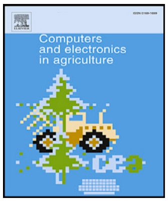
Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag



Detection of tomato plant phenotyping traits using YOLOv5-based single stage detectors

Angelo Cardellicchio [a], Firozeh Solimani [a], Giovanni Dimauro [b], Angelo Petrozza [c], Stephan Summerer [c], Francesco Cellini [c], Vito Renò [a,*]

[a] *Institute of Intelligent Industrial Technologies and Systems for Advanced Manufacturing, National Research Council of Italy, Via Amendola 122, D/O, Bari, 70126, Italy*

[b] *University of Bari, Department of Computer Science, Via E. Orabona, 4, Bari, 70125, Italy*

[c] *ALSIA Centro Ricerche Metapontum Agrobios, s.s. Jonica 106, km 448.2, Metaponto, 75010, Italy*

ARTICLE INFO

Keywords:
Tomato plant phenotyping
YOLOv5
Nodes
Fruit
Flower identification

ABSTRACT

Plant phenotyping is the study of complex plant traits to evaluate its status depending on the life-cycle conditions. Often, these evaluations are carried out by human operators, and the accuracy could be biased by their experience and skill, especially when dealing with huge amounts of data produced by high-throughput phenotyping (HTP) platforms. With the rapid development of key enabling technologies, HTP is only made possible by the vast amounts of data made available by computer vision systems. In this scenario, artificial intelligence algorithms play a key role in the automation, standardization, and quantitative analysis of large data. This paper focuses on detecting tomato plants phenotyping traits using single-stage detectors (either stand-alone or ensemble) based on YOLOv5, aiming to effectively identify nodes, fruit, and flowers on a challenging dataset acquired during a stress experiment conducted on multiple tomato genotypes. Results demonstrate that the models achieve relatively high scores, considering the particular challenges of the input images in terms of object size, similarity between objects, and their color.

1. Introduction

The exponential population growth experienced over the last century has increased the need for a sustainable food supply. However, adverse factors related to disturbances in plant growth, development, tolerance, resistance, architecture, physiology, and ecology have reduced crop yields in agriculture, causing irretrievable losses to agricultural production. Furthermore, extensive agriculture has led to environmental crises and the consequent depletion of natural resources. Consequently, new technologies have been developed to address the challenges of unpredictable agricultural ecosystems to reduce disasters and improve farming practices (Vasconez et al., 2020). For example, precision agriculture, for which the use of aims to reduce costs, increase production, and reduce environmental impact.

Many researchers have been working on applying high-throughput phenotyping by providing advanced systems, such as crop phenotyping platforms, often including environmental sensor networks, robots, unmanned vehicles, tractors, fixed-point monitoring stations, drones, or satellites. These systems have been used to produce and exploit a massive amount of data to correctly identify and classify plants in a reasonable time while keeping costs relatively low (Arunachalam and Andreasson, 2021). Consequently, image processing and artificial intelligence techniques have been used on data acquired by satellites, drones, and RGB cameras for non-destructive testing of crops (Mirhaji et al., 2021).

An interesting set of applications involves deep architectures, such as convolutional neural networks (CNNs). Specifically, two types of architectures have been used as detectors: *two-stages detectors*, such as R-CNN and Fast R-CNN (Girshick, 2015), and *single-stage* detectors, such as YOLO and its successors (Redmon et al., 2016).

Both these types have been successfully exploited in phenotyping-oriented scenarios. However, many studies have favored the YOLO-based algorithms due to their high accuracy and faster detection speed (Wang and Liu, 2021). For example, YOLO and its successors have been used for detecting the inter-node length of cucumbers (Boogaard et al., 2020), kiwifruit (Suo et al., 2021), grapes (Santos et al., 2020), cherries (Gai et al., 2021), apple varieties (Fan et al., 2022) and tomatoes (Wang and Liu, 2021; Lawal, 2021a; Magalhães et al., 2021). Furthermore, the YOLO architecture has been constantly evolved towards denser, richer, and more complex models. At the time of writing, the denser model is called YOLOv5, which has led to a

[*] Corresponding author.
E-mail address: vito.reno@stiima.cnr.it (V. Renò).

https://doi.org/10.1016/j.compag.2023.107757
Received 29 July 2022; Received in revised form 28 February 2023; Accepted 4 March 2023
Available online 9 March 2023

Table 1
Results achieved by state-of-the-art approaches in plant detection. The *Time* column represents the inference time, that is, the time required by the network to perform object detection on an image. This value is not adjusted to deal with different hardware and image size.

Ref.	Model	Plant	mAP (%)	Time (ms)	Focus	Images(#)
Fan et al. (2022)	YOLOv4 P	Apple	93.74	8.36	NIR Images	5700
Fu et al. (2022)	YOLO-Banana	Banana	92.19	35.33	Shrubs/Stems	120
Gai et al. (2021)	YOLOv4-dense	Cherry	N.A.	467	Mature/immature	400
Lawal (2021b)	YOLODenseNet	Tomato	98.3	21.88	Detection	485
	YOLOMixNet	Tomato	98.4	21.1		
Li et al. (2021)	YOLOv4-tiny	Pepper	95.11	11.24	Detection	145
Liu et al. (2020)	YOLO-Tomato	Tomato	94.58	54	Occlusions/lighting	966
	YOLO-Tomato	Occlusions	90.10	54		
Liu and Wang (2020)	YOLOv3	Tomato	92.39	20.39	Diseases/pests	15 000
Liu et al. (2022)	Improved YOLOv5	Citrus	98.4	19	Detection	16 000
Mirhaji et al. (2021)	YOLOv4	Orange	90.8	23.6	Stress/Daytime	30 059
Qi et al. (2022)	SE-YOLOv5	Tomato	94.10	50.63	Diseases	1036
Roy and Bhaduri (2022)	YOLOv4	Tomato	96.29	14.24	Diseases	12 000
Ruparelia et al. (2022)	YOLOv3	Tomato	81.28	60.38	Ripe/unripe, infections	2000
	YOLOv4	Tomato	78.49	68.12		
Sozzi et al. (2022)	YOLOv5	Grapes	76.1	32.26	Detection	2985
Wang and He (2021)	YOLOv5s	Apple	N.A.	8	Detection	3165
Wang and Liu (2021)	YOLOv3	Tomato	96.41	20.28	Detection	3165
Yao et al. (2021)	YOLOv5-Ours	Kiwifruit	94.7	100	Diseases/deformations	1600
Zhang and Li (2022)	EPSA-YOLOv5s	Canola	99.6	547	Seedlings survival	3368
Zheng et al. (2022)	RC-YOLOv4	Tomato	94.44	93.37	Mature/immature	1698

noticeable increase in recognition speed and accuracy compared to its predecessors, making it an optimal choice for target recognition (Qi et al., 2022).

This research aims to use a deep learning approach based on the state-of-the-art YOLOv5 for identifying fruits, flowers, and nodes from tomato images. Tomato is one of the most widely grown fruits and vegetables in the world (Wang and Liu, 2021), and also an iconic crop in Italy. Therefore, the question of how to increase the production of this fruit has also become one of the main challenges for digital agriculture. For example, tomatoes may be subject to problems such as shading, resulting in inadequate fruit quality. Furthermore, it is essential to distinguish between ripe and unripe fruits quickly. To address such challenges, it is useful to extract meaningful information from images, concentrating on phenotyping traits such as tomato fruits, flowers, and stem nodes, accurately monitoring them during the whole growth of the plant. This information can be exploited to provide a constant assessment of the quality of the tomatoes, therefore increasing the overall production. As such, computer vision combined with a deep CNN will make it possible to monitor the response to external nutritional inputs (e.g. fertilization) or the response to abiotic stresses (e.g. drought and salt) through variation in the phenotypic traits under study. In cereals, these techniques have been successfully used to detect plant pests and diseases (Wang and Su, 2022). In addition to the ability to quickly obtain information on plant organs and abiotic stresses, and the ability to segment crops from weeds. The application can also be extended to the counting of leaves showing obvious symptoms, as in the case of leaf mines produced by *Tuta absoluta* in tomato. Infested leaves are phenotypically different from healthy ones in both shape, color as well as color patterns. Tomato fruit size and number (Mesa et al., 2022; Panthee et al., 2018), stem and internode elongation (Litvin et al., 2016) and flower number and setting (Panthee et al., 2018) are traits that are strongly influenced by both biotic and abiotic stresses. This work could be useful for the detection of biotic or abiotic stress in tomato plants. In the past work has involved single models for identifying single traits such as the use of machine learning for identification of tomato plants nodes (Yamamoto et al., 2016) or others using either machine learning (Yamamoto et al., 2014) and deep learning (Mu et al., 2020) to identify tomato fruit on plants. This work is focused on the identification of flowers, fruits and nodes all with a single CNN model.

This work has analyzed a challenging image dataset of the aerial part of tomato plants. The dataset has been labeled by domain experts, which have provided bounding boxes for the nodes on the main stem, flowers, and fruits.

The rest of this paper is organized as follows. In Section 2, the state-of-the-art applications of single-stage object detectors in phenotyping has been discussed. In Section 3, the framework used for the experiment has been described, while achieved results have been discussed in Section 4. Finally, in Section 5, the conclusions, along with a perspective of future works which can be performed to improve the system, has been provided.

2. Related work

Several studies have recently focused on identifying and classifying agricultural products based on models belonging to the YOLO family. An overview of the such studies, which are described in the following, has been provided in Table 1.

In general, models belonging to the YOLO family achieved more density over time and, therefore, more parameters to be automatically learned over consecutive iterations. This means YOLOv5 is denser than YOLOv4, which in turn is denser than YOLOv3, and so on. As expected, higher densities have implied improvements in average accuracy, while lower densities have allowed reaching higher detection speeds, usually paid for with lower accuracy values. This has been shown by a comparison performed in Ruparelia et al. (2022) on 2000 images of ripe, unripe, and infected tomatoes, where YOLOv4 achieves an overall mean average precision (mAP) of 81.28%, outperforming YOLOv3 which achieved an mAP of 78.49%. However, YOLOv3 has been proven to be slightly faster, performing a single detection in 60.38 ms if compared to the 68.12 ms required by YOLOv4. For this reason, authors have considered reasonable and advisable to consider both time and accuracy variables together carefully, to choose the most suitable approach for every different application.

Improved versions of base models have also shown robust performance in several scenarios. For example, the authors in Liu et al. (2020) have used a modified version of YOLOv3 called *YOLO-Tomato*, which proposes two main differences. First, *circular* bounding boxes have been used in place of normal rectangular bounding boxes to match the shape of tomatoes. Second, dense layers have been embedded within the backbone of the network. With those improvements, the authors have claimed a 94.58% accuracy under slight occlusion conditions on a dataset of 609 tomatoes, while 90.10% accuracy under severe occlusions on 303 tomatoes. Another series of improvements over the YOLOv3 model has been proposed by Liu and Wang (2020). Specifically, the first proposal is *feature fusion pyramid*, that is, low-level features gathered by the network have been fused with high-level features to preserve fine-grained details and semantic significance. Afterward, nine sets of prior boxes used to predict the coordinates of the bounding box have been extracted using K-means clustering; finally, multi-scale training has been used, by replacing the fully connected layer in the head of the model with a convolution layer with size 1 × 1 × 4096 to avoid image resizing at the input stage. With these improvements, authors have achieved an overall 92.39% accuracy on a dataset of 15 000 samples and 146 912 labeled boxes representing 12 different types of diseases and pests that can manifest on tomatoes. The model has been also proven to achieve a quite fast inference time of 20.39 ms. A similar set of improvements has been proposed by the same authors in Wang and Liu (2021), where dense connections have been used in the backbone, along with K-means for computing the size of anchor boxes and multiscale training. The model has been tested against the same dataset used in Liu and Wang (2020), achieving an improved accuracy of 96.41% with a reduced detection time of 20.28 ms. YOLOv3 has also been exploited as the basis for models used in automatic robotic platforms. Specifically, in Lawal (2021b), two variants of YOLOv3 have been proposed, called *YOLODenseNet* and *YOLOMixNet*. The two main difference between the two variants lies in their backbone: YOLODenseNet has been developed using a DenseNet-based backbone, while YOLOMixNet has been based on a mixture of DenseNet and DarkNEt. Still, both architectures have exploited a common set of machine learning techniques, such as image pyramid and complete IoU. Both networks have been tested against a dataset of 485 images for tomato detection, achieving an mAP of 98.3% and 98.4% for YOLODenseNet and YOLOMixNet, respectively, with an average detection speed of 21.88 and 21.1 ms.

As for models based on denser architectures, in Zheng et al. (2022), the authors have modified the standard CSPDarkNet53 backbone of YOLOv4 by adding residual layers, achieving an overall average accuracy of 94.44% on a dataset composed of 1698 images of mature and immature tomatoes. Also, Roy and Bhaduri (2022) has used a modified version of YOLOv4 with two types of layers within the backbone to enhance the receptive field and preserve fine-grain localized information for tomato disease detection. Specifically, the network has been tested against a dataset containing 12 000 images of four different plant diseases, that is, early and late blight, Septoria leaf spot, and leaf mold achieving an overall mAP of 96.29%. Finally, the authors in Qi et al. (2022) have used YOLOv5 with a modified backbone that exploited human attention mechanisms and a Squeeze-and-Excite module, achieving an mAP of 94.10% on a dataset composed by 1036 images representing different tomato virus diseases. One of the main insights about related work analysis is that there is the need to customize, change and evolve the models that rely on single-stage detectors such as YOLO, not only to achieve the results previously reported and recapped in Table 1 but also to deal with the specific challenges that characterize every dataset. Therefore, even though the working principle of such YOLO-based models has remained the same, a substantial effort must be put in performing adequate deep learning model customization by proposing meaningful architectural modifications.

Interestingly, other studies have also applied models from the YOLO family to perform detection on other agricultural products.

For example, the authors in Li et al. (2021) have used a modified version of the YOLOv4 tiny architecture with attention layers and an adaptive feature pyramid for green pepper detection. The network has been applied to detect 602 green peppers within a dataset of 145 images, achieving an mAP of 95.11%. In Gai et al. (2021), the authors have replaced the backbone of YOLOv4 with DenseNet, training the obtained network with a variable number of images, from ten to 400, representing three different stages of the lifecycle of cherry, that is, mature, semi-mature and immature stages. Results have been reported in terms of F_1 scores, which is 0.947 for the proposed architecture. Fruit stress estimation has also been explored by Mirhaji et al. (2021), which has proposed a comparison between three different versions of the YOLO family, specifically YOLOv2, v3, and v4. As expected, the latest model has outperformed the others, achieving an mAP of 90.8% on a dataset composed of 30 059 oranges acquired under varying illumination conditions (that is, different times of night and day). In Fu et al. (2022), authors have been focused on model performance, proposing a pruned variant of YOLOv4 called *YOLO-Banana* to investigate banana shrubs and stems in the wild. To this end, the authors have tested a dataset containing 163 banana bunches and 141 stalks in 120 images, achieving an mAP of 92.19%. As for apple detection, in Fan et al. (2022) the authors have proposed a variation of the original YOLOv4 model, which uses model pruning and non-maximum suppression to refine predictions and improve inference speed. The network has been tested on a dataset containing 5700 images of different cultivars of apples acquired in the NIR, with labeled regions belonging to three classes: stems, calyxes, and defects. This modified version of YOLO has achieved an overall mAP of 93.74%. The latest version of YOLO, i.e., YOLOv5, has also been extensively used. For example, in Sozzi et al. (2022), the authors have compared different architectures based on YOLOv5 with YOLOv4 on a dataset composed of 2985 images of grapes, showing how the first can achieve an overall mAP of 76.1%, which is noticeably higher than the one achieved by its predecessors. In Yao et al. (2021), the authors have modified the base model of YOLOv5 by adding two layers, that is, a small object detection layer to allow for small defects detection and an attention layer to consider the importance of different channels, introducing a new loss function called *CIoU*. The modified network has been tested against a dataset containing 1600 images containing data about disease, mold, speckle, and deformation on kiwifruit, achieving an overall mAP of 94.7%. Pruning has also been used on YOLOv5 by authors in Wang and He (2021), achieving results in terms of recall, precision, and F1 score of 87.6%, 95.8% and 91.5%, respectively, on a dataset composed by 3165 apple fruitlet images. The ESPA attention mechanism has been introduced by Zhang and Li (2022) to determine the survival rate of canola seedlings at different growth stages. The network has been tested on a dataset containing 3368 images, achieving an average accuracy of 99.6%. YOLOv5 has also been applied to citrus fruit detection in Liu et al. (2022), where the authors have shown the improvements achieved by the original architecture by adding attention mechanisms and replacing the PANet multiscale feature fusion network with a BiFPN. When tested against a dataset composed of 16 000 images (with data augmentation), the network has shown an mAP value of 98.4%.

Considering that few studies have been conducted on tomato identification and classification based on the YOLOv5 model, this work proposes an investigation on the developed models of this architecture for single and multi-classification of tomato cultivars.

3. Materials and methods

In this section, the materials and methods used for the proposed approach are described. First, in Section 3.1 the system and tools for HTP have been described. Then, the acquisition settings has been detailed in Section 3.2, while the dataset itself has been presented in Section 3.3. Afterward, in Section 3.4, an overview on the experimental setup has been provided, while a brief description of the principles underlying the operation of the YOLO architecture has been provided in Section 3.5. Finally, a brief overview on the metrics used for results evaluation has been provided in 3.6.

3.1. Plant phenotyping description

The High Throughput Plant Phenomics Platform (HTP) that has been used in this work is located within the PhenoLab at the ALSIA Metapontum Agrobios Research Centre is based on a LemnaTec Scanalyzer3D system. The system is equipped with the following components:

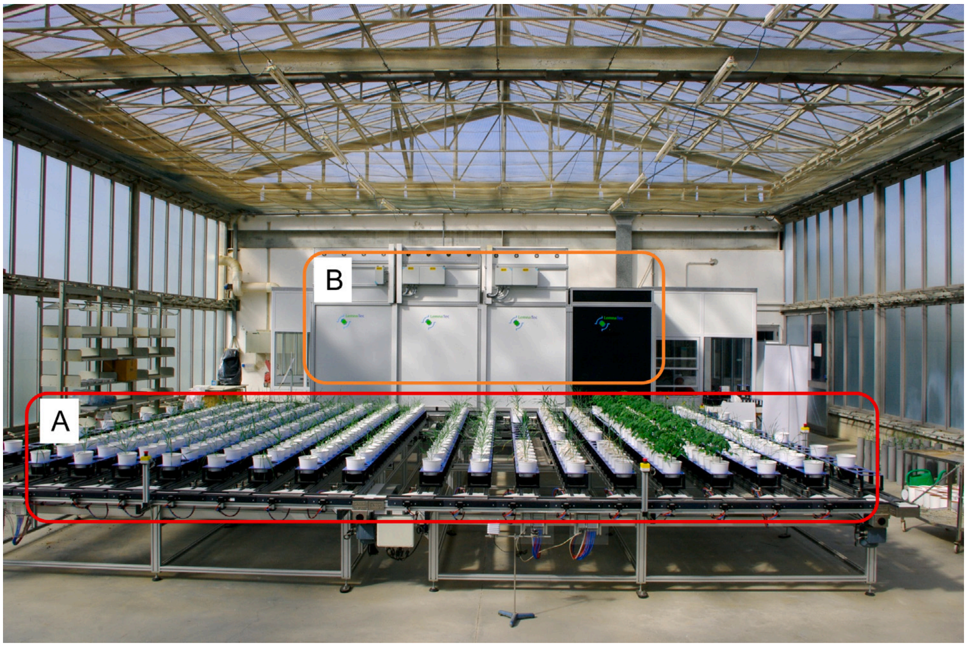


Fig. 1. The setup used for dataset gathering. The foreground of the image (A) contains the plant storage system with conveyor belts that carry the plants to the imaging chambers. The background of the image (B) contains the imaging chambers, which are for, from left to right (the actual direction of plant travel), *soil NIR*, *fluorescence*, visible light and *plant NIR imaging*.

- An automated belt conveyor system, capable of accommodating 494 plants in pots, with a tracking system based on bar code and RFID for safe identification of single plants.
- Four sequential camera stations used to acquire 3D images of plants using near-infrared (NIR), ultraviolet (UV), visible light (RGB), and a special NIR camera dedicated to capturing roots.
- An automated watering system with a weighting station.
- An ICT infrastructure for data acquisition, management, and processing.

The platform, shown in Fig. 1, allows the quantitative, non-destructive analysis of different crops or model plants under high-throughput conditions. Each plant has been imaged sequentially in multiple Scanalyzer3D camera units, employing the available wavelengths, resulting in a high number of reproducible and significant data points related to any aspect of the development of the plant.

3.2. Acquisition settings

Tomato plants have been automatically conveyed to the imaging chamber. Three images have been acquired per plant: one from above the plant (Top View, TV) and two from the lateral sides (Side View, SV) with a relative angle of 90 degrees. The plants have been illuminated by standard fluorescence light tubes (35 W/865 cool daylight) and recorded with a Basler Scout camera. As for the RGB camera, it has been based on a SONY ICX274 CCD sensor, with the following characteristics: a **KAI** 2093 sensor whose size is 1688 × 1248, with a global shutter and resolution of about 2.11 MPs, and pixel size of 8.50 × 6.80 mm. Finally, the lens has a 2/3 format, with *C* mount, 12.5 – 75.0 mm focal lens, a max aperture of 1 : 1.8 with a type 1 motor with 6 V and a maximum of 36 mA.

3.3. Dataset description

This work is focused on the SV image dataset, as the side viewpoint better highlights the classes to be identified, especially the nodes. In detail, the dataset has been acquired during a stress experiment conducted on 15 tomato genotypes using the HTP phenotyping platform. Plants have been grown in 3.2 liter pots containing 1.8 kg of sand-peat mixture. Drought stress has been applied through a 70% reduction of irrigation water in two stress cycles followed by recovery phases. During the 6-weeks trial, RGB images have been acquired on 11 different dates, each one corresponding to a specific experimental point, to obtain digital phenotypes of the control and drought-stressed plants.

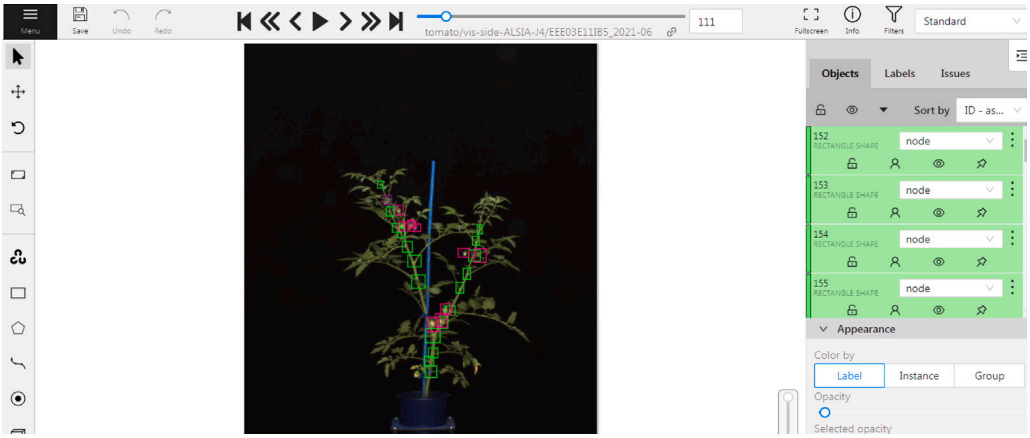
The SV image dataset contains 1683 images, with a fixed resolution of 1624 × 1234. Once gathered, images have been labeled by domain experts using *CVAT* (Sekachev et al., 2020), whose main interface is shown in Fig. 2. Specifically, three different classes of bounding boxes related to phenotyping traits, that is, *flowers*, *fruits* and *nodes* have been provided.

It is worth noticing that the first characteristic of the labeled dataset is that provided classes are not balanced, as shown in Fig. 3, which recaps the total number of labels within the training set. This bias could lead to model overfitting towards nodes (i.e., the largest class) if not properly handled; therefore, data augmentation techniques have been used, such as random affine transforms (i.e., rotation, scale, translation, and shear), HSV augmentation, and random horizontal split.

A final remark about the dataset images should be given about the appearance of the three classes in the images shown in Fig. 4. As the size of the plant is neither constant nor similar for all the images, the bounding box size spans from very small (e.g., (a), (b) in Fig. 4) to large (e.g., (c), (d) in Fig. 4). As such, nodes of early-stage plant growth could be extremely challenging to highlight. Furthermore, the color of the fruit ranges from yellow to green, as there are no red tomatoes among the images, and the flowers can have a completely different appearance.

3.4. Experimental setup

Experiments have been performed using the YOLOv5 library (Jocher et al., 2022). For the experiments, a machine with an Intel Core i9-11900HK CPU, 32 GBs of RAM, and an Nvidia GeForce RTX 3080 GPU with 10GBs of RAM has been used.



tomato/vis-side-ALSIA-J4/EEE03E11IB5_2021-06

111

Standard

Sort by ID - as...

152 RECTANGLE SHAPE node

153 RECTANGLE SHAPE node

154 RECTANGLE SHAPE node

155 RECTANGLE SHAPE node

Fig. 2. The Graphical User Interface provided by the Computer Vision Annotation Tool (CVAT). The user can manually insert bounding boxes on relevant object, and afterward labeling them with a proper annotation. In this case, a domani expert labeled some examples of flowers, fruit, and nodes.

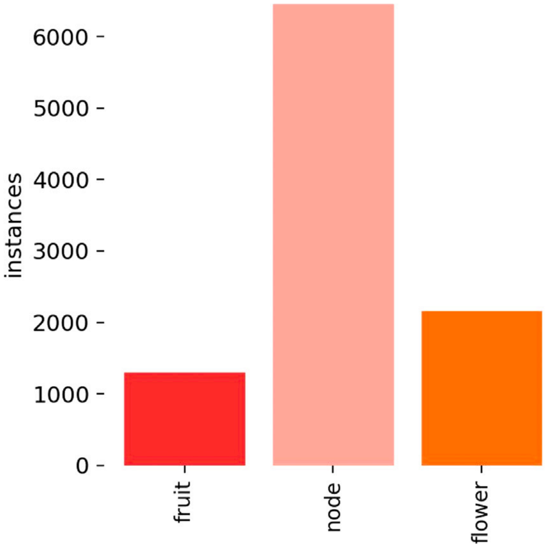


Fig. 3. Number of instances per class. Overall, 1862 fruits, 9276 nodes, and 3111 flower labels have been provided. Consequently, the dataset is imbalanced.

3.5. YOLOv5 for object detection

The process used by state-of-the-art object detectors based on deep CNN can be defined to properly introduce the architecture used in this work. Specifically, the operations of these detectors are as follows:

1. The input image is divided into a set of grids.
2. Each grid is fed to a *backbone* CNN that extracts features from it.
3. Such features are combined by the *neck* to model global relationships.
4. Both relationships and features are used by the *head* for the detection results.

Two types of detectors have been proposed. The first type, called *two-stages detector*, decouple the localization of the object from its classification, that is, in the first stage, objects are localized, while in the second, objects are classified. On the other hand, *one-stage detectors* combine the localization and the classification of the object into a single step. It has been shown that two-stages detectors provide a slightly higher accuracy, while single-stage detectors achieve higher detection speed.

The YOLOv5 architecture, which is one of the latest evolution of the YOLO object detectors, is a single-stage detector based on the following elements:

- As for the backbone, it is based on the latest version of CSP (*Cross-Stage-Partial-connections*)-Darknet53.
- As for the neck, it is composed by SPPF (Spatial Pyramid Pooling - Fast) and the latest version of CSP-PAN (Pyramid Attention Network).
- As for the head, which is the same used on previous architectures.

The underlying operation of YOLO-based architectures is summarized in Fig. 5. Specifically, images are split into an $S \times S$ grid, with S fixed value. Each grid predicts B bounding boxes, each one provided with a confidence score and C conditional class probabilities. By combining the confidence score of each bounding box with the relative class probability map, final detections are provided.

In this work, six different versions of the YOLOv5 architecture have been selected for comparison. These models are based on the same underlying architecture, as described in Tan et al. (2020), and differ mainly for their density, with the lowest given by the smallest and oldest version, that is, YOLOv5s, while the higher provided by YOLOv5x6. The need for comparing different versions of the same underlying architecture is directly related to the achievable accuracy. That is, denser networks usually provide better results in terms of evaluation metrics, given an adequate amount of data is provided for training, at the cost of slightly increases in inference time, which can be as high as 26.2 ms per picture for denser models. However, in this work, the focus is only on the accuracy, as the performance evaluation in terms of inference time will be tested in the future directly on the implementation testbed.

Due to the limited amount of available data, training from scratch the whole network has led to sub-optimal results. As a consequence, transfer learning has been used: the weights of the neurons within the backbone have been frozen, allowing the head to specialize on the problem itself while retaining the knowledge acquired by the backbone on larger datasets. Each run has required 300 epochs of transfer learning, after which results have been reported in terms of precision, recall, and mAP, as described in the next subsection. A standard 60/20/20 split has been used for training, validation, and test sets. Images have been resized to 1280 × 1280 pixels at the input stage.

3.6. Evaluation metrics

Let us briefly describe the evaluation metrics used to assess the detection results. To this end, let us briefly define the following:

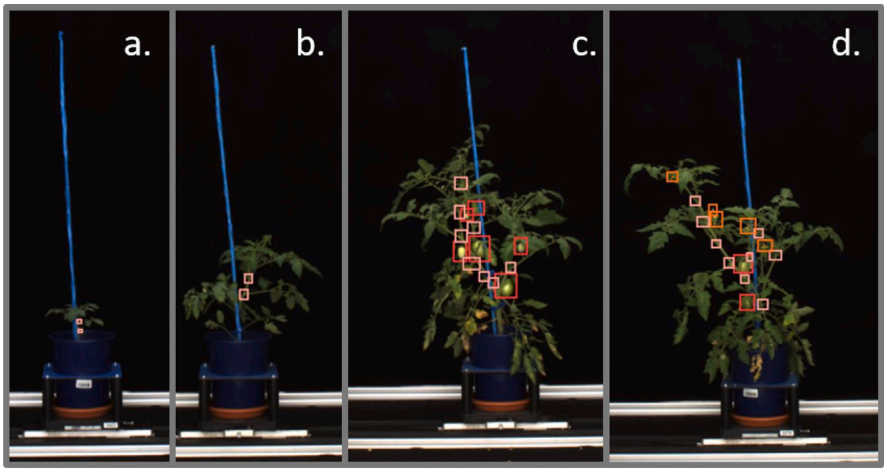


Fig. 4. Dataset labeled images example. The pink bounding boxes refer to the nodes, the orange ones to the flowers, and the red ones to the fruits.

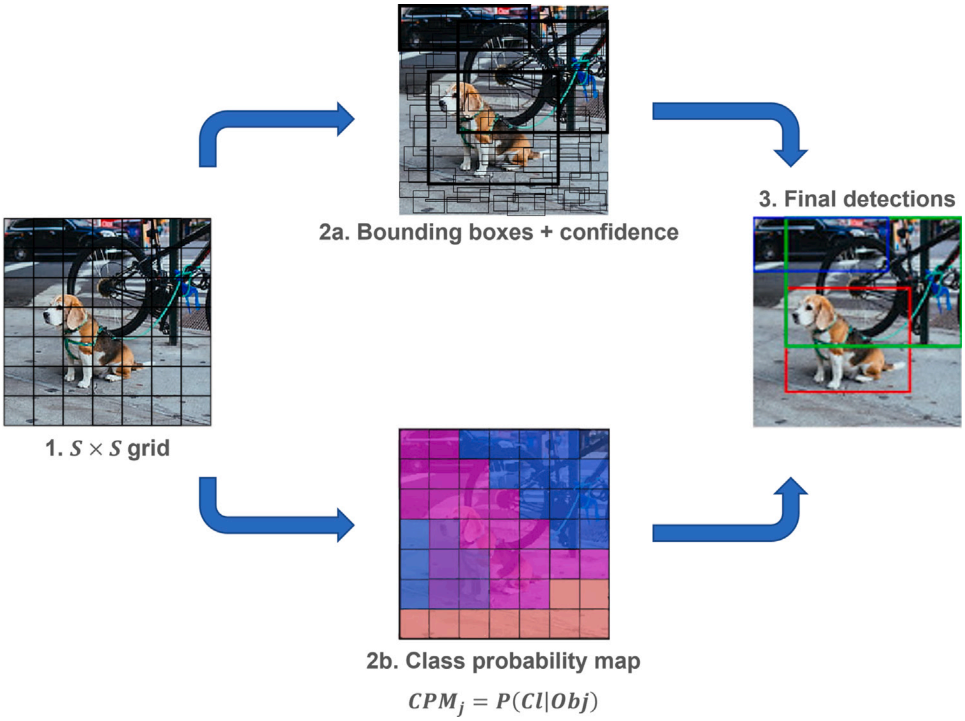


Fig. 5. Working principles of YOLO-based architectures. First, the detector divides the image in a set of $S \times S$ bounding boxes. Afterward, both a class probability map and a confidence score is computed per each estimated bounding box. In the last step, the confidence score is used to provide the final detections.

- A prediction labeled as $c(i,i)$ defines the case where the detector correctly classifies an instance of class i as belonging to the same class. This can be seen as a *true positive*.
- A prediction labeled as $c(j,i)$ defines the case where the detector incorrectly classifies an instance of class j as being an instance of class i. This can be seen as a case of *false positive*.
- A prediction labeled as $c(i,j)$ defines the case where the detector incorrectly classifies an instance of class i as being an instance of class j. This can be seen as a case of *false negative*.

From this, the *precision* of the detector associated with objects of class i is defined as:

$$P_i = \frac{c(i,i)}{\sum_j c(j,i)} \tag{1}$$

This means the precision describes the ratio between the number of times an object of class i has been *correctly* classified over the *total* number of times an instance of class i has been detected. *Recall* is defined as follows:

$$R_i = \frac{c(i,i)}{\sum_j c(i,j)} \tag{2}$$

That is, recall defines the ratio between the number of times an object of class i has been correctly identified over the total instances of objects of class i available within the dataset.

Let us note that the provided definition for precision and recall only account for class i. In the case of multiple classes, a weighted average over all classes is considered to provide the overall precision and recall, where the weight is usually provided by the number of

Table 2
Results achieved on tomato recognition. As expected, the wider architectures, that is, YOLOv5l6 and YOLOv5x6, provide the best results.

Class	Model	TP	M	B-FP	B-FN
Fruit	YOLOv5s	64%	1%	7%	35%
	YOLOv5m	73%	3%	7%	24%
	YOLOv5l	75%	3%	7%	23%
	YOLOv5l6	76%	3%	7%	22%
	YOLOv5x	75%	4%	8%	22%
	YOLOv5x6	78%	4%	8%	19%
Nodes	YOLOv5s	50%	0%	48%	50%
	YOLOv5m	64%	0%	54%	36%
	YOLOv5l	69%	0%	59%	31%
	YOLOv5l6	66%	0%	58%	34%
	YOLOv5x	69%	0%	59%	31%
	YOLOv5x6	68%	0%	59%	31%
Flowers	YOLOv5s	48%	1%	45%	51%
	YOLOv5m	56%	1%	39%	44%
	YOLOv5l	59%	2%	34%	38%
	YOLOv5l6	64%	2%	34%	34%
	YOLOv5x	59%	1%	33%	39%
	YOLOv5x6	64%	2%	33%	34%

instances belonging to each class i over the total number of instances within the dataset.

Precision and recall can be synthesized using the *F1 score*, defined as follows:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{3}$$

To assess the results achieved by the network, the *mean average precision* (*mAP*) is used. This metric is defined as follows:

$$mAP = \frac{1}{N} \sum_i AP_i \tag{4}$$

Where AP_i is the *average precision* for the class i, that is, the *area under curve* provided by the precision–recall plot for the detection of instances of class i.

4. Experiments and results

4.1. Results evaluation

4.1.1. Transfer learning on barebone architectures

Let us start this discussion with the results of the bare YOLOv5 architectures trained using transfer learning, as reported in Table 2.

In Table 2, **TP** represents *true positives*, i.e., the labeled objects which are correctly found by the network, while **M** represents *mismatches*, i.e., the labeled objects to which the network has assigned incorrect labels. As for **B-FP**, it represents *background false positives*, that is, boxes that have been found by the model but to which no labels provided by domain experts are found. Finally, **B-FN** represents *background false negatives*, that is, labeled bounding boxes that the network has not found.

A first remark about the low percentage of mismatches for all three classes should be given. This behavior shows that the YOLOv5-based detector on the proposed dataset can effectively process the images and identify fruits, flowers, and nodes. Let us underline that, in the proposed dataset, the appearance of fruit and flower can be extremely similar, hence even expert human operators can provide incorrect labels. This does not hold for nodes whose visual appearance differs from the other two classes.

It is clear how the best results are achieved using the denser networks, with a slight advantage in terms of overall accuracy for the YOLOv5l6 and YOLOv5x6 models. Furthermore, while fruit recognition shows low percentages regarding false positives and negatives, both nodes and flowers show significantly higher values. This result can be explained as follows.

Table 3
Results achieved on tomato recognition using the ensemble provided by YOLOv5x and YOLOv5x6, namely YOLOv5x+6. The overall improvement with respect to the bare versions of the architecture is about 3% for fruit, 6% for nodes, and 7% for flowers.

Class	TP	M	B-FP	B-FN
Fruit number	81%	4%	9%	15%
Nodes	75%	0%	56%	25%
Flowers	71%	2%	27%	34%

- Nodes present high levels of B-FP since domain experts labeled only nodes on the main stem. However, the network cannot differentiate between the main stem and the ancillary ones. Therefore, the high value in terms of B-FP is probably related to the identification of such nodes.
- Flowers, yellow chlorotic leaf tissue pixels may be misidentified as yellow flower pixels, thus leading to higher values for B-FP.
- Background false negatives are greatly reduced by using denser models. This suggests that these models can capture visual relationships at higher abstraction, properly characterizing objects.
- Interestingly, the previous point may also be confirmed by the fact that denser models have higher values of B-FP for nodes. This suggests that the model is finding more fine-grained nodes on auxiliary stems.

Let us further discuss the results achieved by the best architectures, YOLOv5l6 and YOLOv5x6, using some other metrics, specifically precision, recall, and F1 score. These results are shown in Figs. 6 and 7. Specifically, starting from the top left and moving clockwise, Figs. 6 and 7 show the *Precision/Confidence*, *Recall/Confidence*, *F1-score/Confidence* and *Precision/Recall* curves for the less dense architecture, that is, YOLOv5l6.

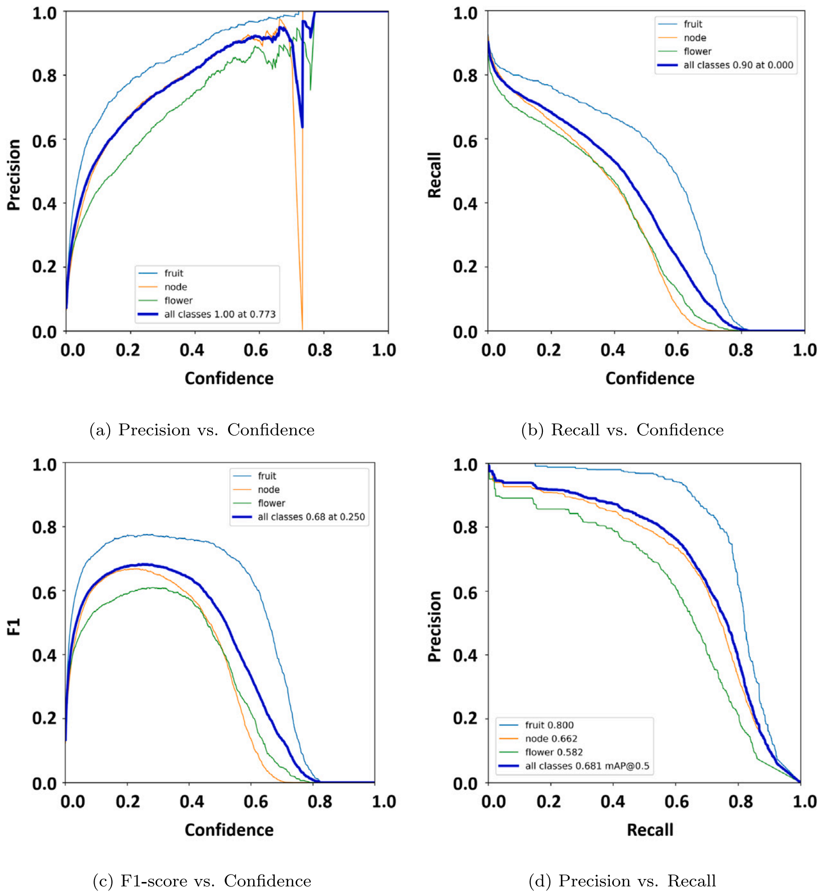
Specifically, it can be seen from Fig. 6(a) a noticeable drop in terms of precision at a confidence score of about 0.7. Hence, this poses a sort of *threshold* in terms of confidence for the network in determining nodes, implying that it will always impose some uncertainty in node identification. This effect is not shown by YOLOv5x6, as shown in Fig. 7(a), which implies that denser models can overcome limitations in terms of confidence posed by smaller models. One last highlight can be taken from the analysis of both figures. Specifically, the F1-score is maximum when the confidence is about 0.4.

4.1.2. Improving test results via TTA and model ensembling

Results achieved by the models trained via transfer learning can be improved using two techniques, that is, *Test-Time Augmentation* (*TTA*) and *model ensembling*. As for TTA, it is the application of data-augmentation techniques at test time, which involves the creation of several slightly modified copies of each test image to be provided to the model for prediction; the final result will be an ensemble of these predictions. Instead, model ensembling is an example of *ensemble learning* (Kotu and Deshpande, 2019), a process where the predictions of several different models are aggregated, e.g. by consensus, to achieve a single final prediction.

Let us start by analyzing results achieved using an ensemble of YOLOv5x and YOLOv5x6, referred to as YOLOv5x+6. Results achieved in validation by the ensemble are shown in Table 3.

As it can be seen, using ensembling has a noticeable impact. To this end, let us compare the performance achieved by the ensemble model with the ones achieved by the one of the most top-performing single model, that is, YOLOv5x6. Specifically, the ensemble improves fruit detection by 3%, node detection by 7%, and flowers detection by 7%. As for mismatches, no significant improvements can be found for flowers; however, there is a significant reduction in false negatives for fruit and nodes, which are reduced by 4% and 6%, respectively. As for false positives, the ensemble model provides comparable results for fruits, with a noticeable improvement for both nodes and flowers.



(a) Precision vs. Confidence

(b) Recall vs. Confidence

(c) F1-score vs. Confidence

(d) Precision vs. Recall

Fig. 6. Precision, recall and F1 score achieved by the YOLOv5l6 architecture after 300 epochs of training.

Table 4
Results achieved on tomato recognition using YOLOv5+. If compared to the ensemble of YOLOv5x and YOLOv5x6, the overall results are improved of about 1% in terms of fruit detection, and 5% for node and flowers detection.

Class	TP	M	B-FP	B-FN
Fruit number	82%	4%	10%	13%
Nodes	80%	0%	55%	20%
Flowers	76%	2%	21%	35%

Table 5
Results achieved on tomato recognition using YOLOv5+TTA. If compared to the same model without Test Time Augmentation, the overall results are improved of about 1% for fruit detection, and 3% for nodes and flowers detection.

Class	TP	M	B-FP	B-FN
Fruit number	82%	7%	9%	11%
Nodes	83%	0%	58%	17%
Flowers	79%	3%	33%	18%

However, let us recall that the high values achieved for these classes are related either to incomplete labeling from domain experts or to visual artifacts related to overlapping.

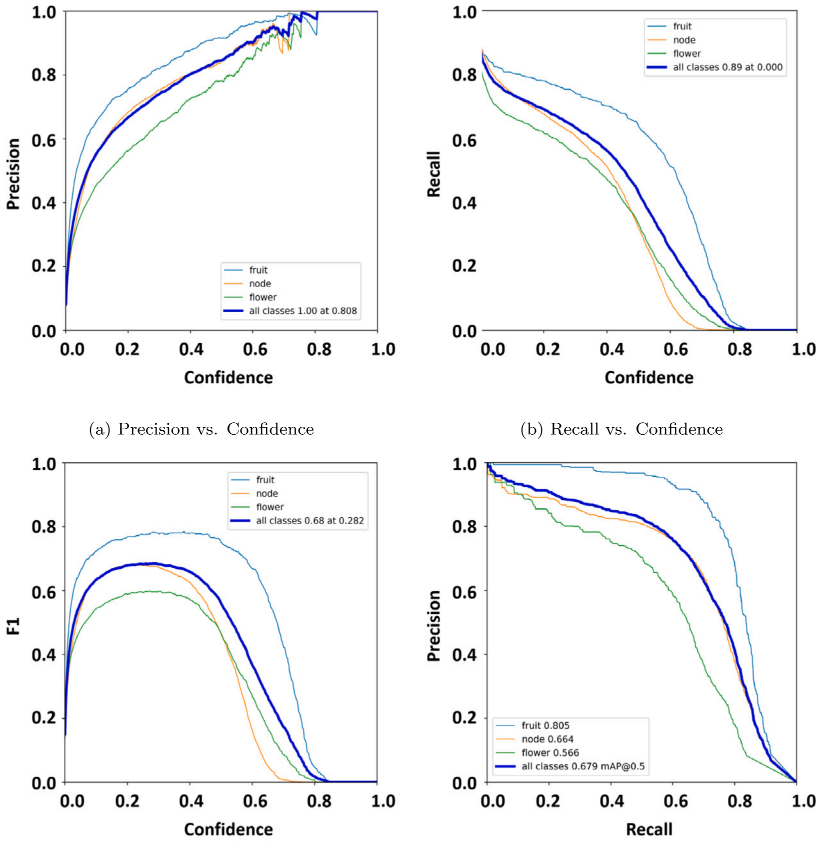
Results achieved by the ensemble in terms of precision, recall, and F1 score are shown in Fig. 8. As expected, precision greatly benefits from the use of ensembling. Furthermore, it can be noted that the maximum value for the F1-score is reached at around 0.5 confidence, meaning that the resulting model can detect items with overall improved accuracy.

Let us consider another ensemble, the one obtained from YOLOv5l, YOLOv5l6, YOLOv5x, and YOLOv5x6, which will be referred to as YOLOv5+, whose results are reported in Table 4. Furthermore, in Table 5, results achieved by YOLOv5+ with TTA are provided. Finally, in Fig. 9, the F1 scores for YOLOv5+ (Fig. 9(a)) and YOLOv5+TTA (Fig. 9(b)) are provided.

The comparison shows that YOLOv5+ achieves the best results if compared to all other models and ensembles. As for true positives, YOLOv5+ improves results in fruit detection by 1%, node and flower detection by 5% without TTA, and 8% with TTA. This improvement comes with a cost in terms of mismatches and false positives, which are slightly higher if compared to other models, especially with TTA. However, let us remark that false positives are mainly related to nodes identified on secondary stems, which are correct from a merely visual perspective because all the models effectively learned what a node is, regardless of its location on the image. Furthermore, the overall number of false negatives is greatly reduced, especially using YOLOv5+TTA, meaning that the model can deal with smaller bounding boxes and reduce the number of missing detections. As for the F1 scores, the trend already shown by the YOLOv5x+ ensemble is confirmed, and, especially for YOLOv5+TTA, it is safe to assume that the network can identify objects within the tomato plant.

4.1.3. *Using different backbones*

The last experiment on the dataset involves using different backbones to compare the impact of various network layers and configurations on achieved results. This experiment has been performed as a state-of-the-art comparison because it is not possible to directly feed images from the dataset to one of the models recapped in the related works avoiding reasonable bias in the results. Moreover, up to the knowledge of the authors, no other models are multi-class networks trained to identify, in addition to the fruits, also the flowers, and the nodes of tomato plants.



(a) Precision vs. Confidence

(b) Recall vs. Confidence

(c) F1-score vs. Confidence

(d) Precision vs. Recall

Fig. 7. Precision, recall and F1 score achieved by the YOLOv5x6 architecture after 300 epochs of training.

Table 6
Results achieved on tomato recognition using different backbones. It can be seen that VGG-16 outperforms the other models in each detection task.

Class	Backbone	TP	M	B-FP	B-FN
Fruit number	MobileNetV3S	64%	1%	7%	35%
	ResNet50V2	73%	3%	7%	24%
	VGG-16	75%	3%	7%	23%
Nodes	MobileNetV3S	50%	0%	48%	50%
	ResNet50V2	64%	0%	54%	36%
	VGG-16	69%	0%	59%	31%
Flowers	MobileNetV3S	48%	1%	45%	51%
	ResNet50V2	56%	1%	39%	44%
	VGG-16	59%	2%	34%	38%

For this experiment, three types of backbone have been selected, that is, MobileNetV3Small (Howard et al., 2017), ResNet50V2 (He et al., 2015), and VGG-16 (Simonyan and Zisserman, 2015). This selection is motivated by the need to compare the original YOLOv5 CSP-Darknet53 backbone with architectures with specific characteristics, such as the use of a deep stack of convolutional layers (VGG-16), a focus on residual layers (ResNet50V2), and a specific configuration for small, mobile applications (MobileNetV3Small). The results achieved using these backbones are shown in Table 6.

Interestingly, the three proposed backbones achieve results comparable with YOLOv5s (for MobileNetV3Small), YOLOv5m (for ResNet50V2), and YOLOv5x (for VGG-16). This suggests that the effectiveness of the network is more influenced by its overall number of parameters instead of specific types of layers, such as residual ones.

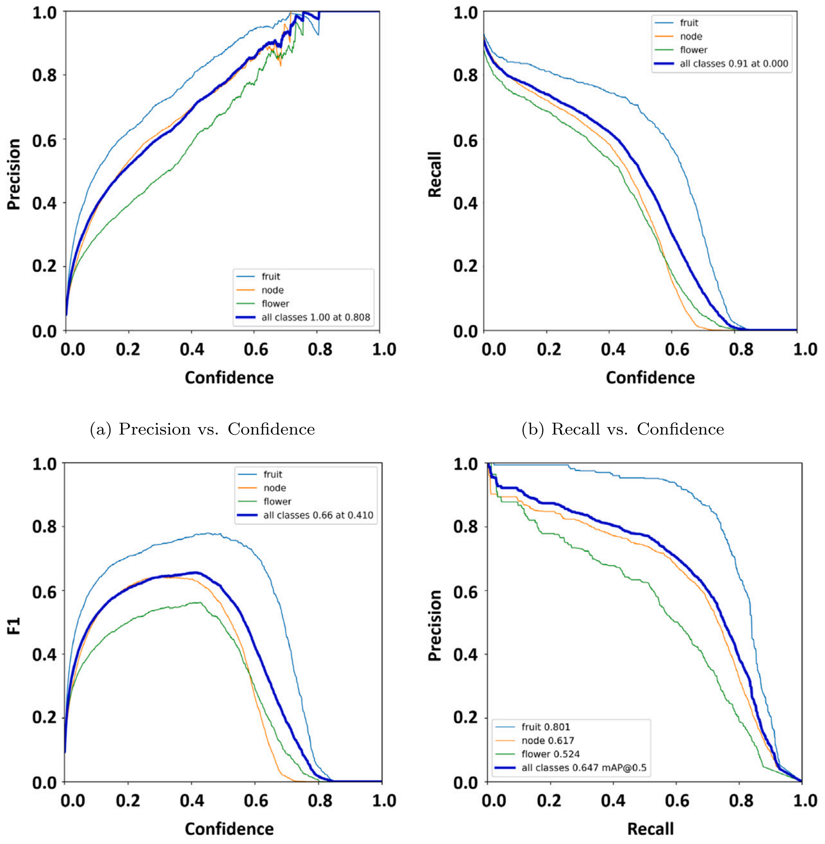
Let us also compare the F1 scores achieved by the different backbones at different confidence scores, as shown in 10.

The previous figure shows that the network can achieve a top value of F1 score of 0.6, which is comparable with the ones achieved by single models and ensembles but at a slightly higher level of confidence in the detection of the bounding boxes.

5. Conclusions and future works

By definition, Plant Phenomics is the science of measuring multiple phenotypic traits of plants at once. Hence, the use of deep CNN to detect and classify traits not easily extracted from images using basic imaging segmentation is of significant importance for plant science. The identification and counting the number of flowers and fruit of a plant can be used to measure fruit set for different plant varieties or the same variety treated with different nutritional products. With multiple images taken during plant growth the timing of flower development can be determined for a measure of plant stress. Similarly, the number of and the distance between leaf nodes on plants is a measure of plant development which can be used to evaluate nutritional products as well as plant varieties under different forms of stress, in particular abiotic stress. The number of nodes is directly related to the number of seed pods thus a measure of productivity (Feng et al., 2021).

Starting from the previous considerations, this paper has proposed a YOLOv5-based single-stage detectors aimed at identifying tomatoes, flowers, and nodes, both stand-alone and in an ensemble fashion. The models have learned how to identify and automatically extract significant phenotyping traits on images of tomato plants at different



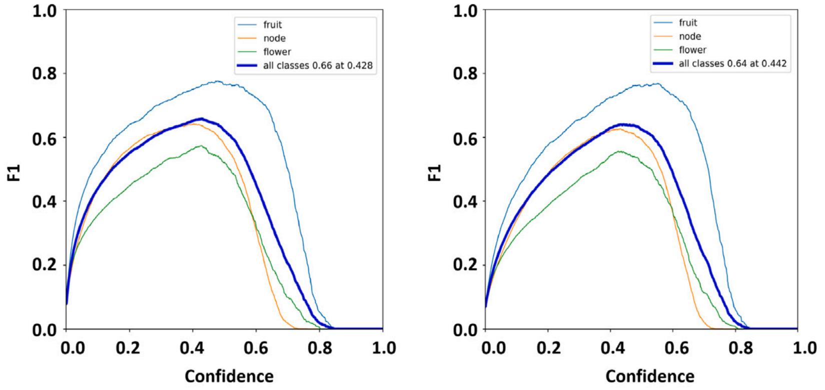
(a) Precision vs. Confidence

(b) Recall vs. Confidence

(c) F1-score vs. Confidence

(d) Precision vs. Recall

Fig. 8. Precision, recall and F1 score achieved by YOLOv5x+6.

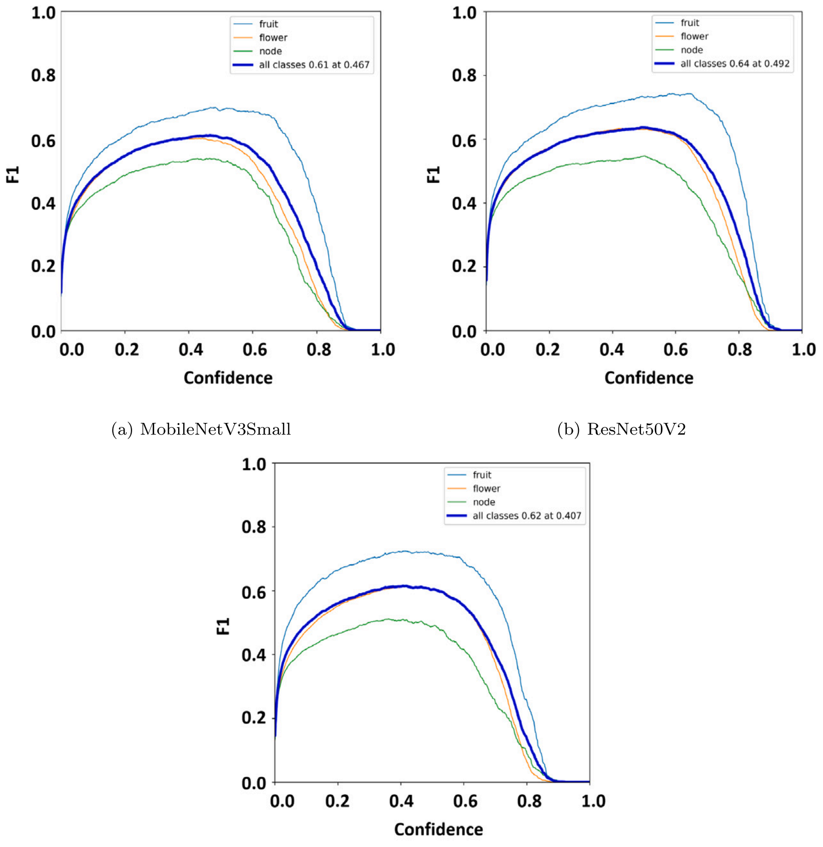


(a) F1-score - YOLOv5+

(b) F1-score - YOLOv5+TTA

Fig. 9. F1 scores achieved by YOLOv5+ and YOLOv5+TTA.

growth stages under various stress conditions. The YOLOv5-based models have achieved relatively high scores in terms of precision, recall, and F1-measure, considering the particular challenges of the input images. A remarkable result is the very low percentage of mismatches, indicating that the proposed models effectively learned a reasonable representation of the features useful to describe and classify the three classes. From a qualitative point of view, the experiments have shown that false positives are ascribable to some class missing instances in the labeled dataset, as for ancillary nodes, or effective similarity between misclassified objects, as for the flowers. Finally, false negative values can be reduced using denser models. For these reasons, future research directions will be devoted to identifying the best trade-off between model complexity and achieved performance, extending the experiments to other plant species when applicable (e.g., for the nodes). Furthermore, the model will be further improved using hyperparameter search, and other mechanisms, such as attention layers, will be tested and implemented. Finally, the dataset will be further extended and made available to other researchers to provide a framework for comparing achieved results and architectural improvements in single-stage object detectors.



(a) MobileNetV3Small

(b) ResNet50V2

(c) VGG-16

Fig. 10. F1 score achieved using different backbones.

CRediT authorship contribution statement

Angelo Cardellicchio: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Firozeh Solimani:** Data Curation, Writing – original draft, Writing – review & editing. **Giovanni Dimauro:** Validation, Writing – review & editing. **Angelo Petrozza:** Validation, Resources, Writing – review & editing. **Stephan Summerer:** Validation, Resources, Writing – review & editing. **Francesco Cellini:** Validation, Resources, Writing – review & editing. **Vito Renò:** Conceptualization, Methodology, Software, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

The activities described in this work are within the research projects *PHENO – Accordo di collaborazione tra ALSIA e CNR STIIMA – ref. prot. CNR STIIMA 3621/2020 and E-Crops – Technology for Sustainable Digital Agriculture*.

References

Arunachalam, A., Andreasson, H., 2021. Real-time plant phenomics under robotic farming setup: A vision-based platform for complex plant phenotyping tasks. Comput. Electr. Eng. 92, 107098. http://dx.doi.org/10.1016/j.compeleceng.2021.107098, URL: https://www.sciencedirect.com/science/article/pii/S0045790621001063.

Boogaard, F.P., Rongen, K.S.A.H., Kootstra, G.W., 2020. Robust node detection and tracking in fruit-vegetable crops using deep learning and multi-view imaging. Biosyst. Eng. 192, 117–132. http://dx.doi.org/10.1016/j.biosystemseng.2020.01.023, URL: https://www.sciencedirect.com/science/article/pii/S1537511020300350.

Fan, S., Liang, X., Huang, W., Jialong Zhang, V., Pang, Q., He, X., Li, L., Zhang, C., 2022. Real-time defects detection for apple sorting using NIR cameras with pruning-based YOLOV4 network. Comput. Electron. Agric. 193, 106715. http://dx.doi.org/10.1016/j.compag.2022.106715, URL: https://www.sciencedirect.com/science/article/pii/S0168169922000321.

Feng, Y.-Y., He, J., Turner, N.C., Siddique, K.H.M., Li, F.-M., 2021. Phosphorus supply increases internode length and leaf characteristics, and increases dry matter accumulation and seed yield in soybean under water deficit. Agronomy 11 (5), 930. http://dx.doi.org/10.3390/agronomy11050930, URL: https://www.mdpi.com/2073-4395/11/5/930. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

Fu, L., Yang, Z., Wu, F., Zou, X., Lin, J., Cao, Y., Duan, J., 2022. YOLO-banana: A lightweight neural network for rapid detection of banana bunches and stalks in the natural environment. Agronomy 12 (2), 391. http://dx.doi.org/10.3390/agronomy12020391, URL: https://www.mdpi.com/2073-4395/12/2/391. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Gai, R., Chen, N., Yuan, H., 2021. A detection algorithm for cherry fruits based on the improved YOLO-v4 model. Neural Comput. Appl. http://dx.doi.org/10.1007/s00521-021-06029-z.

Girshick, R., 2015. Fast R-CNN. pp. 1440–1448, URL: https://openaccess.thecvf.com/content_iccv_2015/html/Girshick_Fast_R-CNN_ICCV_2015_paper.html.

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep residual learning for image recognition. http://dx.doi.org/10.48550/arXiv.1512.03385, URL: http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

Howard, A.G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., Adam, H., 2017. MobileNets: Efficient convolutional neural networks for mobile vision applications. http://dx.doi.org/10.48550/arXiv.1704.04861, URL: http://arxiv.org/abs/1704.04861. arXiv:1704.04861 [cs].

Jocher, G., Nishimura, K., Mineeva, T., Vilariño, R., 2022. yolov5. URL: https://github.com/ultralytics/yolov5.

Kotu, V., Deshpande, B., 2019. Chapter 2 - data science process. In: Kotu, V., Deshpande, B. (Eds.), Data Science (Second Edition). Morgan Kaufmann, pp. 19–37. http://dx.doi.org/10.1016/B978-0-12-814761-0.00002-2, URL: https://www.sciencedirect.com/science/article/pii/B9780128147610000022.

Lawal, M.O., 2021a. Tomato detection based on modified YOLOv3 framework. Sci. Rep. 11 (1), 1447. http://dx.doi.org/10.1038/s41598-021-81216-5, URL: https://www.nature.com/articles/s41598-021-81216-5. Number: 1 Publisher: Nature Publishing Group.

Lawal, O.M., 2021b. Development of tomato detection model for robotic platform using deep learning. Multimedia Tools Appl. 80 (17), 26751–26772. http://dx.doi.org/10.1007/s11042-021-10933-w.

Li, X., Pan, J., Xie, F., Zeng, J., Li, Q., Huang, X., Liu, D., Wang, X., 2021. Fast and accurate green pepper detection in complex backgrounds via an improved Yolov4-tiny model. Comput. Electron. Agric. 191, 106503. http://dx.doi.org/10.1016/j.compag.2021.106503, URL: https://linkinghub.elsevier.com/retrieve/pii/S0168169921005202.

Litvin, A.G., Iersel, M.W.v., Malladi, A., 2016. Drought stress reduces stem elongation and alters gibberellin-related gene expression during vegetative growth of tomato. J. Am. Soc. Horticult. Sci. 141 (6), 591–597. http://dx.doi.org/10.21273/JASHS03913-16, URL: https://journals.ashs.org/jashs/view/journals/jashs/141/6/article-p591.xml. Publisher: American Society for Horticultural Science Section: Journal of the American Society for Horticultural Science.

Liu, X., Li, G., Chen, W., Liu, B., Chen, M., Lu, S., 2022. Detection of dense citrus fruits by combining coordinated attention and cross-scale connection with weighted feature fusion. Appl. Sci. 12 (13), 6600. http://dx.doi.org/10.3390/app12136600, URL: https://www.mdpi.com/2076-3417/12/13/6600. Number: 13 Publisher: Multidisciplinary Digital Publishing Institute.

Liu, G., Nouaze, J.C., Touko Mbouembe, P.L., Kim, J.H., 2020. YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3. Sensors 20 (7), 2145. http://dx.doi.org/10.3390/s20072145, URL: https://www.mdpi.com/1424-8220/20/7/2145. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

Liu, J., Wang, X., 2020. Tomato diseases and pests detection based on improved yolo V3 convolutional neural network. Front. Plant Sci. 11, URL: https://www.frontiersin.org/articles/10.3389/fpls.2020.00898.

Magalhães, S.A., Castro, L., Moreira, G., dos Santos, F.N., Cunha, M., Dias, J., Moreira, A.P., 2021. Evaluating the single-shot MultiBox detector and YOLO deep learning models for the detection of tomatoes in a greenhouse. Sensors 21 (10), 3569. http://dx.doi.org/10.3390/s21103569, URL: https://www.mdpi.com/1424-8220/21/10/3569. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Mesa, T., Polo, J., Arabia, A., Caselles, V., Munné-Bosch, S., 2022. Differential physiological response to heat and cold stress of tomato plants and its implication on fruit quality. J. Plant Physiol. 268, 153581. http://dx.doi.org/10.1016/j.jplph.2021.153581, URL: https://www.sciencedirect.com/science/article/pii/S0176161721002200.

Mirhaji, H., Soleymani, M., Asakereh, A., Abdanan Mehdizadeh, S., 2021. Fruit detection and load estimation of an orange orchard using the YOLO models through simple approaches in different imaging and illumination conditions. Comput. Electron. Agric. 191, 106533. http://dx.doi.org/10.1016/j.compag.2021.106533, URL: https://www.sciencedirect.com/science/article/pii/S0168169921005500.

Mu, Y., Chen, T.-S., Ninomiya, S., Guo, W., 2020. Intact detection of highly occluded immature tomatoes on plants using deep learning techniques. Sensors 20 (10), 2984. http://dx.doi.org/10.3390/s20102984, URL: https://www.mdpi.com/1424-8220/20/10/2984. Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

Panthee, D.R., Kressin, J.P., Piotrowski, A., 2018. Heritability of flower number and fruit set under heat stress in tomato. HortScience 53 (9), 1294–1299. http://dx.doi.org/10.21273/HORTSCI13317-18, URL: https://journals.ashs.org/hortsci/view/journals/hortsci/53/9/article-p1294.xml. Publisher: American Society for Horticultural Science Section: HortScience.

Qi, J., Liu, X., Liu, K., Xu, F., Guo, H., Tian, X., Li, M., Bao, Z., Li, Y., 2022. An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease. Comput. Electron. Agric. 194, 106780. http://dx.doi.org/10.1016/j.compag.2022.106780, URL: https://www.sciencedirect.com/science/article/pii/S0168169922000977.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection. pp. 779–788, URL: https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Redmon_You_Only_Look_CVPR_2016_paper.html.

Roy, A.M., Bhaduri, J., 2022. Real-time growth stage detection model for high degree of occultation using DenseNet-fused YOLOv4. Comput. Electron. Agric. 193, 106694. http://dx.doi.org/10.1016/j.compag.2022.106694, URL: https://www.sciencedirect.com/science/article/pii/S0168169922000114.

Ruparelia, S., Jethva, M., Gajjar, R., 2022. Real-time tomato detection, classification, and counting system using deep learning and embedded systems. In: Thakkar, F., Saha, G., Shahnaz, C., Hu, Y.-C. (Eds.), Proceedings of the International E-Conference on Intelligent Systems and Signal Processing. In: Advances in Intelligent Systems and Computing, Springer, Singapore, pp. 511–522. http://dx.doi.org/10.1007/978-981-16-2123-9_39.

Santos, T.T., de Souza, L.L., dos Santos, A.A., Avila, S., 2020. Grape detection, segmentation, and tracking using deep neural networks and three-dimensional association. Comput. Electron. Agric. 170, 105247. http://dx.doi.org/10.1016/j.compag.2020.105247, URL: https://www.sciencedirect.com/science/article/pii/S0168169919315765.

Sekachev, B., Manovich, N., Zhiltsov, M., Zhavoronkov, A., Kalinin, D., Hoff, B., Computer vision annotation tool. URL: https://github.com/openvinotoolkit/cvat.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. http://dx.doi.org/10.48550/arXiv.1409.1556, URL: http://arxiv.org/abs/1409.1556. arXiv:1409.1556 [cs].

Sozzi, M., Cantalamessa, S., Cogato, A., Kayad, A., Marinello, F., 2022. Automatic bunch detection in white grape varieties using YOLOv3, YOLOv4, and YOLOv5 deep learning algorithms. Agronomy 12 (2), 319. http://dx.doi.org/10.3390/agronomy12020319, URL: https://www.mdpi.com/2073-4395/12/2/319. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute.

Suo, R., Gao, F., Zhou, Z., Fu, L., Song, Z., Dhupia, J., Li, R., Cui, Y., 2021. Improved multi-classes kiwifruit detection in orchard to avoid collisions during robotic picking. Comput. Electron. Agric. 182, 106052. http://dx.doi.org/10.1016/j.compag.2021.106052, URL: https://www.sciencedirect.com/science/article/pii/S0168169921000703.

Tan, M., Pang, R., Le, Q.V., 2020. EfficientDet: Scalable and efficient object detection. http://dx.doi.org/10.48550/arXiv.1911.09070, URL: http://arxiv.org/abs/1911.09070. Number: arXiv:1911.09070 arXiv:1911.09070 [cs, eess].

Vasconez, J.P., Delpiano, J., Vougioukas, S., Auat Cheein, F., 2020. Comparison of convolutional neural networks in fruit detection and counting: A comprehensive evaluation. Comput. Electron. Agric. 173, 105348. http://dx.doi.org/10.1016/j.compag.2020.105348, URL: https://www.sciencedirect.com/science/article/pii/S016816991932232X.

Wang, D., He, D., 2021. Channel pruned YOLO v5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning. Biosyst. Eng. 210, 271–281. http://dx.doi.org/10.1016/j.biosystemseng.2021.08.015, URL: https://www.sciencedirect.com/science/article/pii/S1537511021001999.

Wang, X., Liu, J., 2021. Tomato anomalies detection in greenhouse scenarios based on YOLO-dense. Front. Plant Sci. 12, URL: https://www.frontiersin.org/articles/10.3389/fpls.2021.634103.

Wang, Y.-H., Su, W.-H., 2022. Convolutional neural networks in computer vision for grain crop phenotyping: A review. Agronomy 12 (11), 2659. http://dx.doi.org/10.3390/agronomy12112659, URL: https://www.mdpi.com/2073-4395/12/11/2659. Number: 11 Publisher: Multidisciplinary Digital Publishing Institute.

Yamamoto, K., Guo, W., Ninomiya, S., 2016. Node detection and internode length estimation of tomato seedlings based on image analysis and machine learning. Sensors 16 (7), 1044. http://dx.doi.org/10.3390/s16071044, URL: https://www.mdpi.com/1424-8220/16/7/1044. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

Yamamoto, K., Guo, W., Yoshioka, Y., Ninomiya, S., 2014. On plant detection of intact tomato fruits using image analysis and machine learning methods. Sensors 14 (7), 12191–12206. http://dx.doi.org/10.3390/s140712191, URL: https://www.mdpi.com/1424-8220/14/7/12191. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.

Yao, J., Qi, J., Zhang, J., Shao, H., Yang, J., Li, X., 2021. A real-time detection algorithm for kiwifruit defects based on YOLOv5. Electronics 10 (14), 1711. http://dx.doi.org/10.3390/electronics10141711, URL: https://www.mdpi.com/2079-9292/10/14/1711. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute.

Zhang, P., Li, D., 2022. EPSA-YOLO-v5s: A novel method for detecting the survival rate of rapeseed in a plant factory based on multiple guarantee mechanisms. Comput. Electron. Agric. 193, 106714. http://dx.doi.org/10.1016/j.compag.2022.106714, URL: https://linkinghub.elsevier.com/retrieve/pii/S016816992200031X.

Zheng, T., Jiang, M., Li, Y., Feng, M., 2022. Research on tomato detection in natural environment based on RC-YOLOv4. Comput. Electron. Agric. 198, 107029. http://dx.doi.org/10.1016/j.compag.2022.107029, URL: https://linkinghub.elsevier.com/retrieve/pii/S0168169922003465.