value
optimizer	SGDM *	Hsv_h (hue)	0.015
max epoch	200	hsv_s (saturation)	0.7
lr0 (initial learning rate)	0.01	hsv_v (lightness)	0.4
momentum	0.937	translate	0.1
weight_decay	0.0005	scale	0.5
minimum batch size	32 (YOLOv5n), 16 (YOLOv5s), 12 (YOLOv5m), 8 (YOLOv5l), 4 (YOLOv5x)	fliplr (flip left–right)	0.5
		mosaic	1.0

* SGDM refers to stochastic gradient descent with momentum.

6) IDE: PyCharm Community 2022.1.
7) Source code: https://github.com/iscyy/yoloair.

2) Dataset Division and Data Augmentation: For cross validation and testing, 384 clip images were randomly selected as the training set (80%) and 48 clip images were used as the validation set (10%). The remaining 48 clip images were used as the testing set (10%) to test the robustness of the final model (see Table III). For comparison, four datasets of different dates with different numbers of samples (see Table III) were used to train, validate, and test the performance of the optimized YOLOv5s models.

The UAV images of the same area from the three periods were combined to create the dataset, which can be considered data augmentation. No additional offline data augmentation was applied, but the online data augmentation was applied to ensure the stability and robustness of the model training and validation. The online methods include the following: illumination distortions of hue, saturation, and lightness; translation; geometric distortions of left–right flipping, scaling, and rotation; and mosaicking of four clip images (see Table IV).

3) Hyperparameter Setup: All the models were trained from scratch with no pretrained weights on the IPT dataset to obtain the IPT detection models, which were subsequently validated on the validation set. The default hyperparameters were used

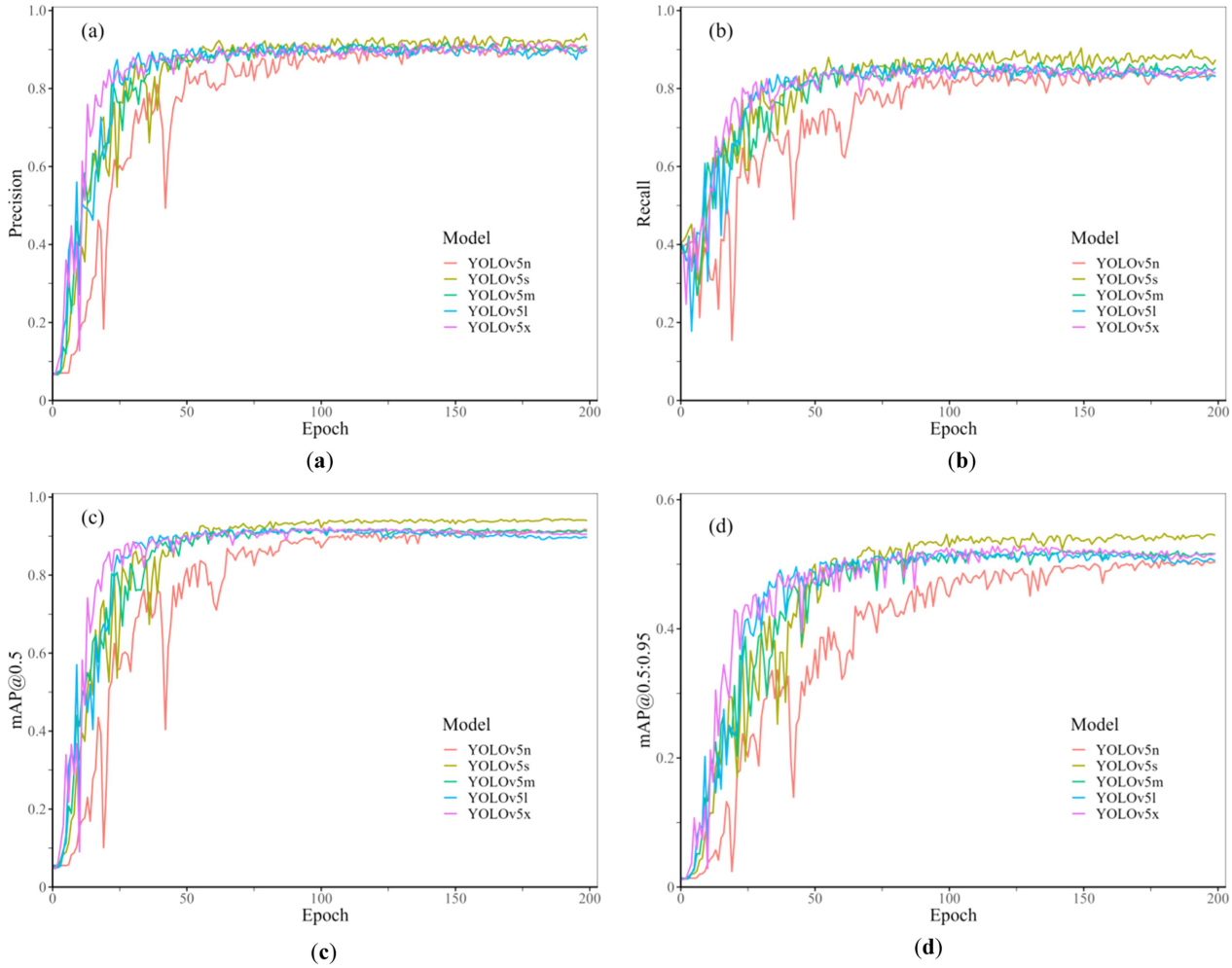

Fig. 10. Visualization of the evaluation metrics of the five YOLOv5 models for training and validation. (a) Training precision. (b) Training recall. (c) Training AP@*0.5*. (d) Training AP@*0.5:0.95*.

to train from scratch with the hyp.scratch-low.yaml file and are mostly shown in Table IV.

To take the full advantage of the memory utilization of the GPU and obtain good accuracy, the maximum number of epochs was set to 200, and the minimum batch size was set to different numbers depending on the needs of the models and the capacity of the GPU. The learning rate was initially set to 0.01 and it decayed dynamically to 0.0002 at the end of the 200th epoch for the standard YOLOv5 models. All the models were trained using a stochastic gradient descent with momentum (SGDM) strategy to avoid overfitting and underfitting.

III. Results

A. Performance of the Five Standard YOLOv5 Models

First, five standard YOLOv5 models were trained and validated on the IPTIS dataset to compare their accuracy metrics and performance. The results are shown as follows.

1) Accuracy Metrics of Training and Validation: As shown in Fig. 10, the precision [see Fig. 10(a)], recall [see Fig. 10(b)], and AP [see Fig. 10(c) and (d)] of the five YOLOv5 models essentially stabilize after 150 epochs; the metric values of YOLOv5s are higher than those of the others. This characteristic was confirmed by the PR curve [see Fig. 11(a)], which demonstrated that the comprehensive performance of YOLOv5s was better than that of all the other models selected. This finding suggested that YOLOv5s has the best accuracy and is the best option for optimizing IPT detection.

Table V and Fig. 11(b) present that the precision, recall, F_1 score, and AP metrics of YOLOv5s are the highest among those of the five standard YOLOv5 models evaluated on the IPTIS validation set, followed by YOLOv5x; the metrics of YOLOv5n are the lowest. Interestingly, the F_1 score and AP metrics do not increase with the scale of YOLOv5, which is not in line with the results of the YOLO series tested on the COCO dataset [51]. The reason for this could be that the size of our dataset is much smaller than that of the COCO dataset. This issue will be discussed in detail in Section IV. The accuracy metrics of YOLOv5s are higher than those of the other four YOLOv5 models [see Table V and Fig. 10(b)], further suggesting that

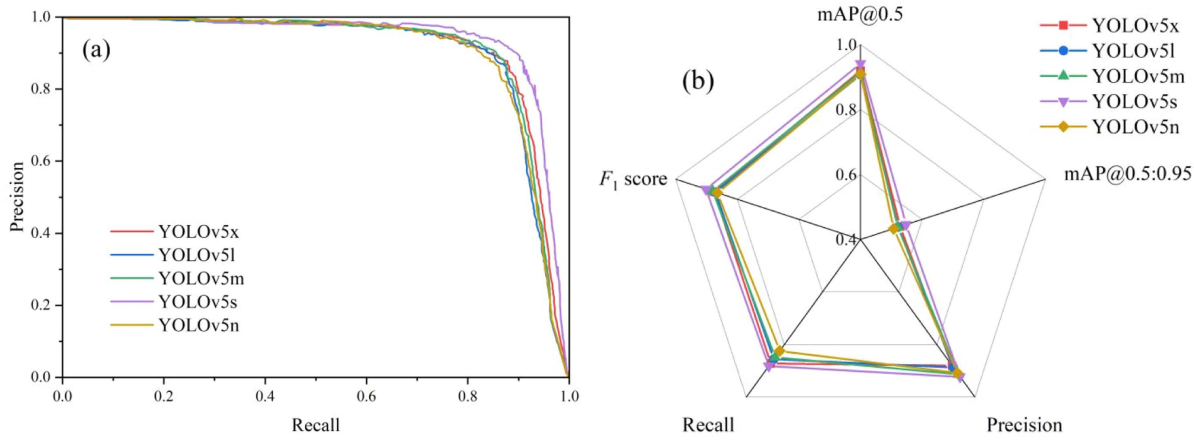

Fig. 11. Comparison of the training and validation results of the five standard YOLOv5 models. (a) PR curve for training results. (b) Evacuation metrics for model validation.

TABLE V
RESULTS OF THE ACCURACY METRICS OF THE FIVE STANDARD YOLOV5 MODELS FOR VALIDATION ON IPTIS

Model	Precision	Recall	F_1 score	*AP*@0.5*	*AP*@0.5:0.95
YOLOv5n	0.9080	0.8243	0.8641	0.9093	0.5075
YOLOv5s	**0.9224**	**0.8817**	**0.9016**	**0.9412**	**0.5485**
YOLOv5m	0.9123	0.8478	0.8789	0.9142	0.5210
YOLOv5l	0.8865	0.8548	0.8704	0.9068	0.5220
YOLOv5x	0.8893	0.8630	0.8759	0.9195	0.5281

* *AP*@0.5 means the average precision when the *IoU* > 0.5 and *AP*@0.5:0.95 means the average precision when the *IoU* lies between 0.5 and 0.95.
The bold entity indicates the maximum value (best result) for each accuracy metric.

TABLE VI
TRAINING TIME AND INFERENCE SPEED OF VARIOUS SCALES OF THE FIVE YOLOV5 MODELS

Model	Training time (h)	Inference speed (FPS*)
YOLOv5n	1.168	73
YOLOv5s	1.230	52
YOLOv5m	1.391	58
YOLOv5l	1.593	47
YOLOv5x	2.181	34

* FPS denotes the frames per second.

YOLOv5s is the most accurate model and is suitable for further optimization to fulfill IPT detection.

2) Training Time and Inference Speeds: As shown in Table VI, as the scale of YOLOv5 increases, the training time increases, and the inference speed decreases, except for YOLOv5m, which is faster than YOLOv5s. All the models are capable of real-time inference (i.e., FPS is greater than 30). Although YOLOv5n consumes fewer computational resources than the other YOLOv5 models, its accuracy metrics are the worst. Considering both performance and consumption, YOLOv5s is the best choice for performing real-time smart orchard management inference tasks. YOLOv5s has the highest evaluation metrics, notably both AP@*0.5* (0.9412, 94.12%) and F_1 score (0.9016, 90.16%), among all the scales of YOLOv5. Therefore, this model was selected as the model for optimization to detect IPTs.

B. Comparison of the Optimized YOLOv5 Models

As mentioned above, seven attention-based YOLOv5 models were trained and validated on the IPTIS dataset. For comparison, faster R-CNN, SDD, YOLOv3, YOLOv4s, YOLOv5s, YOLOv7-tiny, YOLOv7, and YOLOv8s were additionally trained and validated on the IPTIS dataset together with the optimized models mentioned above. The results are shown as follows.

1) Accuracy Metrics of Training and Validation: As shown in Fig. 12, the precision [see Fig. 12(a)], recall [see Fig. 12(b)], and AP [see Fig. 12(c) and (d)] metrics of the trained models remain almost steady after the 100th epoch, and the metric values of YOLOv5s and its optimized models are greater than those of the other models. In the end, the F_1 score and AP metrics (see Fig. 13) of YOLOv5s and its optimized models for validation are all greater than those of the other state-of-the-art models, such as faster R-CNN, SSD, YOLOv7, and YOLOv8s, which illustrates that the comprehensive performance of YOLOv5s and its optimized models surpasses that of all the other models used.

However, from Figs. 12 and 13, it is difficult to determine the best model according to these five metrics because no optimized YOLOv5s model has an advantage over the others in terms of all the commonly used accuracy metrics in the present study. To compare the overall accuracy of these models, a comprehensive metric that averages the F_1 score, AP@*0.5*, and AP@*0.5:0.95* (OA for short) was proposed to help determine the best model. The OA for validation was calculated for each model used. The results showed that the optimized YOLOv5s-CA had the highest OA (0.801), resulting from the highest F_1 score (0.906, 90.6%) and AP@*0.5:0.95* (0.554, 55.4%) and the third highest AP@*0.5* (0.943, 94.3%) among all seven optimized YOLOv5s models.

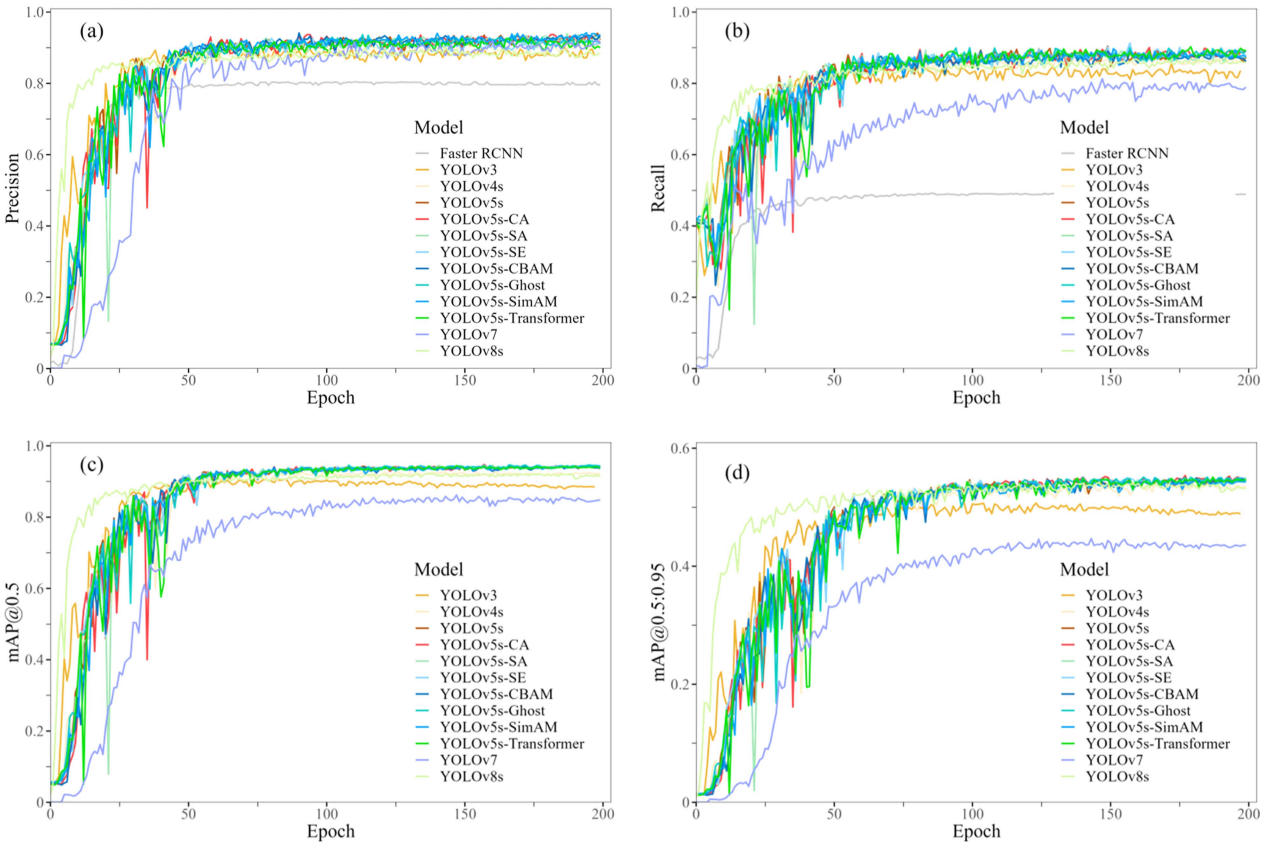

Fig. 12. Comparison of the training evaluation metrics between the optimized YOLOv5 model and the other selected object detection models. (a) Training precision. (b) Training recall. (c) Training AP@*0.5*. (d) Training AP@*0.5:0.95*.

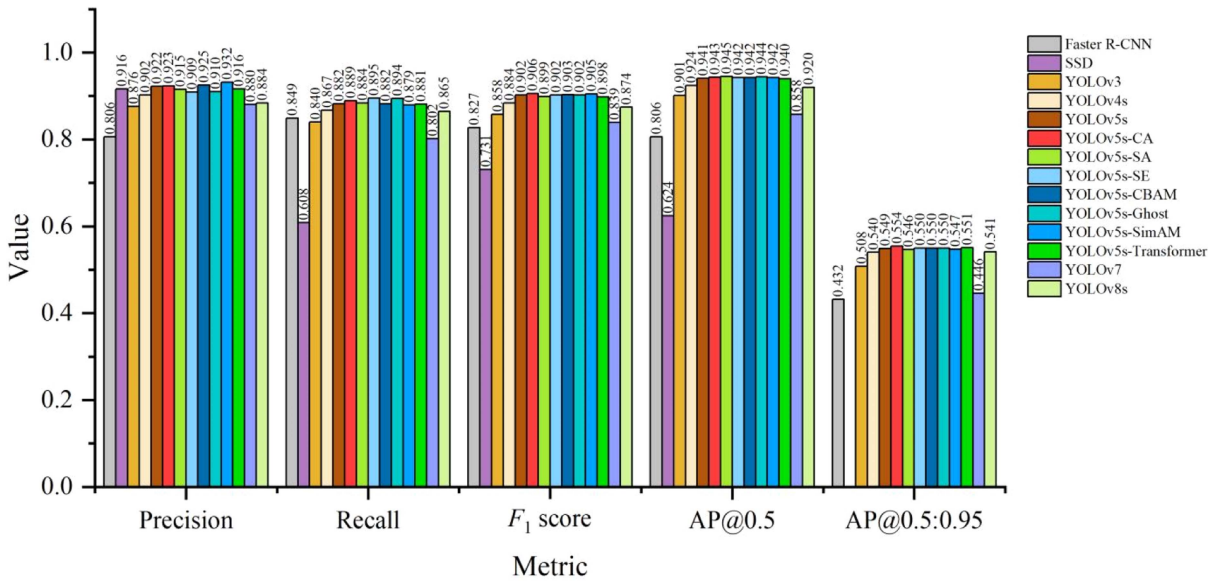

Fig. 13. Comparison of the accuracy evaluation metrics among the optimized YOLOv5s and the state-of-the-art object detection models selected for validation.

TABLE VII
TRAINING TIME, INFERENCE SPEED, AND COMPLEXITY OF THE OPTIMIZED YOLOV5S AND OTHER OBJECT DETECTION MODELS USED

Model	Training time/h	Inference speed/FPS*	Number of layers	Number of parameters/M	FLOPs* /G
Faster R-CNN	1.479	21	153	180.0	46.7
YOLOv3	1.774	41	333	61.5	155.3
YOLOv4s	1.337	66	330	9.1	20.8
YOLOv5s	1.230	52	270	7.0	15.9
YOLOv5s-CA	1.231	80	278	7.0	16.0
YOLOv5s-SA	1.195	61	274	7.0	15.9
YOLOv5s-SE	1.233	74	277	7.0	16.0
YOLOv5s-CBAM	1.239	81	282	7.0	16.0
YOLOv5s-Ghost	1.219	71	306	7.6	16.4
YOLOv5s-SimAM	1.231	74	272	7.0	15.9
YOLOv5s-Transformer	1.230	64	194	9.2	16.7
YOLOv7	1.697	46	415	37.2	105.1
YOLOv8s	1.347	104	255	11.2	28.6

*FPS denotes the frames per second, and FLOPs denote the floating-point operations. The YOLOv5s model we used here is not the original YOLOv5s listed in Table I, but a new modified version from the YOLO air package.[4]

YOLOv5s-CA is, thus, the best model for IPT detection in terms of overall accuracy.

2) Training Time and Inference Speeds: As shown in Table VII, the training times of all the selected models do not differ greatly except for those of faster R-CNN, YOLOv3, and YOLOv7, whose inference speeds are slower than those of any optimized YOLOv5s model. All the models except faster R-CNN are capable of real-time inference. YOLOv8s has the fastest inference speed, although it has more parameters and FLOPs than YOLOv5s and its optimized versions. According to the inference speeds, the seven attention-optimized models are all faster than the standard YOLOv5s. YOLOv5s-CA is one of the optimized models with the fastest inference speed. Based on its performance and speed, YOLOv5s-CA is the best option for performing real-time IPT inference tasks for smart orchards.

C. Mapping the IPTs

Based on its overall accuracy and inference speed, YOLOv5s-CA, henceforth referred to as FruitNet, was selected as the model for predicting IPTs at the two study sites.

Two thematic maps (see Figs. 14 and 15) were made to show the spatial distribution, number, and planting area of the detected IPTs via ArcGIS based on the predictions of FruitNet. As shown in the regions of the two square boxes inserted [see Fig. 14(b) and (c)] in Fig. 14, IPTs of different sizes at site A were almost always detected accurately. This finding suggested that FruitNet has a sufficiently high accuracy to complete the IPT detection task in a large orchard area. The background of site B was more complex than that of site A, with more false influences from other trees than from site A. Most of the IPTs at site B were accurately predicted, but some other trees were falsely detected as IPTs (see Fig. 15). These results verified our hypothesis. With the aid of deep learning, the location of every IPT can be accurately inferred. Based on the mosaic DOM tessellated with the UAV images, the distribution map of all the IPTs in a large orchard can be generated by using ArcGIS. To our knowledge, the integration of deep learning with UAV remote sensing is a groundbreaking way to detect and make thematic maps of the spatial distributions of IPTs.

The planting area and the number of detected IPTs in the experimental study area were counted with ArcGIS software and are shown in the thematic map within an inserted table. Notably, the total planting area was obtained by summing the area of each IPT and omitting land with no detected pomelo trees. The total number of IPTs and planting area of site A were 7183 and 10.70 ha, respectively (see Fig. 14), and those of site B are 2817 and 5.52 ha, respectively (see Fig. 15). Although the land area of site B is greater than that of site A, the planting area and number of IPTs are smaller than those of site A because most of the land lies in the hilly regions densely covered by other trees in site B. The number of manually labeled IPTs is 6830 for site A, whereas that predicted by FruitNet is 7182. The percentage error of the number of IPTs is 5.15% for the prediction of the whole DOM at site A. The error for site B cannot be obtained due to the lack of manual labels for the entire area. The clipped images from site B were partly labeled to supplement the IPTIS dataset and used for testing the robustness of FruitNet.

These maps are two experimental cases demonstrating the application of FruitNet to large planting areas of pomelo trees in Meizhou city. Our aim was to promote the use of the FruitNet in the whole city of Meizhou by making an entire distribution map of IPTs in the city for smart orchard planning and management.

IV. DISCUSSION

A. Robustness of the Optimized Models

The seven optimized models were tested on the testing subset of IPTIS to validate their robustness and generalizability compared with the other models selected. As shown in Fig. 16, the five accuracy metrics of the seven optimized models for testing all reach relatively high values, which are all greater than those of faster R-CNN, YOLOv3, YOLOv4s, and YOLOv7, except for the recall metric of YOLOv4s. The precision metrics are all greater than 0.870, the recall metrics are greater than 0.844, the F_1 score metrics are greater than 0.863, the AP@*0.5* metrics are greater than 0.844, and the AP@*0.5:0.95* metrics are greater than 0.550. All these methods have relatively high robustness in inferring IPTs. However, comparing Fig. 16 with Fig. 13, we can see that these metrics are all less than those of the validation set. Hence, the robustness and generalizability of the optimized models still need to be improved to increase the precision of the IPT inference in other orchards.

Interestingly, the F_1 score, AP@*0.5*, and AP@*0.5:0.95* metrics for the YOLOv5s-transformer model are the highest among all the optimized models, implying that the YOLOv5s-transformer model has great potential for performing the IPT

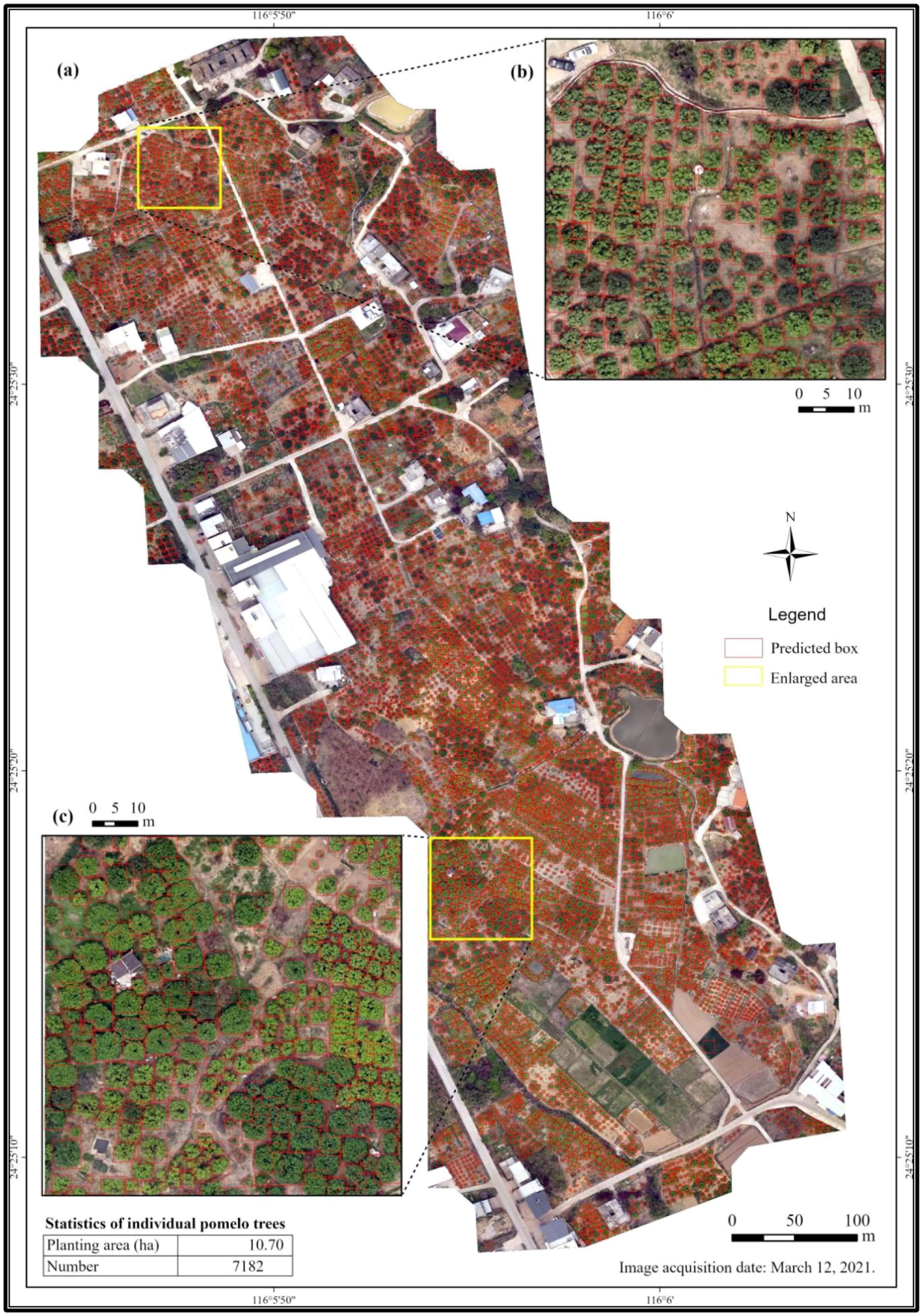

Fig. 14. Thematic map showing the spatial distribution, planting area, and number of IPTs detected using FruitNet and the (a) mosaic image of site A acquired on 12 March 2021, with two enlarged rectangular regions inserted as (b) and (c).

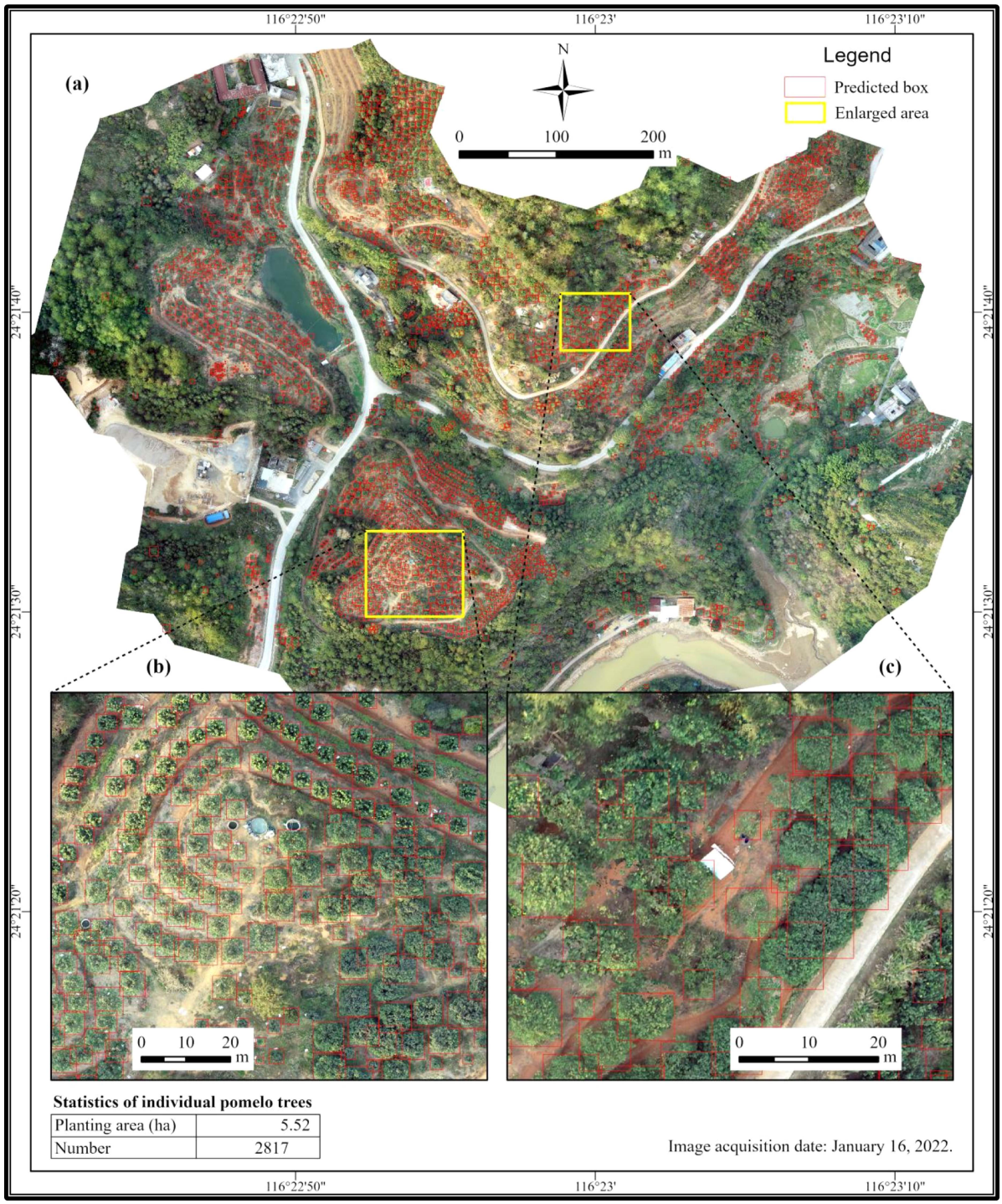

Fig. 15. Thematic map showing the spatial distribution, planting area, and number of IPTs detected using FruitNet and a (a) mosaic image of site B acquired on 16 January 2022, with two enlarged rectangular regions inserted as (b) and (c).

inference task. However, these values are slightly greater (0.1%–0.6%) than those of the other six attention-optimized YOLOv5 models, especially for YOLOv5s-CA (0.1%–0.2%).

The robustness and feasibility of FruitNet were further tested on the D_J dataset of site B in Sanxiang. For brevity, only the evaluation metrics of the standard and optimized YOLOv5 models for testing on the D_J dataset were compared. Table VIII presents that all five metrics except the recall of YOLOv5s-CA are greater than those of YOLOv5s. The optimized YOLOv5s-CA model is more robust than the original YOLOv5s model and is more feasible for inferring IPTs in other orchards.

To observe the IPT inference results visually, two clipped images [see Fig. 17(a) and (b)] from the D_J dataset of the pomelo orchards at site B in Sanxiang together with one clipped

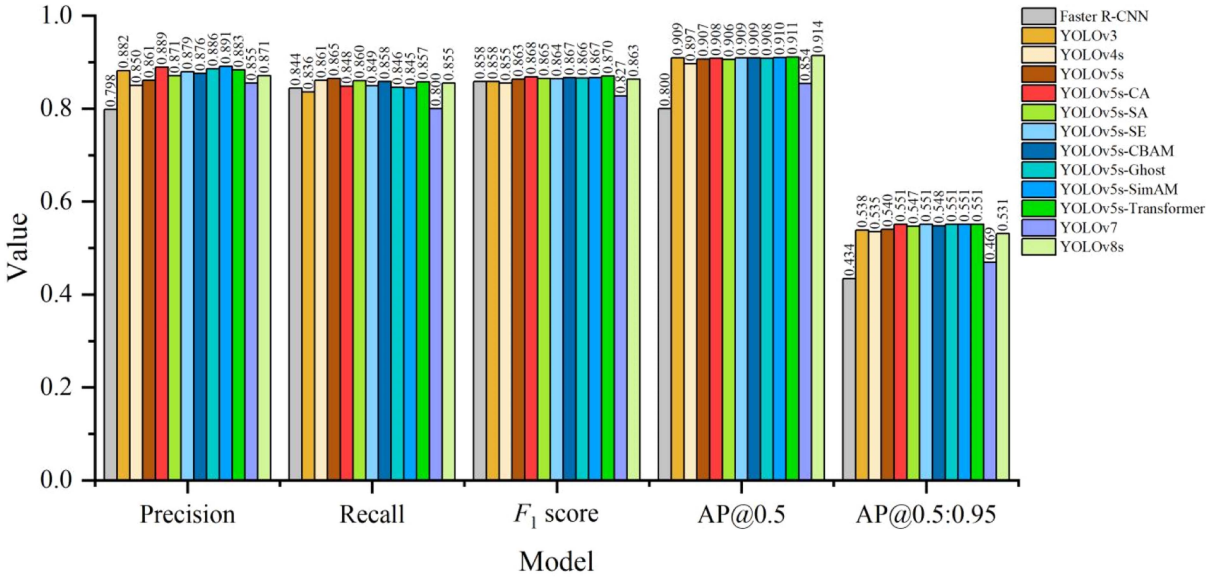

Fig. 16. Comparison of the accuracy evaluation metrics among the optimized YOLOv5s and the state-of-the-art object detection models selected for testing. AP@*0.5* denotes the average precision when the IoU $>$ 0.5, and AP@*0.5:0.95* denotes the average precision when the IoU lies between 0.5 and 0.95.

TABLE VIII
COMPARISON OF THE ACCURACY METRICS BETWEEN THE STANDARD AND OPTIMIZED YOLOV5S-CA MODELS ON THE D_J DATASET

Model	Precision	Recall	F_1 score	*AP*@0.5*	*AP*@0.5:0.95*
YOLOv5s	0.785	0.820	0.863	0.851	0.548
YOLOv5s-CA	0.844	0.765	0.868	0.855	0.563

* *AP*@0.5 denotes the average precision when the *IoU* $>$ 0.5, and the *AP*@0.5:0.95 denotes the average precision when the *IoU* lies between 0.5 and 0.95.

image [see Fig. 17(c)] from the other orchard near site A in Shishan were collected to validate the robustness and generalizability of FruitNet. As shown in Fig. 17, the IPTs in the three images captured under different lighting conditions are almost all correctly predicted by both FruitNet and YOLOv5s, although a few IFTs are repeatedly overlaid [see Fig. 17(a1), (a2), (a3), (b2), and (b3)], falsely predicted [see Fig. 17(b1)], or fully omitted [see Fig. 17(b2)]. Furthermore, FruitNet and YOLOv5s can detect sparsely and densely distributed IPTs of different sizes with relatively high confidence (see Fig. 17). Background conditions, such as those of other kinds of similar trees or shrubs, may slightly affect the detection results. It is challenging for a model to detect IPTs with uneven and dense distributions from UAV images captured in complex environments. Most of the confidence scores of FruitNet are greater than those of YOLOv5s. Therefore, it is necessary to append more high-quality images of different complex environments across various pomelo orchards to the IPTIS dataset to establish a more reliable basis and make more in-depth improvements to the deep learning model optimization to increase its performance and robustness for the detection of IPTs in larger regions and even the entire Meixian city.

B. Evaluation of the Datasets of Different Dates Acquired

In addition to the IPTIS dataset, based on the composite of the UAV-based images taken during the four months at the two study sites (A and B) mentioned above (see Table III), four separate datasets were created using the UAV-based images of each month. The sizes of these four datasets are all very small, each containing no more than 165 clip images. The standard YOLOv5s and optimized YOLOv5s-CA models were trained and validated on the three datasets at site A. The results (see Fig. 18) show that the accuracy metrics almost all display increasing trends for the two models. The two exceptions are the precision metric of YOLOv5s, which decreases on the validation dataset from February 2021 (D_F) to March 2021 (D_M) and then increases on the validation dataset of April 2021 (D_A), and the recall metric of YOLOv5s, which decreases slightly on the validation datasets from D_M to D_A. This could be caused by the image quality and number of datasets. The images in the D_F dataset were captured in cold winter weather. Many plants, including pomelo trees, were damaged by cold weather in Meizhou in January 2021. Many pomelo trees dried up and did not recover, causing the image's color to appear yellow [see Fig. 4(a)]. This approach reduced the richness of image feature extraction and representation in the model. From February to April 2021, the pomelo trees at site A grew better, and their features were, thus, easier to recognize. Therefore, although the size of the D_M is smaller than that of the D_F, the F_1 score and AP metrics of both YOLOv5s and YOLOv5s-CA validated on the D_M are all greater than those validated on the D_F. Moreover, these metrics of the two models validated on the D_A are greater than those validated on both D_M and D_F.

Surprisingly, there was no improvement in performance when YOLOv5s-CA was compared with YOLOv5s when validated on any of the datasets, and the metrics showed a slight decrease

Fig. 17. Clip images overlaid with the ground truth boxes (a0, b0, and c0) and the detected IPT boxes and confidence scores predicted by YOLOv5s (a1, b1, and c1) and YOLOv5s-CA (a2, b2, and c2), respectively. (a) IPTs in sparse alignments. (b) IPTs in dense alignments interspersed with other trees. (c) IPTs in a dense and random distribution. Images (a) and (b) were selected from site B in the town of Sanxiang, and image (c) was deliberately selected from the other orchard near site A in the town of Shishan, Meizhou. These three images were not fed into the models for training. Note: White rectangles denote the manually labeled ground truth boxes, red rectangles denote the predicted bounding box with the class *P* and confidence score shown above it, brown rectangles denote the undetected IPTs, yellow ellipses denote the repeatedly overlaid IPTs, and cyan triangles denote the falsely predicted IPTs.

on all the datasets except for a small increase in the precision validated on the D_F (see Fig. 18). This finding suggested that the optimization of YOLOv5s could not improve the accuracy of the validated IPT detection on a small dataset. Two methods could be used to improve the performance of the model. One is to enlarge the size (number) of the dataset, which can enhance the feature representation of IPTs. The other is to try other network models or optimize the model, which can strengthen the feature extraction and generalization ability of the model used.

C. Comparison With Other Related Work

As mentioned in Section I, in our previous study, we proposed a YOLOx-nano pomelo tree detection method based on an

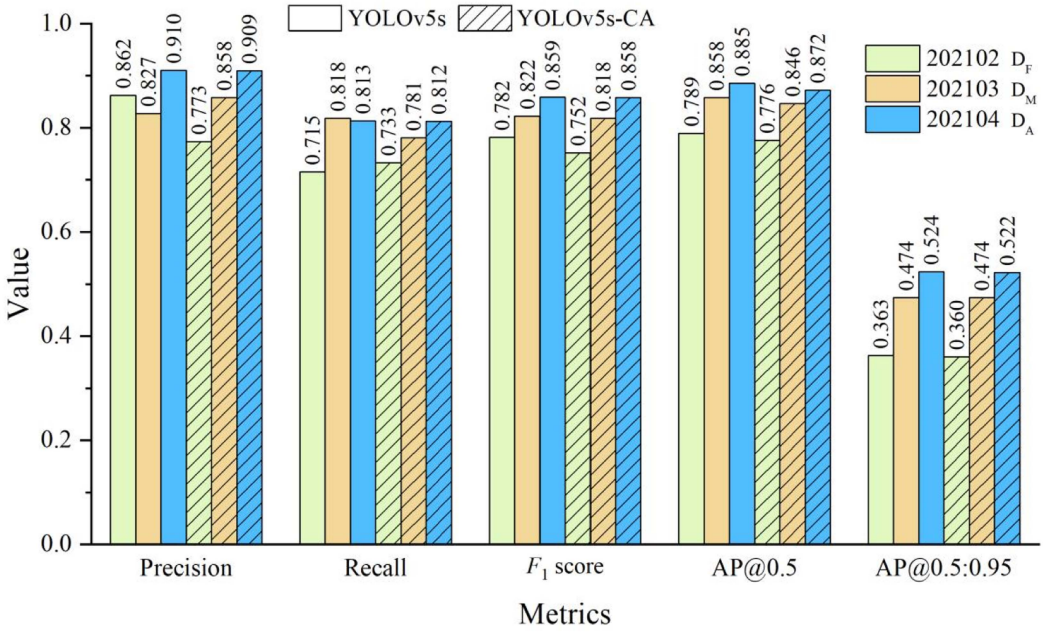

Fig. 18. Comparison of the accuracy metrics of the standard YOLOv5s and the optimized YOLOv5s-CA models for validation on the datasets of different dates.

attention mechanism and cross-layer feature fusion and showed that this method was more suitable for pomelo tree detection than other state-of-the-art object detection algorithms. The present study showed that YOLOv5s and its attention-optimized models can detect IPTs with high accuracy, in line with the results of Yuan et al. [49]. Although the structure of the network proposed by Yuan et al. [49] was lightweight, the AP value reached 93.74%. Our optimized YOLOv5s models fully outperformed this network in terms of AP, which were all higher than 94.00%, with the highest AP value of 94.50%. These two studies applied attention mechanisms to optimize the models used, increasing the ability of these models to identify small targets [49]. This finding implies that improving the deep learning model could increase the detection accuracy of IPTs.

Several other works detected citrus trees, which have crowns similar to those of pomelo trees, using the connected-components labeling algorithm [66], mask R-CNN [48], and YOLOv5s [67] based on the high-resolution UAV images. These methods (see Table IX) all demonstrated high accuracy, indicating the feasibility of using different algorithms for IFT detection tasks. The authors in [66] and [67] used citrus tree images without interference from other trees for training and validation, which are relatively simple object detection tasks. Our proposed model, FruitNet, was trained and validated on the IPTIS dataset, which consists of many images in which pomelo trees coexist with other trees, crops, roads, ponds, villages (houses), and overhead power lines. This complex natural background may have increased the difficulty of IPT detection.

Nevertheless, our model also has high-accuracy metrics, even greater than those of [49] and [67]. Therefore, our optimized YOLOv5s model has strong performance and is suitable for IPT detection. Furthermore, more deep learning algorithms could be used and optimized to detect IPTs or other IFTs for better performance and robustness. Integrating UAV remote sensing and deep learning could help accurately and efficiently detect and map IFTs in every orchard on the Earth for precision and smart orchard management.

TABLE IX
COMPARISON OF THE ACCURACY EVALUATION METRICS OF FRUITNET WITH OTHER RELATED WORKS

Model	Precision	Recall	F_1 score	$AP@0.5^*$
Yuan et al. [49]	0.924	0.871	0.879	0.937
Donmez et al. [66]	0.970	0.950	0.960	Nan*
Wang et al. [48]	0.910	0.910	0.910	Nan*
Tian et al. [67]	0.910	0.900	0.905	Nan*
Our FruitNet	**0.923**	**0.889**	**0.906**	**0.943**

* $AP@0.5$ denotes the average precision when the $IoU > 0.5$, and Nan denotes no data reported.
The bold entity indicates the maximum value (best result) for each accuracy metric.

D. Limitations and Future Work

Despite much effort, there are still several limitations in dataset creation, model selection, and hyperparameter optimization, as stated above. First, although we acquired both RGB and multispectral images simultaneously via UAV remote sensing, only UAV-based RGB images were used to construct the IPTIS dataset in the present study. UAV-based multispectral images will be used in a future study with the hope of improving the accuracy of the models. Other UAV-based high-resolution images, such as hyperspectral or LiDAR imagery, could be better options for detecting IPTs because their spectral information

or highly effective point cloud data [5], [43] could reveal more detailed features and improve the performance of CNNs that help distinguish IPTs from the images accurately. Second, although a comparative study of YOLOv5 and its optimized counterparts was carried out for object detection, many other machine learning and deep learning models for semantic or instance segmentation should be analyzed to overcome the shortcomings of object detection. More state-of-the-art CNN models, such as U-Net, mask R-CNN, and YOLOv8 [21], [44], [45], [46], could be optimized and trained to obtain better performance. More modifications and optimizations of the architectures of the selected models could be made to further improve accuracy and robustness [49], [68] for more precise applications of smart orchard management. Third, hyperparameter optimization and data augmentation need to be further carried out to enhance the robustness and performance of the proposed method. All of these questions deserve further in-depth research.

In the future, spatial and attribute data on IFTs (i.e., location, planting area, and number) in a large orchard (for example, a pomelo orchard) could be obtained through our proposed approach. These data could be easily integrated into a smart orchard management system that could provide rapid growth monitoring of IFTs, accurate fruit yield estimation, real-time disease prevention and control, and precise cultivation and management. In addition, town-, county-, and city-level thematic maps of IPTs could be generated through the proposed approach in a forthcoming study. A precise pomelo yield estimation study based on a thematic map of IPTs will be a key topic of future research.

V. Conclusion

The present study proposed an optimized deep learning approach for detecting and mapping IFTs with UAV remote sensing imagery, taking pomelo trees in Meizhou city as experimental examples. UAV remote sensing technology was applied to acquire high-spatial-resolution images of the study areas. These images were preprocessed to produce DOMs and DSMs using DJI Terra. The novel IPTIS dataset was, subsequently, created through visual interpretation and fieldwork investigation. Five different scales of YOLOv5 (i.e., YOLOv5n, YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x) object detection models were used to train and validate on the IPTIS dataset. The evaluation results show that YOLOv5s performs the best with the highest accuracy among the five models. Consequently, YOLOv5s was selected as the baseline model to be optimized using seven widely used attention mechanisms in computer vision. The accuracy evaluation results show that the CA-based YOLOv5s model, namely, YOLOv5s-CA, outperforms the other six attention-optimized YOLOv5s models. YOLOv5s-CA was, thus, selected and named FruitNet to detect the IPTs in the whole mosaic orthographic images of the study areas. Finally, after postprocessing, two spatial distribution thematic maps of the IPTs at sites A and B were successfully made based on the detection results of FruitNet. Our results demonstrate that the attention-optimized YOLOv5s-CA model slightly increases the accuracy of IPT detection compared with its base model (YOLOv5s) and that FruitNet is suitable for accurately and efficiently detecting IPTs integrated with UAV remotely sensed imagery for precision agriculture and smart orchards. Our hypothesis was confidently verified in the present study, which could provide reference information for smart orchard management and related research.

Acknowledgment

The authors would like to thank Prof. L. Xu, Prof. G. Zhong, Prof. L. Zhang, Dr. Z. Xie, Dr. Y. Liu, C. Zhu, and X. Ma at Jiaying University for their great help in pomelo orchard selection, image acquisition, and discussions on the experimental design, and the editors and reviewers for their valuable comments and suggestions that greatly improved our article. The growers of the selected orchards are appreciated for their kind help.

References

[1] C. B. Kukunda, J. Duque-Lazo, E. González-Ferreiro, H. Thaden, and C. Kleinn, "Ensemble classification of individual Pinus crowns from multispectral satellite imagery and airborne LiDAR," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 65, pp. 12–23, 2018.

[2] Y. Ampatzidis, V. Partel, and L. Costa, "Agroview: Cloud-based application to process, analyze and visualize UAV-collected data for precision agriculture applications utilizing artificial intelligence," *Comput. Electron. Agriculture*, vol. 174, 2020, Art. no. 105457.

[3] H. Masood et al., "Recognition and tracking of objects in a clustered remote scene environment," *Comput., Mater. Continua*, vol. 70, pp. 1699–1719, 2022.

[4] A. Hamza et al., "An integrated parallel inner deep learning models information fusion with Bayesian optimization for land scene classification in satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 9888–9903, Oct. 2023.

[5] D. Jaskierniak et al., "Individual tree detection and crown delineation from unmanned aircraft system (UAS) LiDAR in structurally complex mixed species eucalypt forests," *ISPRS J. Photogramm. Remote Sens.*, vol. 171, pp. 171–187, 2021.

[6] D. Li and M. Li, "Research advance and application prospect of unmanned aerial vehicle remote sensing system," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 39, no. 5, pp. 505–513, 2014.

[7] C. Zhang, J. Valente, L. Kooistra, L. Guo, and W. Wang, "Orchard management with small unmanned aerial vehicles: A survey of sensing and analysis approaches," *Precis. Agriculture*, vol. 22, no. 6, pp. 2007–2052, 2021.

[8] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[9] L. P. Osco et al., "A review on deep learning in UAV remote sensing," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, 2021, Art. no. 102456.

[10] V. Dillshad, M. A. Khan, M. Nazir, O. Saidani, N. Alturki, and S. Kadry, "D2LFS2Net: Multi-class skin lesion diagnosis using deep learning and variance-controlled marine predator optimisation: An application for precision medicine," *CAAI Trans. Intell. Technol.*, to be published, doi: 10.1049/cit2.12267.

[11] M. A. Khan et al., "TS2HGRNet: A paradigm of two stream best deep learning feature fusion assisted framework for human gait analysis using controlled environment in smart cities," *Future Gener. Comput. Syst.*, vol. 147, pp. 292–303, 2023.

[12] M. A. Khan et al., "HGRBOL2: Human gait recognition for biometric application using Bayesian optimization and extreme learning machine," *Future Gener. Comput. Syst.*, vol. 143, pp. 337–348, 2023.

[13] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, 2020.

[14] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[15] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[16] A. V. Etten, "You only look twice: Rapid multi-scale object detection in satellite imagery," 2018, *arXiv:1805.09512*.

[17] J. Yan et al., "Recognition of Rosa roxbunghii in natural environment based on improved faster R-CNN," *Trans. Chin. Soc. Agricultural Eng.*, vol. 35, no. 18, pp. 143–150, 2019.
[18] Y. Xiong, Q. Chen, M. Zhu, Y. Zhang, and K. Huang, "Accurate detection of historical buildings using aerial photographs and deep transfer learning," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2020, pp. 1592–1595.
[19] J. Liu and X. Wang, "Tomato disease and pest detection algorithm based on YOLO convolutional neural network," *China Cucurbits Vegetables*, vol. 33, no. 9, pp. 18–38, 2020.
[20] J. Xiong et al., "Visual detection of green mangoes by an unmanned aerial vehicle in orchards based on a deep learning method," *Biosyst. Eng.*, vol. 194, pp. 261–272, 2020.
[21] T. Jintasuttisak, E. Edirisinghe, and A. Elbattay, "Deep neural network based date palm tree detection in drone imagery," *Comput. Electron. Agriculture*, vol. 192, 2022, Art. no. 106560.
[22] T. Liu et al., "Pineapple (*Ananas comosus*) fruit detection and localization in natural environment based on binocular stereo vision and improved YOLOv3 model," *Precis. Agriculture*, vol. 24, no. 1, pp. 139–160, 2023.
[23] W. Li, H. Fu, L. Yu, and A. Cracknell, "Deep learning based oil palm tree detection and counting for high-resolution remote sensing images," *Remote Sens.*, vol. 9, no. 1, 2017, Art. no. 22.
[24] W. Li, R. Dong, H. Fu, and L. Yu, "Large-scale oil palm tree detection from high-resolution satellite images using two-stage convolutional neural networks," *Remote Sens.*, vol. 11, no. 1, 2019, Art. no. 11.
[25] G. B. Weinstein, S. Marconi, S. Bohlman, A. Zare, and E. White, "Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1309.
[26] M. Freudenberg, N. Nölke, A. Agostini, K. Urban, F. Wörgötter, and C. Kleinn, "Large scale palm tree detection in high resolution satellite images using U-net," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 312.
[27] A. Koirala, K. B. Walsh, Z. Wang, and C. McCarthy, "Deep learning—Method overview and review of use for fruit detection and yield estimation," *Comput. Electron. Agriculture*, vol. 162, pp. 219–234, 2019.
[28] J. Wu et al., "Extracting apple tree crown information from remote imagery using deep learning," *Comput. Electron. Agriculture*, vol. 174, 2020, Art. no. 105504.
[29] B. G. Weinstein, S. Marconi, S. A. Bohlman, A. Zare, and E. P. White, "Cross-site learning in deep learning RGB tree crown detection," *Ecol. Inform.*, vol. 56, 2020, Art. no. 101061.
[30] A. Pleşoianu, M. Stupariu, I. ****andric, I. Pătru-Stupariu, and L. Drăguţ, "Individual tree-crown detection and species classification in very high-resolution remote sensing imagery using a deep learning ensemble model," *Remote Sens.*, vol. 12, no. 15, 2020, Art. no. 2426.
[31] M. Culman, S. Delalieux, and K. Van Tricht, "Individual palm tree detection using deep learning on RGB imagery to support tree inventory," *Remote Sens.*, vol. 12, no. 21, 2020, Art. no. 3476.
[32] J. Zheng et al., "Cross-regional oil palm tree counting and detection via a multi-level attention domain adaptation network," *ISPRS J. Photogramm. Remote Sens.*, vol. 167, pp. 154–177, 2020.
[33] M. P. Ferreira et al., "Individual tree detection and species classification of Amazonian palms using UAV images and deep learning," *Forest Ecol. Manage.*, vol. 475, 2020, Art. no. 118397.
[34] M. Brandt et al., "An unexpectedly large count of trees in the West African Sahara and Sahel," *Nature*, vol. 587, no. 7832, pp. 78–82, 2020.
[35] N. P. Hanan and J. Y. Anchang, "Satellites could soon map every tree on Earth," *Nature*, vol. 587, pp. 42–43, 2020.
[36] H. Liu, P. Dong, C. Wu, P. Wang, and M. Fang, "Individual tree identification using a new cluster-based approach with discrete-return airborne LiDAR data," *Remote Sens. Environ.*, vol. 258, 2021, Art. no. 112382.
[37] X. Xu, Z. Zhou, Y. Tang, and Y. Qu, "Individual tree crown detection from high spatial resolution imagery using a revised local maximum filtering," *Remote Sens. Environ.*, vol. 258, 2021, Art. no. 112397.
[38] T. Yun et al., "Individual tree crown segmentation from airborne LiDAR data using a novel Gaussian filter and energy function minimization-based approach," *Remote Sens. Environ.*, vol. 256, 2021, Art. no. 112307.
[39] J. Mäyrä et al., "Tree species classification from airborne hyperspectral and LiDAR data using 3D convolutional neural networks," *Remote Sens. Environ.*, vol. 256, 2021, Art. no. 112322.
[40] T. Yin, J. Zeng, X. Zhang, and X. Zhou, "Individual tree parameters estimation for Chinese Fir (Cunninghamia lanceolate (Lamb.) Hook) plantations of South China using UAV oblique photography: Possibilities and challenges," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 827–842, Nov. 2021.
[41] H. Luo, K. Khoshelham, C. Chen, and H. He, "Individual tree extraction from urban mobile laser scanning point clouds using deep pointwise direction embedding," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 326–339, 2021.
[42] L. Yao, T. Liu, J. Qin, N. Lu, and C. Zhou, "Tree counting with high spatial-resolution satellite imagery based on deep neural networks," *Ecol. Indicators*, vol. 125, 2021, Art. no. 107591.
[43] J. Hu et al., "A robust deep learning approach for the quantitative characterization and clustering of peach tree crowns based on UAV images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4408613.
[44] A. A. D. Santos et al., "Assessment of CNN-based methods for individual tree detection on images captured by RGB cameras attached to UAVs," *Sensors*, vol. 19, no. 16, 2019, Art. no. 3595.
[45] A. Safonova, E. Guirado, Y. Maglinets, D. Alcaraz-Segura, and S. Tabik, "Olive tree biovolume from UAV multi-resolution image segmentation with mask R-CNN," *Sensors*, vol. 21, no. 5, 2021, Art. no. 1617.
[46] K. Yu et al., "Comparison of classical methods and mask R-CNN for automatic tree detection and mapping using UAV imagery," *Remote Sens.*, vol. 14, no. 2, 2022, Art. no. 295.
[47] M. Mohan et al., "Optimizing individual tree detection accuracy and measuring forest uniformity in coconut (Cocos nucifera L.) plantations using airborne laser scanning," *Ecol. Model.*, vol. 409, 2019, Art. no. 108736.
[48] H. Wang, N. Han, C. Lu, W. Mao, M. Li, and L. Li, "Recognition and segmentation of individual citrus tree crown based on mask R-CNN," *Trans. Chin. Soc. Agricultural Mach.*, vol. 52, no. 5, pp. 169–174, 2021.
[49] H. Yuan, K. Huang, C. Ren, Y. Xiong, J. Duan, and Z. Yang, "Pomelo tree detection method based on attention mechanism and cross-layer feature fusion," *Remote Sens.*, vol. 14, no. 16, 2022, Art. no. 3902.
[50] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, 2010.
[51] Ultralytics, "YOLOv5 in PyTorch," 2022, Accessed on: Mar. 1, 2022. [Online]. Available: https://github.com/ultralytics/yolov5
[52] Ultralytics, "YOLOv8 in PyTorch," 2023, Accessed on: Jan. 28, 2023. [Online]. Available: https://github.com/ultralytics/ultralytics
[53] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.
[54] X. Zhu, S. Lyu, X. Wang, and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios," in *Proc. IEEE Int. Conf. Compu. Vis.*, 2021, pp. 2778–2788.
[55] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13708–13717.
[56] E. Y. Obsie, H. Qu, Y.-J. Zhang, S. Annis, and F. Drummond, "Yolov5s-CA: An improved Yolov5 based on the attention mechanism for mummy berry disease detection," *Agriculture*, vol. 13, no. 1, 2023, Art. no. 78.
[57] H. Zhang et al., "ResNest: Split-attention networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2735–3745.
[58] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[59] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," *Comput. Vis.—ECCV Lecture Notes Compu. Sci.*, vol. 11211, 2018.
[60] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1577–1586.
[61] L. Yang, R.-Y. Zhang, L. Li, and X. Xie, "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proc. 38th Int. Conf. Mach. Learn.*, 2021, pp. 11863–11874.
[62] J. Schulman et al., "Introducing ChatGPT," 2022, Accessed on: Dec. 20, 2022. [Online]. Available: https://openai.com/blog/chatgpt
[63] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, and A. Ku, "Image transformer," in *Proc. Int. Conf. Mach. Learn.*, vol. 80, 2018, pp. 4055–4064.
[64] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.
[65] O. Csillik, J. Cherbini, R. Johnson, A. Lyons, and M. Kelly, "Identification of citrus trees from unmanned aerial vehicle imagery using convolutional neural networks," *Drones*, vol. 2, no. 4, 2018, Art. no. 39.
[66] C. Donmez, O. Villi, S. Berberoglu, and A. Cilek, "Computer vision-based citrus tree detection in a cultivated environment using UAV imagery," *Comput. Electron. Agriculture*, vol. 187, 2021, Art. no. 106273.

[67] H. Tian et al., "Extraction of citrus trees from UAV remote sensing imagery using YOLOv5s and coordinate transformation," *Remote Sens.*, vol. 14, 2022, Art. no. 4208.
[68] M. H. Saleem, J. Potgieter, and K. M. Arif, "Automation in agriculture by machine and deep learning techniques: A review of recent developments," *Precis. Agriculture*, vol. 22, no. 6, pp. 2053–2091, 2021.

Yongzhu Xiong (Member, IEEE) received the B.Sc. degree in geology and the M.Sc. degree in paleontology and stratigraphy from the Chengdu University of Technology, Chengdu, China, in 1997 and 2003, respectively, and the Ph.D. degree in human geography from the University of Chinese Academy of Sciences, Beijing, China, in 2007.

He is currently a Professor with the School of Geography and Tourism, Jiaying University, Meizhou, China. He has authored or coauthored more than 60 papers in scientific journals, such as *Remote Sensing*, *Energies*, and *Remote Sensing Applications: Society and Environment*, and a textbook titled *Introduction to Remote Sensing Fundamentals* (Jilin University Press, 2017) as the Chief Editor. He was elected as a member of the "Thousand–Hundred–Ten" Talents Program of Guangdong Province in 2010. His research interests include ecosystem service and sustainability, UAV and agricultural remote sensing, environmental remote sensing, and geospatial big data and artificial intelligence.

Xiaofeng Zeng received the B.Sc. degree in geographic information science from Jiaying University, Meizhou, China, in 2021.

He is currently a Software Developer with Guangdong Airace Technology Company Ltd., Huizhou, China. His research focuses on application platform development of geographic information system and remote sensing.

Weiqian Lai received the B.Sc. degree in geographic information science from Jiaying University, Meizhou, China, in 2021.

He is currently a Remote Sensing Artificial Intelligence Engineer with Satxspace Technology Company Ltd., Dong-guan, China. His research focuses on remote sensing artificial intelligence algorithm development and application.

Jiawen Liao, photograph and biography not available at the time of publication.

Yankui Chen received the B.Sc. degree in geographic information system from Jiaying University, Meizhou, China, in 2010, and the M.Sc. degree in cartography and geographic information system from Guangxi Normal University, Nanning, China, in 2013.

He is currently an Experimentalist with the School of Geography and Tourism, Jiaying University, Meizhou, China. He has authored or coauthored more than 20 papers in scientific journals. His research interests include land use planning, land information technology, and remote sensing application.

Mingyong Zhu received the B.Sc. degree in geography from Lanzhou University, Lanzhou, China, in 2001, and the M.Sc. and Ph.D. degrees in ecology from the University of Chinese Academy of Sciences, Beijing, China, in 2007 and 2010, respectively.

He is currently an Associate Professor with the School of Geography and Tourism, Jiaying University, Meizhou, China. He has authored or coauthored more than 30 papers in scientific journals, such as *Environmental Research*, *Energies*, *Ecohydrology*, *Heliyon*, *Chemosphere*, and *Journal of Environmental Radioactivity*, and a book titled *Ten-Year Change Assessment of Ecological Environment in the Energy and Chemical Industry Zone in the Middle and Upper Reaches of the Yellow River (2000–2010)* (Science Press, 2017) as the lead author. His research interests include soil erosion and remote sensing application.

Kekun Huang (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in applied mathematics from Sun Yat-Sen University, Guangzhou, China, in 2002, 2005, and 2016, respectively.

He is currently a Professor with the Department of Mathematics, Jiaying University, Meizhou, China. He has authored or coauthored more than 20 papers in top journals, such as IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON CYBERNETICS, and *Pattern Recognition*. He was elected as a Nanyue Excellent Teacher of Guangdong province. His research interests include image processing and pattern recognition.