Expert Systems With Applications 257 (2024) 124972

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Vision foundation model for agricultural applications with efficient layer aggregation network

Jianxiong Ye [a,b,1], Zhenghong Yu [a,c,*,1], Jiewu Lin [a,b], Hongyuan Li [a], Lisheng Lin [b]

[a] College of Robotics, Guangdong Polytechnic of Science and Technology, Zhuhai 519090, China
[b] School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, China
[c] College of Engineering, South China Agricultural University, Guangzhou 510642, China

ARTICLE INFO

Dataset link: http://github.com/Ye-Sk/TasselELANet.git

Keywords:
Agricultural vision
Deep learning
Maize tassels
Wheat ears
Rice panels
Concise and efficient

ABSTRACT

Agricultural production is transitioning from traditional tools to IoT-connected automation devices. The integration of computer vision and agricultural automation is becoming closer with the rapid development of deep learning technology. Agricultural vision tasks, nevertheless, are relatively simple and less diverse compared to large-scale vision tasks, and there is a scarcity of plant data. This is also the reason why researchers need to make significant adjustments when directly applying advanced algorithms to the agricultural domain. In practice, downstream vision models should be designed to be concise and efficient. In this paper, we meticulously designed a deep convolutional network called TasselELANet for agricultural vision applications. It features a concise and efficient global architecture, including a 16-fold downsampling layer encoder and a decoder utilizing only 2 feature layers, which is greatly different from the general approaches. At its core is the efficient layer aggregation network, which utilizes the cross-stage fusion strategy to effectively optimize the gradient propagation path, thereby enhancing the learning ability and inference speed of the network. We validated the performance of TasselELANet on three challenging publicly available plant detection and counting datasets: maize tassels, wheat ears, and rice panels. The experimental tests achieved an average Accuracy of 0.899 and an average determination coefficient R^2 of 0.867, remarkably outperforming other advanced computer vision methods in terms of accuracy and efficiency, and demonstrating sufficient generalizability. We firmly believe that TasselELANet can serve as a powerful and reliable vision tool for agricultural practitioners while also providing a solid foundation for future research. The code, datasets can be accessed at https://github.com/Ye-Sk/TasselELANet.git.

1. Introduction

Traditional agricultural tools and machinery in agricultural production activities are gradually being replaced by IoT-connected automation devices (Padmavathi, BhagyaLakshmi, Vishnupriya, & Datchanamoorthy, 2024). Through information-physical systems, sensors, embedded terminal systems, smart control systems, and communication facilities form an intelligent network system, enabling comprehensive perception of environmental information for planting and breeding, real-time monitoring of individual behavior, and tracking the working status of agricultural equipment (Gajjar, Gajjar, Thakor, Patel, & Ruparelia, 2021).

Computer vision technology plays a crucial role in various fields, including agriculture (Li, Bao, & Qi, 2022), ecology (Gage, Miller, Spalding, et al., 2017), and plant science (Ye & Yu, 2024). As agricultural automation becomes increasingly prominent, vision application, as a key area of interdisciplinary research between computer vision and plant science, provides agricultural practitioners with a non-invasive, efficient, and reproducible method to understand and monitor plant distribution and quantity.

Vision is a fundamental task in the agricultural domain (Xu et al., 2022), providing essential prior information to practitioners to accomplish various related tasks, such as predicting plant growth stages (Jurado-Ruiz et al., 2024), machine harvesting (Lawal & Zhao, 2021), and estimating plant quantity (Yu, Ye, Li, Zhou, & Li, 2023). With the continuous advancement of computer vision and machine learning technology, practitioners can automate agricultural vision applications through image processing and analysis algorithms.

However, due to the non-rigid nature of plants and numerous vision challenges, achieving reliable automation remains a challenging and

* Corresponding author at: College of Robotics, Guangdong Polytechnic of Science and Technology, Zhuhai 519090, China.
E-mail address: honger1983@gmail.com (Z. Yu).
[1] These authors contributed equally to this work and are co-first authors.

https://doi.org/10.1016/j.eswa.2024.124972
Received 2 February 2024; Received in revised form 5 July 2024; Accepted 1 August 2024
Available online 10 August 2024

limiting task. This can be largely attributed to the influence of both internal and external plant growth environments on recognition:

1. The complexity of the plant growth environment interferes with image acquisition and analysis, including factors like lighting, shadows, rainwater, and wind.
2. Variations in size, shape, color, and texture among different plants increase the difficulty of plant recognition.
3. Self-occlusion and occlusion between plants can lead to errors in the recognition process.
4. Low-quality images or incorrect annotation data can result in training failures or inaccurate predictions of the model.
5. Plants often exhibit uneven distribution, requiring the technology to effectively recognize different vegetation densities.
6. In certain application scenarios, such as agricultural production, there is a high demand for real-time monitoring of plant status and quantity.

Benefiting from the rapid development of the third wave of artificial intelligence—deep learning, new solutions have emerged to address the challenges faced by agricultural vision applications. These solutions are primarily driven by the success of deep convolutional neural networks (CNNs) (Chai, Nie, Zhou, & Zhou, 2024; Duan, Luo, & Zhang, 2024; He, Zhang, Ren, & Sun, 2015). In recent years, research on agricultural vision applications can be categorized into three main areas: plant counting, plant segmentation, and plant detection.

Regarding plant counting research, Lu et al. (2022) proposed a local counting framework called TasselNetV3, which introduced an upsampling operator to supervise the reassignment of counts, thereby improving the visualization of outputs. Soon after, Li, Wang, Qiao, et al. (2023) developed a lightweight deep regression network called RapeNet for dense rape flower cluster counting, combining the advantages of Bayesian loss function and mainstream regression estimation methods. On the other hand, Bai et al. (2023) designed a deep network named RPNet to enhance rice plant counting performance. RPNet utilizes dense exploitation of shallow and deep features, enabling accurate counting in high-throughput images captured by unmanned aerial vehicles (UAV).

In the field of plant phenotyping research, researchers use segmentation methods to accurately segment plants or organs in image sequences. To facilitate precise crop breeding and management, Yu, Yin, Nie, Ming, et al. (2022) dynamically monitor the maize tassel area using the U-Net segmentation model and high-resolution images. Depending on the specific scenario, either Vgg16 or MobileNet is selected as the backbone network to strike a balance between accuracy and speed. Considering the use of U-Net convolutional architecture for peak identification, Misra et al. (2020) propose a non-destructive method named SpikeSegNet for monitoring and counting spikes in wheat plants. Furthermore, Ye, Yang et al. (2023) analyze the performance of the Mask R-CNN model in cabbage plants based on UAV images, including feature extraction and quantity estimation. According to their reported results, this method shows promise in providing technical references for monitoring crop growth information in artificially cultivated fields.

Compared to counting methods, segmentation and detection methods provide known location information, making plant counting an easily achievable additional task. Under the need for acquiring plant location information, the implementation cost of detection methods is significantly lower than segmentation methods. Moreover, detection methods are equally important and beneficial for segmentation tasks. Detection can serve as prior knowledge and constraints for segmentation algorithms, aiding in more accurate localization of object boundaries. By detecting the position of objects, segmentation algorithms only need to perform operations within the detected object regions, avoiding pixel-level computations on the entire image (He, Gkioxari, Dollár, & Girshick, 2017; Woo et al., 2023). Existing research has shown that detection methods offer better upward and downward compatibility (Lin et al., 2017; Ye, Yu, Wang, Lu, & Zhou, 2024) and can be used as useful auxiliary clues. The extension and application of advanced detection algorithms to agricultural systems are of great interest to current researchers. It has been demonstrated to be a promising approach for various practical applications in monitoring and controlling plant growth, including anomaly detection (Pei et al., 2022), disease detection (Gómez-Flores, Garza-Saldaña, & Varela-Fuentes, 2019), pest monitoring (Wen et al., 2022), and crop yield estimation (Chlingaryan, Sukkarieh, & Whelan, 2018), among others.

Despite achieving success, we have also identified some noteworthy issues. Often, when researchers introduce advanced vision technologies into agricultural scenarios, they tend to focus only on narrow applications at the vision representation level. A good study should demonstrate broader continuity, profound depth, and greater generality. In other words, models suitable for large-scale high-level vision tasks may not exhibit good adaptability in agricultural scenarios, mainly due to the relative simplicity and limited diversity of agricultural vision tasks. What is more, the relatively high cost of collecting plant image data results in a limited availability of data. Faced with these constraints, complex vision models are difficult to train and lack sufficient robustness. This is also why researchers need to make notable adjustments when directly applying advanced algorithms to the agricultural field. In reality, the design starting point for vision models in downstream tasks should be concise and efficient.

Current CNN architecture designs mostly follow the common perception in deep networks (He et al., 2015), which involves extracting low-level features from shallow layers and high-level features from deep layers. After being processed by the backbone network, the size of the feature map is usually reduced to 1/32 of the original image, which is expected to capture larger plant receptive fields. The mainstream feature extraction methods include feature pyramid network (FPN) (Lin et al., 2017) or path aggregation network (PAN) (Liu, Qi, Qin, Shi, & Jia, 2018). Generally, 4–5 feature layers are used, which is often employed to mitigate the impact of appearance changes in plants during different growth stages or variations in external lighting conditions. Encouragingly, as stronger backbone networks are designed, these initial design intentions can be overcome by the powerful representation capacity of CNN.

In recent study on detection and counting maize tassels, we proposed a novel deep convolutional network called TasselLFANet (Yu et al., 2023), which integrates advanced techniques from machine learning pioneers. Carefully designed, we constructed a very concise architecture comprising an encoder with a 16-fold downsampling layer and a decoder utilizing only 2 feature layers, allowing for optimization and diagnostics of specific modules. During practical applications, when aiming to develop TasselLFANet into a general learning framework, we encountered some model bottlenecks. Like traditional methods, TasselLFANet still relied on predefined anchors for plant predictions, limiting the model's robustness and further performance improvement. To overcome this issue, we introduced the state-of-the-art computer vision concept of predicting anchors based on pixel values (Duan et al., 2019), along with re-adjusted loss functions. Apart from that, through analyzing the computational bottlenecks, we noticed that a considerable amount of computations concentrated on the efficient layer aggregation network (ELAN) module (Wang, Liao and Yeh, 2023). Inspired by the compound scaling strategy in neural network design (Tan & Le, 2019; Wang, Liao and Yeh, 2023), for a concatenation-based model, when the depth factor is scaled, the width factor is also adjusted. We optimized ELAN when no performance degradation was observed. Overall, these adjustments markedly reduced the model parameters, greatly improved operational efficiency, and enhanced the model's robustness. Although we retained TasselLFANet's global architecture, the core of the entire network lies in the efficient layer aggregation network, and thus, we named the new model TasselELANet. TasselELANet inherits the advantages of TasselLFANet while addressing its limitations, resulting in superior performance and efficiency in plant recognition tasks.

To demonstrate the model's generality, we validated TasselELANet's performance on three challenging publicly available plant detection and counting datasets: maize tassels, wheat ears, and rice panels. Extensive experimental results show that TasselELANet's accuracy and efficiency are pronouncedly superior to other state-of-the-art computer vision methods. Our ultimate goal is to provide agricultural practitioners with a powerful and reliable vision tool, enabling them to efficiently tackle real-world problems. Concurrently, we hope to offer valuable vision architecture references for future researchers. In line with this, we have made our implementation available online. In summary, our contributions are as follows:

1. TasselELANet: An efficient agricultural vision tool that can serve as a solid foundation for future research and applications.
2. A concise and efficient global architecture, comprising an encoder with a 16-fold downsampling layer and a decoder utilizing only 2 feature layers.
3. The efficient layer aggregation network, making the model lightweight while maintaining excellent performance and robustness.
4. Reporting state-of-the-art performance on three publicly available plant datasets.

2. Materials and methods

2.1. Plant datasets

To better generalize our approach, we conducted experiments on three publicly available plant datasets with corresponding annotation boxes. Here, we provide a brief introduction to the characteristics and challenges of these datasets.

The Maize Tassels Detection and Counting (MTDC) (Zou, Lu, Li, Li, & Zhang, 2020) dataset was collected from four experimental fields in China, covering six maize varieties. The dataset comprises 186 and 175 images for training and testing, respectively, with image resolutions of 3648 $\times$2736, 4272$\times$ 2848, and 3456 $\times$ 2304. It is worth noting that the test set was intentionally designed to be entirely different sequences by the researchers, resulting in remarkable variations in data distribution. This poses a substantial challenge for the model's domain adaptation. Such dataset characteristics demand the model to possess strong generalization capabilities for practical applications, to adapt to various scenes and conditions.

The Wheat Ears Detection Update (WEDU) (Lu, Ye, Wang, & Yu, 2023) dataset is an updated version of the WED dataset initially proposed by Madec et al. (2019). It contains 165 and 71 images for training and testing, respectively, with an image resolution of 6000 $\times$ 4000. However, inconsistencies between annotation labels and images in the initial dataset have hindered researchers from keeping up with global research progress. In our previous research work, we developed a neural network for wheat ear recognition and used the results to update the annotation boxes in the WED dataset (Ye, Yu, Wang et al., 2023). Despite our efforts, the accuracy of the annotation boxes is still limited by the model's performance, which prevents the complete elimination of possible noise. Compared to other manually curated datasets, the WEDU dataset poses significant challenges.

The Diverse Rice Panicle Detection (DRPD) (Teng et al., 2023) dataset was collected from experimental fields in multiple geographical regions (China, the United States, and Japan) and includes 229 rice varieties. Aerial images were captured from rice fields at three different altitudes (7 m, 12 m, and 20 m), resulting in corresponding image sets. In this study, we selected the most challenging images captured at an altitude of 20 m for testing. This subset of the dataset contains 224 images for training and 334 images for testing. It is worth noting that the aerial images were cropped by researchers into 512 $\times$ 512 pixel sizes and have high density, posing challenges for the model in handling low-resolution images and dense predictions. Example images from the three plant datasets are shown in Fig. 1.

2.2. TasselELANet

Here, we will first introduce the core component of TasselELANet, which is ELAN. Next, we will elaborate on its global architecture and various design details. Alongside this, we will provide explanations for the rationale behind these design choices.

2.2.1. Efficient layer aggregation network

By controlling the shortest and longest gradient paths, a deeper network can effectively learn and converge. The ELAN module achieves this by parallelizing more gradient flow branches, obtaining richer gradient information, and thus achieving higher precision and more reasonable latency. As shown in Fig. 2(a), ELAN is characterized by local connections, reassembling feature representations, and eliminating redundant information. Its implementation is straightforward and efficient: firstly, the input feature map, x, is split into two parts, with one part passing through the computational block, and the other part directly traversing the entire stage. The output, denoted as $y = F(x) \otimes x$, is obtained by integrating the two parts, while $F(x)$ represents the stack in computational block operation. The computational block consists of a series of convolutional, batch normalization, and activation functions, and by truncating the repeated gradient flow of these blocks, it can make the information learned in adjacent stages more diverse. In TasselLFANet's ELAN (Yu et al., 2023), a certain level of stability can be achieved regardless of the length of the gradient path and the number of computational block stacks. Nonetheless, if more computational blocks are infinitely stacked, this stable state may be disrupted, leading to a decrease in parameter utilization. In practice, as depicted in Fig. 2(b) for concatenation-based models, when the number of computational block stacks is scaled, the overall ratio of input channels and output channels changes, potentially causing a reduction in hardware utilization. Ma, Zhang, Zheng, and Sun (2018) reported that excessive fragmentation operations are not friendly to parallel acceleration. To reduce feature redundancy and improve computational efficiency in TasselELANet, we control the number of computational block stacks in ELAN to be two.

2.2.2. Global architecture

The global architecture of TasselELANet is illustrated in Fig. 3, consisting of an encoder for generating the feature set, a decoder for feature refine, and a detector for visual output.

(1) *Encoder*: Given an RGB image of size $x \in R^{h\times w\times 3}$ as input, the encoder defines a transformation function E:

$$R^{h\times w\times 3} \longrightarrow R^{(h/r)\times(w/r)\times c} \tag{1}$$

Eq. (1) used to encode x and generate feature map X, where c represents the number of feature channels, and r is the downsampling rate of the encoder. In particular, each downsampling stage in the encoder reduces the height and width of the input feature map by a factor of $1/r$. Furthermore, the feature maps from the 2nd to the 4th layers will undergo the feature remapping operation of ELAN, denoted as R:

$$R^{(h/r)\times(w/r)\times c} \longrightarrow R^{(h/r)\times(w/r)\times c'} \tag{2}$$

With three channel numbers set as $c'_1 = 64$, $c'_2 = 128$, and $c'_3 = 256$. As a result, the feature maps X' after ELAN will have their height and width reduced by a factor of r, and their channel numbers will be c'_1, c'_2, and c'_3, respectively.

(2) *Decoder*: Given the output feature maps $y \in R^{(h/16)\times(w/16)\times 256}$ from the encoder, the decoder further defines a transformation function D:

$$R^{(h/16)\times(w/16)\times 256} \longrightarrow R^{(h/8)\times(w/8)\times 128} \tag{3}$$

Specifically, the operations in the decoder include processing the feature maps y through the spatial pyramid pooling fast (SPPF) (He, Zhang, Ren, & Sun, 2014) for context encoding. After that, the ELAN feature remapping is applied to obtain the output layer. Subsequently,

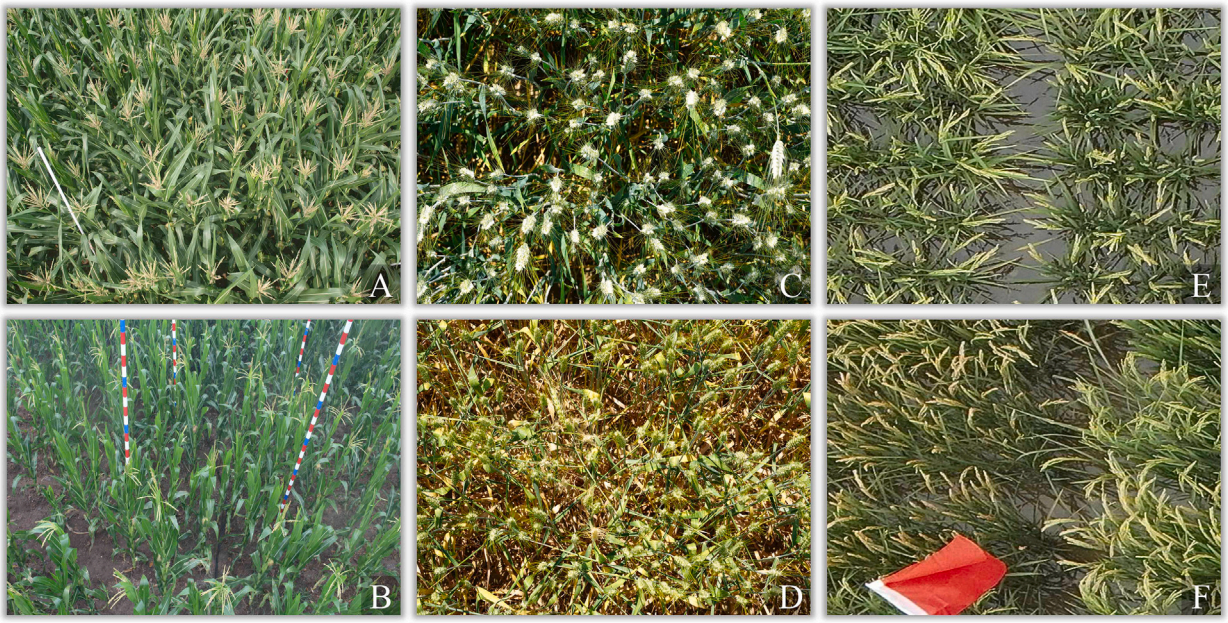

Fig. 1. Example images on three plant datasets. Panels (A, B) are from the maize tassels detection and counting (MTDC) dataset, panels (C, D) are from the wheat ears detection update (WEDU) dataset, and panels (E, F) are from the diverse rice panicle detection (DRPD) dataset.

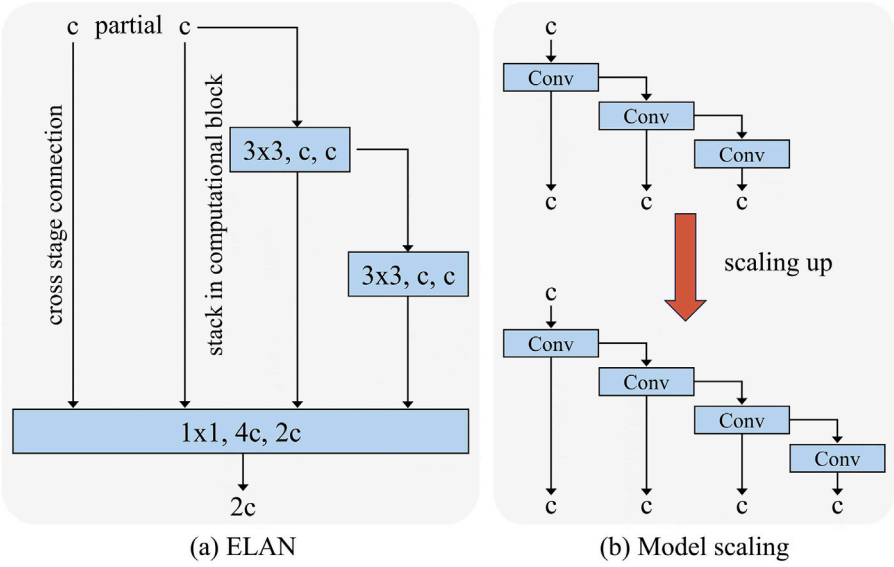

Fig. 2. (a) Efficient layer aggregation networks. (b) Model scaling for concatenation-based models.

we merge the feature maps $X_3 \in R^{(h/8)\times(w/8)\times 128}$ reduced by 1/8 from the encoder to counteract scale and perspective transformations in the image. Next, we use the lightweight data-related upsampling operator content-aware reassembly of features (CARAFE) (Wang et al., 2019) on the decoder's output layer to incorporate cascade information in the spatial dimension, resulting in an output feature map of the same size as X_3. Following the CARAFE operation, we apply a multi-efficient channel attention (Mlt-ECA) (Yu et al., 2023) to adaptively adjust the feature responses. The feature maps after the concat operation serve as the second output layer of the decoder. Finally, both output layers undergo reparameterization conv (Ding et al., 2021) to enhance the non-linearly transformed output feature maps $Y_1 \in R^{(h/16)\times(w/16)\times 256}$ and $Y_2 \in R^{(h/8)\times(w/8)\times 128}$.

(3) *Detector*: Given the output feature map, the main task of the detector is to merge sub-image detection results from different stages of feature maps and fuse the encoded information back into the original feature map. It adopts pixel-level prediction to determine the object's position by regressing the distances between each anchor and the four edges of the object bounding box. The detector operates in steps such as feature extraction, plant detection, feature fusion, and network layer fusion to accurately detect objects in the input image and output their position and class information. Finally, Non-maximum suppression (NMS) is applied to filter the generated prediction boxes. The intersection over union (IoU) metric is used to measure the overlap between two prediction boxes. By comparing the IoU values between prediction boxes, it determines whether they belong to the same object, thereby eliminating redundant detection results.

In conclusion, the aforementioned transformations are modeled by a CNN. The encoder of TasselELANet is implemented using a simple network, represented as follows:

$$2 \cdot c_3^2(32, 64) \rightarrow E(64) \rightarrow M_2^2 \rightarrow E(128) \rightarrow M_2^2 \rightarrow E(256) \quad (4)$$

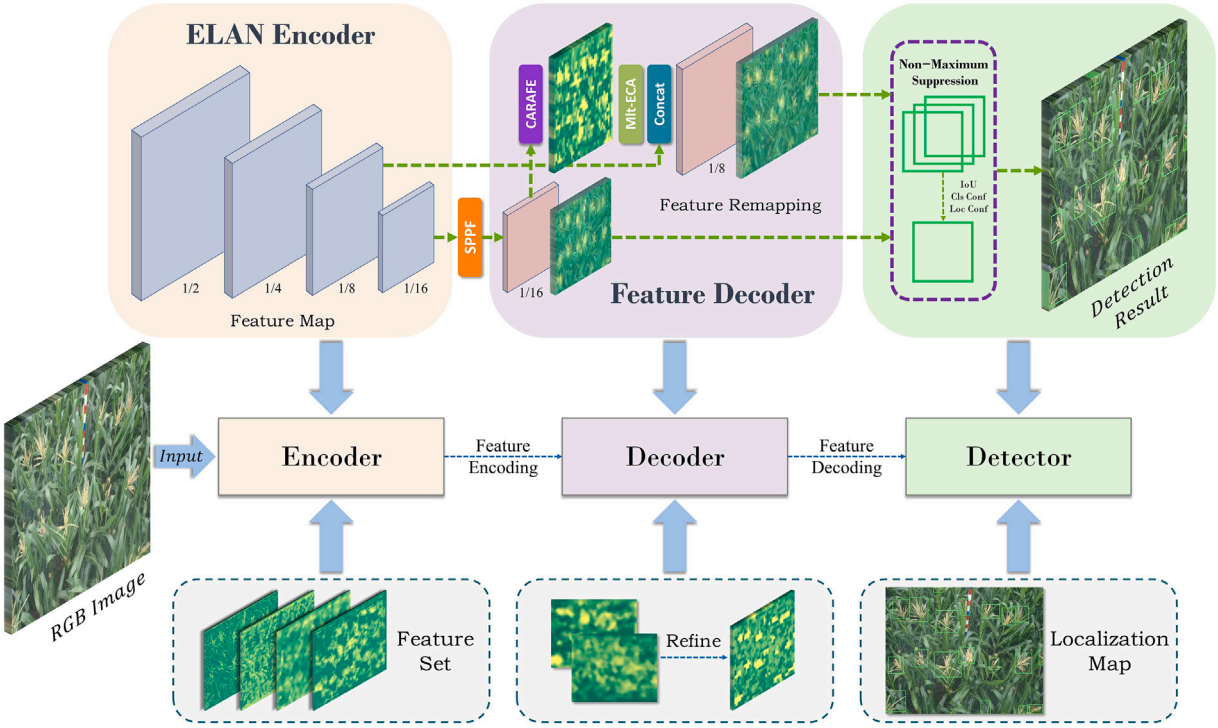

Fig. 3. The global architecture of TasselELANet consists of encoder, decoder, and detector components.

$n \cdot c_k^s(c_0, c_n)$ consists of n consecutive convolutional layers, each parameterized by a s-stride $k \times k$ kernel and outputting channels ranging from c_0 to c_n. After each convolutional layer, batch normalization (BN) and gaussian error linear units (GELU) are applied. $E(c)$ is a ELAN module with c output channels, and M_k^s is a max pooling layer with a s-stride $k \times k$ kernel. According to this definition, the decoder can directly utilize the final feature layer output by the encoder, which is the feature map y mapped to a new tensor space $R^{(h/16)\times(w/16)\times 256}$ by $E(256)$. Another output layer is obtained by mapping y back to the tensor space $R^{(h/8)\times(w/8)\times 128}$ using the CARAFE upsampling operator to combine feed-forward features. Beyond, each output layer also applies a structural reparameterization module, consistent with TasselLFANet. During training, the network maps features through multiple gradient flow paths to capture different receptive fields, enriching feature information. During inference, an Op fusion strategy is employed to transform all network layers into 3 × 3 convolutions for improved computational efficiency (Yu et al., 2023).

As a whole, TasselELANet features a highly streamlined global architecture, comprising an encoder with a 16-fold downsampling layer and a decoder utilizing only 2 feature layers, along with an efficient post-processing detector. This design inspiration comes from an in-depth observation of modern neural networks. To reduce feature redundancy and enhance computational efficiency, we first limit the utilization of encoder feature layers to 2. When restoring the feature map size with a single upsampling operation, we move back one downsampling layer to maintain semantic information continuity. This design also restores a more compact global architecture. In practice, the 16-fold downsampling layer encoder retains more spatial information in the feature map, facilitating precise localization and capturing fine-grained object features. In contradistinction, the 32-fold downsampling layer encoder provides a larger receptive field, suitable for capturing broader scene information and detecting/classifying larger objects. Another notable difference is that the latter is faster and more efficient, while the former has lower model capacity. In subsequent experiments, we further validate that the 16-fold downsampling layer encoder demonstrates superior comprehensive performance in agricultural vision tasks under other unchanged conditions.

The above provides an overview of the core architecture of TasselELANet, and more detailed parameter configurations can be found in our published code.

2.3. Loss function

Loss function is the key component in optimizing deep learning models, carrying the interaction information between data and the model. Its design is crucial for the convergence and performance of the model. In TasselELANet, the loss function is aimed at jointly optimizing the classification accuracy and bounding box localization in plant detection tasks. It consists of the following two parts.

2.3.1. Classification loss

Classification loss is used to measure the model's accuracy in classifying different categories. TasselELANet employs the cross-entropy loss function to calculate the loss between predicted categories and object categories. It is a common binary classification loss function used to assess the learning disparity between positive and negative samples. Given the object values (label values) as $g \in \{0, 1\}^n$ and the predicted results as $y \in R^n$, the classification loss is defined as:

$$L_{cls} = -\frac{1}{n}\sum_{i=1}^{n}(g_i log(p_i) + (1 - g_i) log(1 - p_i)) \tag{5}$$

where n is the batch size, and p represents the probability values of the predicted results y after being converted by the sigmoid function:

$$p = \frac{1}{1 + exp(-y)} \tag{6}$$

2.3.2. Regression loss

Regression loss consists of two components: distribution focal loss (DFL) (Li et al., 2020) and scylla-IoU (SIoU) (Gevorgyan, 2022), both of which are used to optimize the plant detection model's bounding box position prediction and morphology learning. SIoU guides the model in learning the degree of bounding box matching by considering the scale and angle information of the bounding box. Assuming the width

DO WE CODE THIS MATHS IN OR FIND A TOOL THAT DOES IT?

and height of the bounding box are represented as w_b and h_b, SIoU is described as:

$$L_{siou} = IoU - ((s_{cw}/w_b)^2 + (s_{ch}/h_b)^2) \tag{7}$$

s_{cw} and s_{ch} respectively represent the scale information of the bounding box center point in the horizontal and vertical directions. DFL, on the flip side, optimizes the bounding box position using the smooth L1 loss. For each positive sample i, it is defined as:

$$dfl_i = Sml(pd_i, gt_i) \times w_i \tag{8}$$

where $Sml(pd_i, gt_i)$ represents the smooth L1 loss for the ith positive sample, and w_i is the weight associated with the ith sample. Let N_{pos} denote the number of positive samples, then the final DFL loss is:

$$L_{dfl} = \frac{1}{N_{pos}} \sum_{i=1}^{N_{pos}} dfl_i \tag{9}$$

By combining the classification loss and regression loss, we obtain the total loss for TasselELANet, denoted as $L_{elan} = \alpha L_{cls} + \beta L_{reg}$, where $L_{reg} = L_{siou} + L_{dfl}$.

3. Results and discussions

3.1. Implementation details

To ensure the objectivity and reliability of the results, we employed the same configuration for both training and testing in our experiments. For data partitioning on the three plant datasets, we followed the method described in Section 2.1. As for parameter settings, we initialized the weights with pre-trained weights on the COCO dataset. To reduce computational burden, we scaled the longest side of the input images to 608 pixels and proportionally scaled the other side to maintain the original aspect ratio of the images. During the training process, we used stochastic gradient descent as the optimizer with a momentum factor of 0.937. The initial learning rate was set to 0.01, and we performed 300 epochs of iterative optimization with a batch size of 4. Considering the small size of the training data, we applied data augmentation techniques such as color distortion, random scaling transformation, and mosaic data augmentation to avoid overfitting. It should be emphasized that the training parameters for other models used in this research followed their default settings and were not adjusted.

3.2. Evaluation metrics

The performance of plant detection is evaluated using two metrics: average precision at 50% IoU (AP_{50}) and average precision at 50%–95% IoU (AP_{50-95}). These metrics provide more accurate measures of the model's localization performance. AP is defined as follows:

$$AP = \int_0^1 P_r(R_e) d(R_e) \tag{10}$$

P_r represents the proportion of correctly predicted objects among all the predicted objects by the model, and R_e represents the proportion of correctly predicted objects among all the true objects in the dataset. The evaluation metrics for plant counting are as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |G_i - P_i| \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (G_i - P_i)^2} \tag{12}$$

where n is the number of images, G_i and P_i represent the predicted count and ground-truth count, respectively, for the ith image. Mean absolute error (MAE) quantifies the accuracy of the model, while root mean square error (RMSE) quantifies the robustness of the model. The lower these two metrics, the better the counting performance.

Table 1
Quantitative results of MTDC dataset.

Method	AP_{50}	AP_{50-95}	MAE	RMSE
TasselNetV3	–	–	4.0	6.9
CSRNet	–	–	6.9	11.5
Faster R-CNN	0.694	0.270	7.89	10.1
CenterNet	0.746	0.328	4.60	6.66
Yolov7-tiny	0.835	0.383	6.33	14.08
TasselLFANet	0.858	0.424	5.57	11.55
TasselELANet	**0.865**	**0.463**	**3.99**	**6.39**

The best performance is in boldface.

Table 2
Quantitative results of WEDU and DRPD datasets.

Method	WEDU dataset				DRPD dataset			
	AP_{50}	AP_{50-95}	MAE	RMSE	AP_{50}	AP_{50-95}	MAE	RMSE
Faster R-CNN	0.521	0.199	28.90	36.11	0.462	0.183	31.27	45.05
CenterNet	0.653	0.375	17.52	24.04	0.673	0.203	29.10	34.67
Yolov7-tiny	0.896	0.456	12.38	19.13	0.839	0.434	23.93	29.18
TasselLFANet	0.926	0.515	8.82	12.33	**0.854**	0.464	20.79	27.37
TasselELANet	**0.931**	**0.547**	**7.03**	**9.03**	0.848	**0.497**	**18.69**	**25.69**

The best performance is in boldface.

3.3. Comparison with state of the art

To validate the superiority of TasselELANet, we compared it with several state-of-the-art vision methods, including Faster R-CNN (Ren, He, Girshick, & Sun, 2017), CenterNet (Duan et al., 2019), Yolov7-tiny (Wang, Bochkovskiy and Liao, 2023), and TasselLFANet (Yu et al., 2023), all of which are applicable to agricultural vision tasks in images. Furthermore, we referenced other published results (MTDC: TasselNetV3 (Lu et al., 2022); CSRNet (Li, Zhang, & Chen, 2018)) in appropriate cases. The quantitative results for the three plant datasets are presented in Tables 1–2, and the qualitative results are shown in Figs. 4–6. Our analysis is as follows:

The MTDC dataset: TasselELANet reports state-of-the-art performance in this case. Compared to the other two datasets, the MTDC dataset exhibits a wide variation in the size of maize tassels. Traditionally, researchers often adopt a strategy of fusing multi-scale features to handle such situations, but this approach often relies on the direct application of prior knowledge. Yet still, our research indicates that by designing a more powerful backbone network, the utilization of feature layers can become concise and efficient, and these initial design intentions can be overcome by the strong representation capabilities of CNNs. As shown in Fig. 4, TasselELANet performs comparably to other methods. Another aspect that needs attention is the significant difference in data distribution between the training and testing sets in the MTDC dataset. This impact can lead to a lack of generalization and a blind performance of the model, especially evident in counting tasks (see Table 1). This is also the main reason for the notable variations in counting metrics among different methods, while the detection task primarily focuses on the model's accuracy and recall in object localization and classification. It is worth mentioning that even when compared to advanced counting methods, our model's performance remains superior. Unfortunately, these counting methods are not capable of precise plant localization.

The WEDU and DRPD datasets: TasselELANet continues to demonstrate superior performance. The wheat ears and rice panicles in the WEDU and DRPD datasets have uniform sizes, with the main difference being that the former is based on high-resolution imaging while the latter is the opposite. Although the WEDU dataset allows the model to observe more detailed information, the presence of noise inevitably hinders effective training and accurate predictions for the models, as observed in Faster R-CNN and CenterNet in Table 2, which are highly sensitive to such anomalies. Despite reporting higher performance, we found that all methods exhibit noticeable false detections when

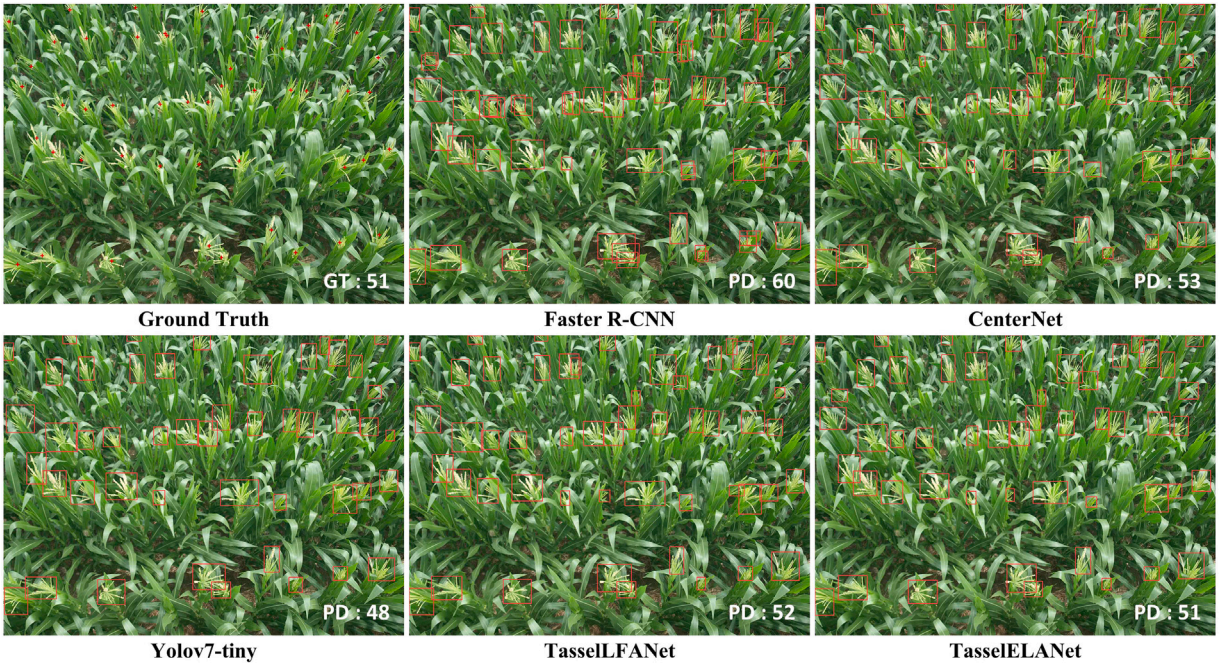

Fig. 4. Qualitative results of MTDC dataset. GT denotes the ground-truth count and PD the predicted count.

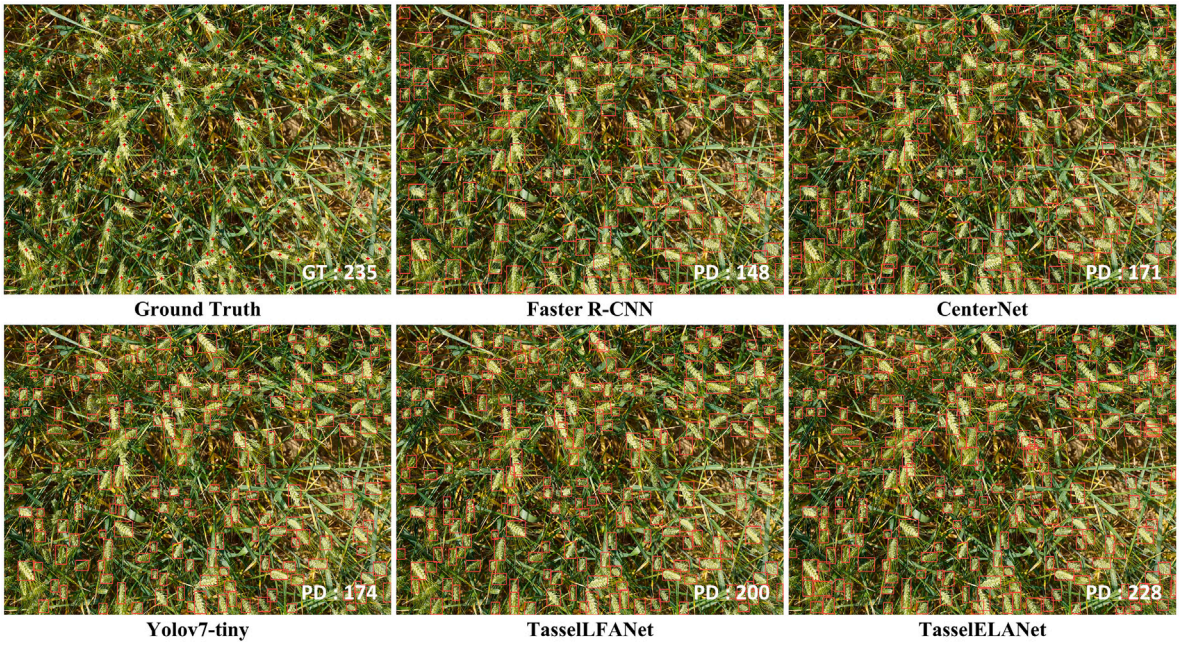

Fig. 5. Qualitative results of WEDU dataset. GT denotes the ground-truth count and PD the predicted count.

induced by cluttered background information (see Fig. 5). This induction may arise from environmental noise, surrounding vision stimuli, interference from other tasks, or cognitive load. Another considerable characteristic of the DRPD dataset is the dense distribution and severe occlusion caused by precision farming practices. The degradation of appearance cues due to low-resolution pixel values from aerial images leads to a rapid decline in model performance (see Table 2). Encouragingly, as shown in Fig. 6, the optimized TasselELANet has greatly improved predictions in these challenging scenarios, and we will present more detailed information later on.

3.4. Robustness analysis

As shown in Fig. 7, the confusion matrices for the three provided datasets (MTDD, WEDU, and DRPD) allow us to evaluate the overall accuracy of the model across different classes. From these three datasets, it can be observed that the model performs well in predicting the background. However, in predicting specific crops such as maize tassels, wheat ears, and rice panicles, the accuracy is relatively low. A major challenge remains the remarkable intrinsic distribution differences that can occur in plant images due to factors such as variety,

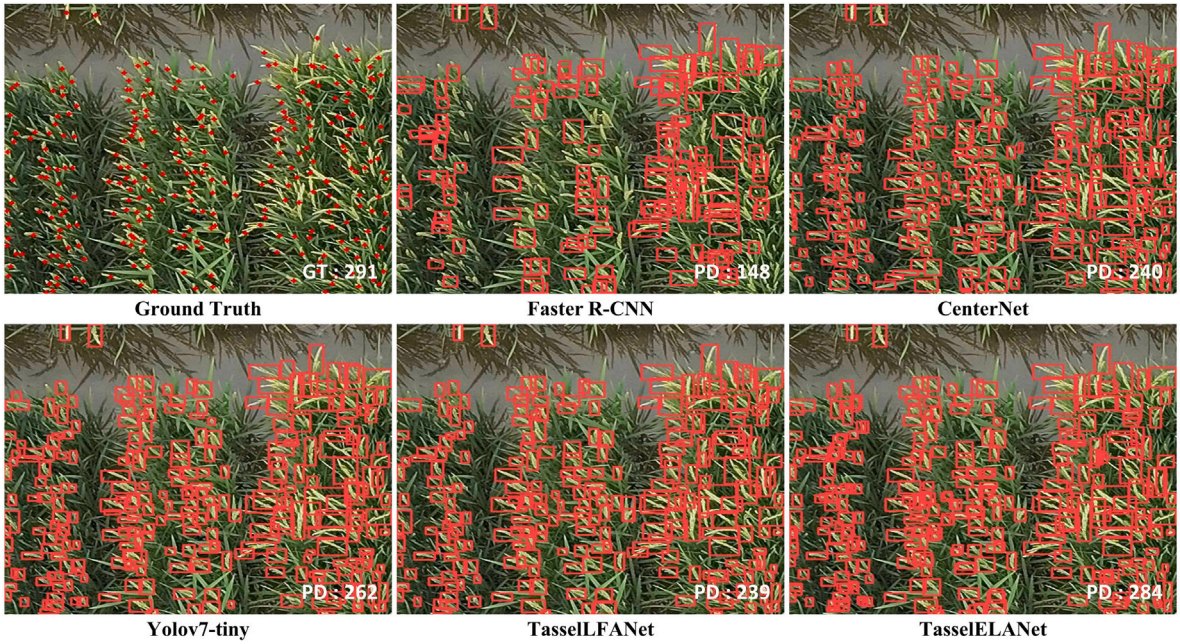

Fig. 6. Qualitative results of DRPD dataset. GT denotes the ground-truth count and PD the predicted count.

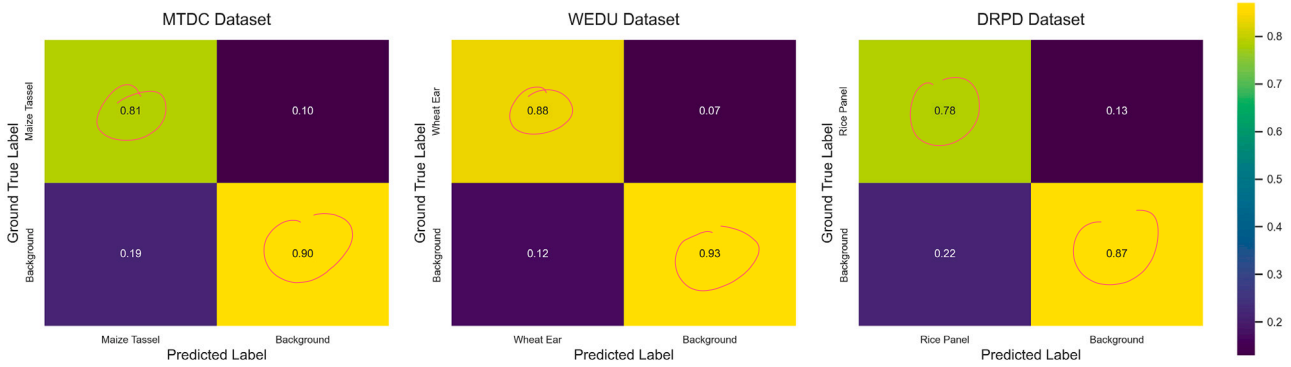

Fig. 7. The confusion matrix of TasselELANet classification on the MTDC, WEDU, and DRPD datasets. An obvious observation is that the classification performance on the WEDU dataset is much better than that on the other two datasets. What is more, TasselELANet achieved a relatively high accuracy on the WEDU dataset, while there is still much room for improvement on the other two datasets.

environment, and growth status (Gupta & Tripathi, 2024; Huang et al., 2023). It can also be observed that, despite the different sources and compositions of these datasets, the model's performance seems similar across all datasets, with the accuracy in predicting the background always being higher than the accuracy in predicting specific crops. In addition, these confusion matrices reveal the tendencies and limitations of TasselELANet in handling different categories. For example, TasselELANet seems to be more prone to misclassifying the background as other classes rather than the other way around. This may be because backgrounds typically have more variability, ambiguity, and complexity.

Furthermore, we also report in Table 3 the three evaluation metrics calculated from the confusion matrices: Accuracy, Sensitivity, and Specificity. These metrics are used to measure the proportion of correctly predicted samples out of all samples, the proportion of correctly predicted positive samples out of all actual positive samples, and the proportion of correctly predicted negative samples out of all actual negative samples, respectively. Comparing the three datasets, it can be seen that although TasselELANet performs well on most datasets, there is still room for improvement. For instance, improving its Sensitivity (Recall) on the DRPD dataset is also a direction for future research, which is related to dense prediction topics (Dang et al., 2024; Yao et al., 2024; Yu, Wang et al., 2024).

Table 3
Performance evaluation of TasselELANet on classification tasks for the MTDD, WEDU, and DRPD datasets.

Dataset	Accuracy	Sensitivity	Specificity
MTDC	0.965	0.957	0.973
WEDU	0.905	0.881	0.928
DRPD	0.827	0.784	0.870

3.5. Linear regression visualization

Linear regression plots have always been an essential tool in our experimental analysis of counting tasks. In Fig. 8, we present the linear regression results of TasselELANet compared to the baseline TasselLFANet. These curves correspond to the predicted results obtained by both models through regression analysis. We also calculated the R^2 score to measure the proportion of variability in the predicted values explained by the influencing factors. The closer the proportion is to 1, the more accurately it describes the true distribution of the data. The linear regression visualization has provided us with effective means to diagnose potential issues, and some regression results with significant biases highlight the challenges posed by these datasets.

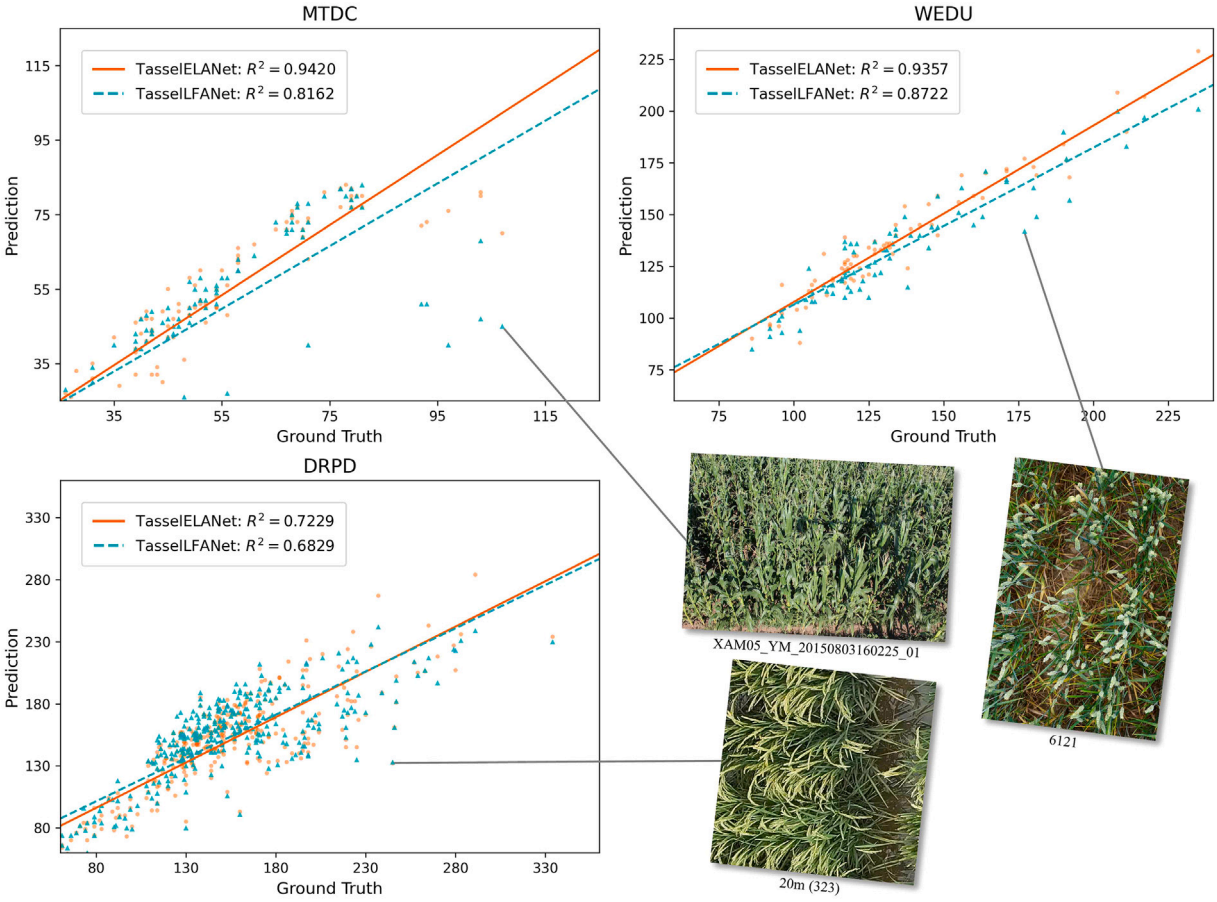

Fig. 8. TasselELANet linear regression results on three datasets, while also showcasing some images with notable prediction errors.

For instance, in the MTDC dataset, the image with relatively large prediction errors from the baseline TasselLFANet corresponds to a challenging vision scenario with severe lighting variations and extremely dense tassel distribution. In the WEDU dataset, the regression error plot could be misled by data noise during the training stage. The DRPD dataset exhibits a high plant density. Here we also note that there is a marked difference in R^2 values, due to the severe occlusion phenomenon, which makes it difficult for the model to accurately identify each rice panicle, leading to a large number of overestimations and underestimations. Overestimation may occur because the model is unable to distinguish adjacent rice panicles and counts them as a single entity. Underestimation may occur because the model fails to identify occluded rice panicles or misclassifies them as background. Moreover, low-resolution images may lead to inaccurate recognition of rice panicle shape, texture, and other information by the model, thereby affecting the counting results. Fortunately, TasselELANet has remarkably improved over the baseline in addressing these challenges.

3.6. Further discussions

Efficiency has always been a key consideration in our design, as it directly impacts whether people with cheap devices can benefit from advanced technology. Particularly in resource-limited agricultural settings, technologies like edge devices, embedded systems, and mobile robots offer practitioners more decision support and production management tools. As shown in Fig. 9, we compare the efficiency of TasselELANet with different methods, where FPS measures the number of image frames the model can process per unit time, and params directly measures model complexity, which is a crucial constraint for deployment. The testing was performed on a mobile device with Nvidia GTX 1650 GPU (4G) and Intel i5-10200H CPU (8G), a low-end hardware configuration with relatively slower processing speed. Overall, TasselELANet not only demonstrates state-of-the-art accuracy but also pronouncedly outperforms other vision methods in efficiency. In other words, even on affordable devices, our insights can still be extended to good practice.

Computer vision problems in the agricultural field remain intricate and challenging. Some issues are widespread. Due to the limited availability of data, when attempting to apply a well-trained model to an unknown environment, performance degradation is often observed. This discrepancy challenges the assumption of consistent data distribution between training and testing in traditional machine learning, leading to domain transfer (Pan & Yang, 2010). One direct solution is the collection and integration of cross-domain plant data, which sometimes requires joint efforts from researchers worldwide, such as the global wheat head detection (GWHD) (David, Serouart, Smith, et al., 2021) and diverse rice panicle detection (DRPD) (Teng et al., 2023) datasets. Another promising aspect is that when the limits of CNN representation capacity are further expanded, domain transfer can be better overcome. What is more, we have noticed the urgent need for improvement in recognition under dense distribution and severe occlusion in the agricultural field. As evident from the linear regression results of the DRPD dataset in Fig. 8, a prominent number of overestimations and underestimations indicate the considerable challenges that these vision patterns pose to rice panicle recognition. Unfortunately, up to this point, TasselELANet remains the best-performing model in this regard. This issue may be alleviated on the focused counting model (Bai et al., 2023; Lu et al., 2022). Another substantial limitation is the high cost and time-consuming nature of collecting agricultural image data due to the natural growth patterns of plants. A single experimental field often yields only 1–2 sequences of images in a year, resulting in very

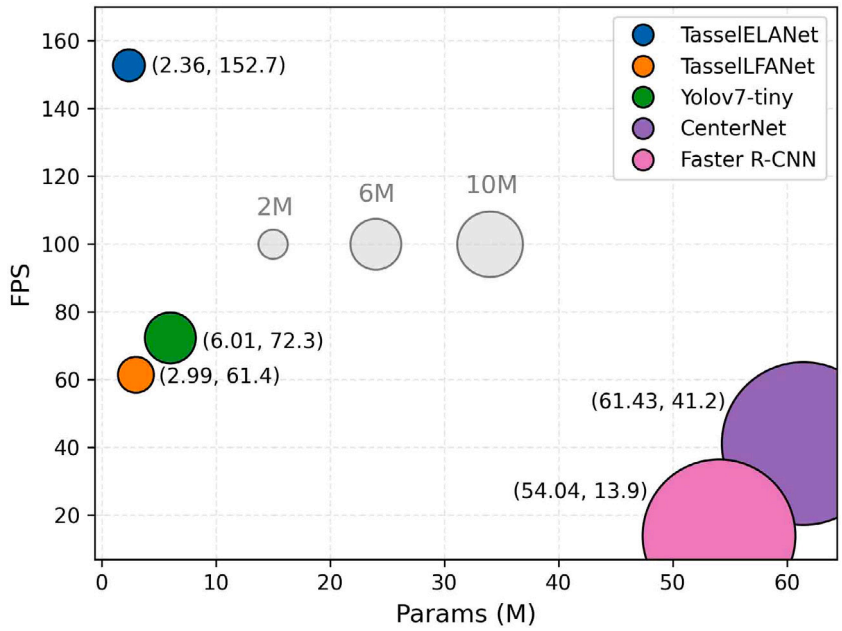

Fig. 9. Efficiency comparison of different methods, TasselELANet emerged as the clear winner. Test mobile devices based on Nvidia GTX 1650 GPU (4G), Intel i5-10200H CPU (8G).

Table 4
TasselELANet's encoder ablation experiments with different downsampling layers r and decoder output channel number c settings.

Dataset	r	c	FLOPs	Params	AP_{50}	AP_{50-95}	MAE	RMSE	R^2
MTDC	1/32	128 + 256 + 512	18.8G	12.84M	0.848	0.440	**3.69**	**6.19**	**0.9470**
	1/32	256 + 512	16.1G	9.47M	0.840	0.430	4.73	6.91	0.9341
	1/16	128 + 256	**14.8G**	**2.36M**	**0.865**	**0.463**	3.99	6.39	0.9420
WEDU	1/32	128 + 256 + 512	18.8G	12.84M	0.928	0.544	7.46	9.10	**0.9362**
	1/32	256 + 512	16.1G	9.47M	0.888	0.492	10.63	16.28	0.8174
	1/16	128 + 256	**14.8G**	**2.36M**	**0.931**	**0.547**	**7.03**	**9.03**	0.9357
DRPD	1/32	128 + 256 + 512	18.8G	12.84M	**0.850**	**0.500**	**17.50**	**24.24**	**0.7553**
	1/32	256 + 512	16.1G	9.47M	0.794	0.439	21.61	27.90	0.6708
	1/16	128 + 256	**14.8G**	**2.36M**	0.848	0.497	18.69	25.69	0.7229

The best performance is in boldface.

limited real-world data availability. Although TasselELANet operates under these constraints in agricultural vision applications, the data used sufficiently covers various field variations. From another perspective, this also provides an architectural reference for achieving excellent representations under the limitation of limited available data. Even so, ample training data remains a key factor for good performance, and exploring the collection of larger and more diverse datasets is still a direction worth investigating in the future.

3.7. Future applications and research suggestions

TasselELANet can serve as a powerful and reliable vision tool for agricultural practitioners. Here, we provide some practical suggestions for practitioners:

1. TasselELANet demonstrates high efficiency in the field of agricultural intelligence, particularly well-suited for tasks such as robotic technology, crop monitoring, pest and disease detection, anomaly recognition, and yield estimation.
2. Due to its ability to provide a comprehensive scene description, TasselELANet exhibits good interpretability, allowing users to understand the model's decision-making process and facilitating optimization and diagnosis of specific parts.
3. For achieving optimal plant image capture, it is advisable to utilize high-resolution cameras under appropriate lighting conditions, minimize background interference, ensure accurate color representation, and capture images from multiple angles.
4. For non-cross-domain applications, collecting images from a single scene is sufficient for training. When deploying on mobile devices, it is preferable to have diverse imaging patterns for plant data.
5. TasselELANet demonstrates tolerance towards weakly supervised learning. However, it is still advisable to prepare high-quality annotated data, especially in automated applications, as predictions with marked errors can be catastrophic.
6. Directly training on UAV datasets captured at high flight altitudes is difficult since TasselELANet's design was not intended for remote sensing applications.

TasselELANet can serve as a powerful benchmark model for researchers. Before continuing with further analysis, we report the contributions of using 16-fold and 32-fold downsampling layer encoders, and the quantitative results are shown in Table 4. The 32-fold downsampling layer encoder follows the same design principles of the deep convolutional neural network, and the two output layers in the final decoder have their channel numbers increased by a factor of one to 256 and 512, compensating for the loss of spatial information due to the downsampling operation, making the comparison fair. We also conducted comparative experiments with a 32-fold downsampling layer that produces outputs at three different scales with channel numbers of 128, 256, and 512. In addition, we report the floating point operations (FLOPs), which are used to measure the total number of floating-point calculations executed by the model during inference and are

commonly used to assess the computational complexity of the model. For future work, we provide the following better research suggestions to researchers:

1. Compared to 32-fold with 2 scales, the advantage of using an encoder with a 16-fold downsampling layer in TasselELANet is significant. Especially in the WEDU and DRPD datasets, where the spatial occupancy of plants is relatively small, this is due to the higher spatial resolution that preserves more detailed information at each pixel position.
2. Although the 32-fold downsampling layer encoder allows for subsequent computations on lower-resolution feature maps, the increase in channel numbers inevitably leads to higher memory consumption and params in the model. In fact, TasselELANet with a 16-fold downsampling layer encoder is more efficient.
3. The design of the 32-fold downsampling layer encoder is often expected to understand information at deeper levels. One direct approach to improve accuracy is to combine data paths (feed-forward paths) with different levels of features and increase the input image resolution. Bai et al. (2023) designed RPNet, which achieved higher counting performance by densely utilizing shallow and deep features, though they reported only 3.57 Hz FPS on the high-performance GPU RTX 3090.
4. Comparative experiments with a 32-fold downsampling layer with 3 scales, as in the ablation experiments, combine more scale-specific feed-forward features by continuing to perform upsampling. Many times, nevertheless, this is not always effective, one possible reason being that, due to the more complex imaging patterns in the agricultural domain, the prediction results from different layers introduce noise, resulting in suboptimal final accuracy when merged, as demonstrated in the study by Yan, Zhao, Cai, et al. (2023). Besides, this pattern comes at the cost of sacrificing more params and FLOPs.
5. Another clever approach is to adopt the compound scaling strategy (Tan & Le, 2019; Wang, Liao and Yeh, 2023). By applying a compound coefficient to uniformly scale all dimensions of depth/width/resolution, this approach consistently achieves better efficiency than existing techniques within a wide range of resource constraints.
6. Furthermore, the core idea of this study revolves around a concise and efficient global architecture based on ELAN for feature encoding. When considering practical applications, choosing a suitable encoder based on specific task scenarios is crucial. This may involve selecting an encoder that sacrifices efficiency to enhance accuracy or one that sacrifices accuracy to improve efficiency. The ELAN we have chosen is a compromise between the two.

Apart from that, due to the constraints imposed by the natural growth patterns of plants, there is a persistent scarcity of agricultural images. Therefore, considering the impact of small sample sizes is very necessary. For future research, we offer the following insights and suggestions:

1. Fine-tuning models pre-trained on large datasets is a common practice (Todescato, Garcia, Balreira, & Carbonera, 2024). Whereas our experience indicates that this method is not effective in the agricultural domain. A key reason is the distribution gap between natural images and plant images. Unlike the geometric and temporal invariance found in many artificial objects in natural images, plant images exhibit considerable intrinsic distribution variations due to factors like species, environment, and growth states.
2. Utilizing data augmentation techniques (Naveed, Anwar, Hayat, Javed, & Mian, 2024) such as image rotation, flipping, scaling, and color transformation can artificially increase the diversity of the training dataset. This helps the model learn a broader data distribution, thereby improving its generalization capability under small sample sizes.
3. Lightweight network design is essential (Sun, Li, & Zhang, 2024; Ye, Yu et al., 2023) because large models often have high complexity, meaning they require more data to train effectively and fine-tune parameters. When the amount of data is insufficient to support the model's complexity, the model's performance can significantly deteriorate.
4. Introducing regularization terms can prevent models from becoming overly complex (Dialameh, Hamzeh, Rahmani, Dialameh, & Kwon, 2024; Qin et al., 2023), thus reducing the likelihood of overfitting. In our previous agricultural work, regularization techniques markedly enhanced the performance of detecting and counting five types of crops (Ye et al., 2024), as well as our work on optimization strategies for regularization (Yu, Ye, Liufu, Lu and Zhou, 2024).
5. Furthermore, employing cross-validation strategies (Jiang, Yan, & Wang, 2024) can provide a more reliable assessment of a model's performance on small sample sets. By dividing the dataset into multiple subsets for training and testing, a stable estimate of model performance can be obtained.

4. Conclusion

The aim of this paper is to provide agricultural practitioners with a powerful and reliable vision tool to help them solve real-world problems more efficiently. Moreover, we hope to offer valuable vision architecture references for future researchers. Through in-depth research and optimization of previous work, we have designed a more powerful vision model, TasselELANet, which features a concise and efficient global architecture. This architecture is highly innovative and notably different from previous work, and it demonstrates generalizability, serving as a solid foundation for future research work. In our experiments, we reported a series of advanced performances of TasselELANet, including a parameter count of 2.36M, an average precision for detection AP_{50} of 0.881, and a mean absolute error for counting MAE of 9.9. In conclusion, we have presented substantial practical directions and suggestions for practitioners and researchers, with the hope of better extending our insights into effective applications in the field.

Funding

This work was supported in part by 2022 key scientific research project of ordinary universities in Guangdong Province under Grant 2022ZDZX4075, in part by 2022 Guangdong province ordinary universities characteristic innovation project under Grant 2022KTSCX251, in part by the Collaborative Intelligent Robot Production & Education Integrates Innovative Application Platform Based on the Industrial Internet under Grant 2020CJPT004, in part by 2020 Guangdong Rural Science and Technology Mission Project under Grant KTP20200153, in part by the Engineering Research Centre for Intelligent equipment manufacturing under Grant 2021GCZX018, in part by the Guangke & Sany Marine Industry Collaborative Innovation Center, and in part by the key task projects of Rural Science and Technology Commissioners in Guangdong Province under Grant KTP20210302.

CRediT authorship contribution statement

Jianxiong Ye: Conceptualization, Methodology, Formal analysis, Writing – original draft. **Zhenghong Yu:** Resources, Software, Review & editing, Funding acquisition, Investigation. **Jiewu Lin:** Visualization, Validation, Data curation. **Hongyuan Li:** Project administration, Supervision. **Lisheng Lin:** Validation, Investigation.

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

The codes, datasets can be accessed at http://github.com/Ye-Sk/TasselELANet.git.

References

Bai, X., Gu, S., Liu, P., Yang, A., Cai, Z., Wang, J., et al. (2023). RPNet: Rice plant counting after tillering stage based on plant attention and multiple supervision network. *The Crop Journal*, *1*(1), 5–10. http://dx.doi.org/10.1016/j.cj.2023.04.005.

Chai, B., Nie, X., Zhou, Q., & Zhou, X. (2024). Enhanced cascade R-CNN for multiscale object detection in dense scenes from SAR images. *IEEE Sensors Journal*, *24*(12), 20143–20153. http://dx.doi.org/10.1109/JSEN.2024.3393750.

Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Computers and Electronics in Agriculture*, *151*, 61–69. http://dx.doi.org/10.1016/j.compag.2018.05.012.

Dang, M., Liu, G., Xu, Q., Li, K., Di Wang, & He, L. (2024). Multi-object behavior recognition based on object detection for dense crowds. *Expert Systems with Applications*, *248*, Article 123397. http://dx.doi.org/10.1016/j.eswa.2024.123397.

David, E., Serouart, M., Smith, D., et al. (2021). Global wheat head detection 2021: An improved dataset for benchmarking wheat head detection methods. *Plant Phenomics*, *2021*(001), 003. http://dx.doi.org/10.34133/2021/9846158.

Dialameh, M., Hamzeh, A., Rahmani, H., Dialameh, S., & Kwon, H. J. (2024). DL-Reg: A deep learning regularization technique using linear regression. *Expert Systems with Applications*, *247*, Article 123182. http://dx.doi.org/10.1016/j.eswa.2024.123182.

Ding, X., Zhang, X., Ma, N., Han, J., Ding, G., & Sun, J. (2021). Repvgg: Making VGG-style ConvNets great again. In *2021 IEEE/CVF conference on computer vision and pattern recognition* (pp. 13728–13737). http://dx.doi.org/10.1109/CVPR46437.2021.01352.

Duan, K., Bai, S., Xie, L., Qi, H., Huang, Q., & Tian, Q. (2019). CenterNet: Keypoint triplets for object detection. In *2019 IEEE/CVF international conference on computer vision* (pp. 6568–6577). http://dx.doi.org/10.1109/ICCV.2019.00667.

Duan, Z., Luo, X., & Zhang, T. (2024). Combining transformers with CNN for multi-focus image fusion. *Expert Systems with Applications*, *235*, Article 121156, doi:10.

Gage, J. L., Miller, N. D., Spalding, E. P., et al. (2017). TIPS: a system for automated image-based phenotyping of maize tassels. *Plant Methods*, *13*(1), http://dx.doi.org/10.1186/s13007-017-0172-8.

Gajjar, R., Gajjar, N., Thakor, V. J., Patel, N. P., & Ruparelia, S. (2021). Real-time detection and identification of plant leaf diseases using convolutional neural networks on an embedded platform. *The Visual Computer*, 1–16. http://dx.doi.org/10.1007/s00371-021-02164-9.

Gevorgyan, Z. (2022). SIoU loss: More powerful learning for bounding box regression. http://dx.doi.org/10.48550/arXiv.2205.12740, arXiv preprint.

Gómez-Flores, W., Garza-Saldaña, J. J., & Varela-Fuentes, S. E. (2019). Detection of huanglongbing disease based on intensity-invariant texture analysis of images in the visible spectrum. *Computers and Electronics in Agriculture*, *162*, 825–835. http://dx.doi.org/10.1016/j.compag.2019.05.032.

Gupta, S., & Tripathi, A. K. (2024). Fruit and vegetable disease detection and classification: Recent trends, challenges, and future opportunities. *Engineering Applications of Artificial Intelligence*, *133*, Article 108260. http://dx.doi.org/10.1016/j.engappai.2024.108260.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision* (pp. 2961–2969). http://dx.doi.org/10.1109/ICCV.2017.322.

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *37*(9), 1904–1916. http://dx.doi.org/10.1007/978-3-319-10578-9_23.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *2015 IEEE international conference on computer vision* (pp. 1026–1034). http://dx.doi.org/10.1109/ICCV.2015.123.

Huang, Y., Qian, Y., Wei, H., Lu, Y., Ling, B., & Qin, Y. (2023). A survey of deep learning-based object detection methods in crop counting. *Computers and Electronics in Agriculture*, *215*, Article 108425. http://dx.doi.org/10.1016/j.compag.2023.108425.

Jiang, Y., Yan, R., & Wang, X. (2024). A deep learning model for predicting non-histone crotonylation sites in plants. *Plant Methods*, *20*, 28. http://dx.doi.org/10.1186/s13007-024-01157-8.

Jurado-Ruiz, F., Nguyen, T.-P., Peller, J., Aranzana, M. J., Polder, G., & Aarts, M. G. M. (2024). LeTra: a leaf tracking workflow based on convolutional neural networks and intersection over union. *Plant Methods*, *20*(11), 1–11. http://dx.doi.org/10.1186/s13007-024-01138-x.

Lawal, O., & Zhao, H. (2021). YOLOFig detection model development using deep learning. *IET Image Processing*, *15*, 3071–3079. http://dx.doi.org/10.1049/ipr2.12293.

Li, Y., Bao, Z., & Qi, J. (2022). Seedling maize counting method in complex backgrounds based on YOLOV5 and Kalman filter tracking algorithm. *Frontiers in Plant Science*, *13*, http://dx.doi.org/10.3389/fpls.2022.1030962.

Li, J., Wang, E., Qiao, J., et al. (2023). Automatic rape flower cluster counting method based on low-cost labeling and UAV-RGB images. *Plant Methods*, *19*, http://dx.doi.org/10.1186/s13007-023-01017-x.

Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., et al. (2020). Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection. In *NIPS'20: proceedings of the 34th international conference on neural information processing systems*. http://dx.doi.org/10.48550/arXiv.2006.04388.

Li, Y., Zhang, X., & Chen, D. (2018). CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 1091–1100). http://dx.doi.org/10.1109/CVPR.2018.00120.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 936–944). http://dx.doi.org/10.1109/CVPR.2017.106.

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 8759–8768). http://dx.doi.org/10.1109/CVPR.2018.00913.

Lu, H., Liu, L., Li, Y.-N., Zhao, X.-M., Wang, X.-Q., & Cao, Z.-G. (2022). TasselNetV3: Explainable plant counting with guided upsampling and background suppression. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 1–15. http://dx.doi.org/10.1109/TGRS.2021.3058962.

Lu, D., Ye, J., Wang, Y., & Yu, Z. (2023). Plant detection and counting: Enhancing precision agriculture in UAV and general scenes. *IEEE Access*, *11*, 116196–116205. http://dx.doi.org/10.1109/ACCESS.2023.3325747.

Ma, N., Zhang, X., Zheng, H.-T., & Sun, J. (2018). ShuffleNet V2: Practical guidelines for efficient CNN architecture design. *Vol. 11218*, In *Proceedings of the computer vision (ECCV) 2018*. Cham: Springer, http://dx.doi.org/10.1007/978-3-030-01264-9_8.

Madec, S., Jin, X., Lu, H., De Solan, B., Liu, S., Duyme, F., et al. (2019). Ear density estimation from high resolution RGB imagery using deep learning technique. *Agricultural and Forest Meteorology*, *264*, 225–234. http://dx.doi.org/10.1016/j.agrformet.2018.10.013.

Misra, T., Arora, A., Marwaha, S., Chinnusamy, V., Rao, A. R., Jain, R., et al. (2020). SpikeSegNet - A deep learning approach utilizing encoder-decoder network with hourglass for spike segmentation and counting in wheat plant from visual imaging. *Plant Methods*, *16*, http://dx.doi.org/10.1186/s13007-020-00582-9.

Naveed, H., Anwar, S., Hayat, M., Javed, K., & Mian, A. (2024). Survey: Image mixing and deleting for data augmentation. *Engineering Applications of Artificial Intelligence*, *131*, Article 107791. http://dx.doi.org/10.1016/j.engappai.2023.107791.

Padmavathi, B., BhagyaLakshmi, A., Vishnupriya, G., & Datchanamoorthy, K. (2024). IoT-based prediction and classification framework for smart farming using adaptive multi-scale deep networks. *Expert Systems with Applications*, *254*, Article 124318. http://dx.doi.org/10.1016/j.eswa.2024.124318.

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1345–1359. http://dx.doi.org/10.1109/TKDE.2009.191.

Pei, H., Sun, Y., Huang, H., Zhang, W., Sheng, J., & Zhang, Z. (2022). Weed detection in maize fields by UAV images based on crop row preprocessing and improved YOLOv4. *Agriculture*, *12*, http://dx.doi.org/10.3390/agriculture12070975.

Qin, C., Zheng, B., Zeng, J., Chen, Z., Zhai, Y., Genovese, A., et al. (2023). Dynamically aggregating MLPs and CNNs for skin lesion segmentation with geometry regularization. *Computer Methods and Programs in Biomedicine*, *238*, Article 107601. http://dx.doi.org/10.1016/j.cmpb.2023.107601.

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. http://dx.doi.org/10.1109/TPAMI.2016.2577031.

Sun, B., Li, Q., & Zhang, Z. (2024). Dynamic context modeling based lightweight high-resolution network for dense prediction. *Engineering Applications of Artificial Intelligence*, *129*, Article 107642. http://dx.doi.org/10.1016/j.engappai.2023.107642.

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the international conference on machine learning* (pp. 6105–6114). http://dx.doi.org/10.48550/arXiv.1905.11946.

Teng, Z., Chen, J., Wang, J., Wu, S., Chen, R., Lin, Y., et al. (2023). Panicle-cloud: An open and AI-powered cloud computing platform for quantifying rice panicles from drone-collected imagery to enable the classification of yield production in rice. *Plant Phenomics*, http://dx.doi.org/10.34133/plantphenomics.0105.

Todescato, M. V., Garcia, L. F., Balreira, D. G., & Carbonera, J. L. (2024). Multiscale patch-based feature graphs for image classification. *Expert Systems with Applications*, *235*, Article 121116. http://dx.doi.org/10.1016/j.eswa.2023.121116.

Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. (pp. 7464–7475). http://dx.doi.org/10.1109/CVPR52729.2023.00721.

Wang, J., Chen, K., Xu, R., Liu, Z., Loy, C. C., & Lin, D. (2019). CARAFE: Content-aware ReAssembly of features. In *Proceedings of the IEEE international conference on computer vision* (pp. 3007–3016). http://dx.doi.org/10.1109/ICCV.2019.00310.

Wang, C.-Y., Liao, H.-Y. M., & Yeh, I.-H. (2023). Designing network design strategies through gradient path analysis. *Journal of Information Science and Engineering (JISE)*, *39*(4), 975–995.

Wen, C., Chen, H., Ma, Z., Zhang, T., Yang, C., Su, H., et al. (2022). PestYOLO: A model for large-scale multiclass dense and tiny pest detection and counting. *Frontiers in Plant Science*, *13*, http://dx.doi.org/10.3389/fpls.2022.973985.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., et al. (2023). ConvNeXt V2: Co-designing and scaling ConvNets with masked autoencoders. In *2023 IEEE/CVF conference on computer vision and pattern recognition* (pp. 16133–16142). http://dx.doi.org/10.1109/CVPR52729.2023.01548.

Xu, X., Wang, L., Shu, M., Liang, X., Ghafoor, A., Liu, Y., et al. (2022). Detection and counting of maize leaves based on two-stage deep learning with UAV-based RGB image. *Remote Sensing*, *14*, http://dx.doi.org/10.3390/rs14215388.

Yan, J., Zhao, J., Cai, Y., et al. (2023). Improving multi-scale detection layers in the deep learning network for wheat spike detection based on interpretive analysis. *Plant Methods*, *19*(46), http://dx.doi.org/10.1186/s13007-023-01020-2.

Yao, M., Li, W., Chen, L., Zou, H., Zhang, R., Qiu, Z., et al. (2024). Rice counting and localization in unmanned aerial vehicle imagery using enhanced feature fusion. *Agronomy*, *14*(4), http://dx.doi.org/10.3390/agronomy14040868.

Ye, Z., Yang, K., Lin, Y., Guo, S., Sun, Y., Chen, X., et al. (2023). A comparison between pixel-based deep learning and object-based image analysis (OBIA) for individual detection of cabbage plants based on uav visible-light images. *Computers and Electronics in Agriculture*, *209*, http://dx.doi.org/10.1016/j.compag.2023.107822.

Ye, J., & Yu, Z. (2024). Fusing global and local information network for tassel detection in UAV imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *17*, 4100–4108. http://dx.doi.org/10.1109/JSTARS.2024.3356520.

Ye, J., Yu, Z., Wang, Y., Lu, D., & Zhou, H. (2024). PlantBiCNet: A new paradigm in plant science with bi-directional cascade neural network for detection and counting. *Engineering Applications of Artificial Intelligence*, *130*, Article 107704. http://dx.doi.org/10.1016/j.engappai.2023.107704.

Ye, J., Yu, Z., Wang, Y., et al. (2023). WheatLFANet: In-field detection and counting of wheat heads with high-real-time global regression network. *Plant Methods*, *19*(1), http://dx.doi.org/10.1186/s13007-023-01079-x.

Yu, Z., Wang, Y., Ye, J., Liufu, S., Lu, D., Zhu, X., et al. (2024). Accurate and fast implementation of soybean pod counting and localization from high-resolution image. *Frontiers in Plant Science*, *15*(1320109), http://dx.doi.org/10.3389/fpls.2024.1320109.

Yu, Z., Ye, J., Li, C., Zhou, H., & Li, X. (2023). TasselLFANet: A novel lightweight multibranch feature aggregation neural network for high-throughput image-based maize tassels detection and counting. *Frontiers in Plant Science*, *14*, http://dx.doi.org/10.3389/fpls.2023.1158940.

Yu, Z., Ye, J., Liufu, S., Lu, D., & Zhou, H. (2024). TasselLFANetV2: Exploring vision models adaptation in cross-domain. *IEEE Geoscience and Remote Sensing Letters*, *21*, 1–5. http://dx.doi.org/10.1109/LGRS.2024.3382871.

Yu, X., Yin, D., Nie, C., Ming, B., et al. (2022). Maize tassel area dynamic monitoring based on near-ground and UAV RGB images by U-Net model. *Computers and Electronics in Agriculture*, *203*, http://dx.doi.org/10.1016/j.compag.2022.107477.

Zou, H., Lu, H., Li, Y., Li, J., & Zhang, L. (2020). Maize tassels detection: a benchmark of the state of the art. *Plant Methods*, *16*, http://dx.doi.org/10.1186/s13007-020-00651-z.