ORIGINAL RESEARCH **Open access**

Attention-Augmented YOLOv11 for Robust Tea Bud Detection in Complex Agricultural Environments

João Silva[1*], Pedro Costa[1], Ana Beatriz[2], Rafael Mendes[3], Carlos Lima[1], Fernanda Souza[2]

Abstract

Automated identification of tea buds is a vital component of intelligent tea-harvesting systems. Yet obstacles such as complex environmental conditions and the strong visual similarity between tea buds and their surroundings often lead to incorrect and overlooked detections, particularly among tiny buds. To overcome these hurdles, this study presents an automatic TBD technique, TBD-Y, grounded in the YOLOv11 detection architecture. To begin with, a Synergistic Object-Spatial Attention (SOSA) module is designed, incorporating a novel Local Context Attention (LCA) mechanism to boost feature representation along both spatial and regional axes. This sharpens the network's focus on tea bud zones and reduces background noise. Next, a Global-local Attention Guided Feature Fusion (GAGFF) strategy is crafted with two distinct paths: one strengthens low-resolution, semantically rich high-level features. At the same time, the other refines high-resolution features that safeguard fine-grained visual cues. The integration of these two paths elevates the descriptive power of the features. The SOSA and GAGFF components are woven into the YOLOv11 architecture, yielding three TBD model configurations with varying parameter scales: TBD-Y-L, TBD-Y-M, and TBD-Y-S. Trials on a self-collected TBD dataset and the public Global Wheat Head Dataset 2021 (GWHD_2021) reveal that TBD-Y-L outperforms current state-of-the-art approaches, achieving the highest detection accuracy. Additionally, the TBD-Y-S variant achieves improved accuracy over YOLOv11-L while requiring fewer parameters and a lower computational budget.

Keywords YOLOv11, Tea bud detection, Synergistic object-spatial attention, Local context attention, Global-local attention guided feature fusion

*Correspondence:
João Silva
joao.silva@ufrj.br
[1] Department of AI for Sugarcane and Soybean Precision Agriculture, Faculty of Agricultural Sciences, Federal University of Rio de Janeiro, Rio de Janeiro, Brazil
[2] Department of Machine Learning for Pest and Weed Identification, Faculty of Agriculture, University of Campinas, Campinas, Brazil
[3] Department of Smart Harvesting Robotics, Faculty of Agriculture, University of São Paulo, São Paulo, Brazil

Introduction

Automatic Tea Bud Detection (TBD) is an essential element of intelligent tea-picking machinery [1], making a key contribution to improving tea quality and reducing production and workforce costs. Tea buds are diminutive, crowded, and heavily occluded, and their coloring closely mimics a cluttered background, making precise detection profoundly difficult.

Lately, Deep Learning has advanced rapidly and achieved major advances across a variety of computer vision domains, including image classification, object detection, and image segmentation. Deep Learning-driven object detectors have steadily emerged and achieved exceptional performance, surpassing conventional approaches by a wide margin. Investigators have devised several TBD solutions anchored in Deep Learning detection frameworks such as the YOLO

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

line and Faster R-CNN (Region-based Convolutional Neural Network), achieving respectable detection rates [2-6]. Still, present-day TBD techniques bear the following inadequacies:

(1) Images of tea buds often depict complex surroundings and tightly packed objects, hampering precise localization and recognition of individual buds. Detection precision demands further uplift.

(2) Due to disparate shooting perspectives, distances, and other variables, tea buds in images can appear at varying scales and orientations. These aspects have not been suitably managed. The outcome is a high incidence of incorrect and missed detections, particularly for small buds.

To remedy the deficiencies above, a TBD pipeline built on the YOLOv11 detection framework is proposed, named TBD-Y. To cater to TBD's distinctive traits and demands, a specialized attention mechanism and a feature fusion scheme are purposefully designed and embedded into YOLOv11 to enhance detection performance. The core contributions of this work are as follows:

(1) A Synergistic Object-Spatial Attention (SOSA) mechanism is advanced, which steers the model into concentrating more intently on tea bud regions whilst curbing background interference. Incorrect and missed detection episodes can thereby be lessened.

(2) A Global-local Attention Guided Feature Fusion (GAGFF) blueprint is formulated that encompasses two pathways. One pathway boosts low-resolution feature maps that carry dense semantic content, whereas the other upgrades high-resolution feature maps that hold richer low-level visual detail. Fusing the feature outputs from these two pathways effectively unites information across scales, better equipping the model to handle tea buds subject to pronounced size and orientation shifts.

Experimental evidence from both the custom TBD dataset and the open Global Wheat Head Dataset 2021 (GWHD_2021) demonstrates that the offered TBD-Y model achieves higher detection performance than prior methods.

The remainder of the manuscript proceeds as follows: Section 2 surveys the prior art; Section 3 elaborates on the introduced approach; Section 4 presents the experimental results and discussion; Section 5 wraps up the study.

Related Works

The recent surge in Deep Learning capabilities has spurred the development of numerous object detection solutions tailored to different crop types. Most of these solutions are built on established general-purpose detection frameworks and fall into two broad classes: single-stage and dual-stage approaches.

Single-stage representatives span the YOLO lineage [7-12], SSD (Single Shot MultiBox Detector) [13], and RetinaNet [14], whereas dual-stage examples cover Faster R-CNN [15], Mask R-CNN [16], and DetectoRS [17].

In their work, Chen *et al.* [2] described a TBD pipeline that weaves the Selective Kernel Attention module alongside an original multi-feature blending design into the YOLOv7 backbone, surpassing the baseline YOLOv7 in detection metrics.

Li *et al.* [18] engineered a lightweight yet highly precise TBD system rooted in YOLOv4. The backbone network is swapped out for GhostNet, standard convolutional operations are replaced with depth-wise variants to reduce parameter volume, and the Convolutional Block Attention Module (CBAM) is embedded within the Path Aggregation Network (PANet). The SIoU loss further guides training. Empirical evaluation confirms that this setup delivers competitive detection figures while maintaining a compact parameter footprint and modest computational load.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

To address the picking point problem, Chen *et al.* [3] formulated a two-step procedure: first, Faster R-CNN isolates One Tip with Two Leaves (OTTL) areas, and then a Fully Convolutional Network (FCN) determines the exact picking coordinates within those areas. Tests on their internally gathered data yield encouraging results.

The Transformer paradigm has likewise begun penetrating tea leaf analysis tasks. Chen *et al.* [19] proposed a blended CNN-Transformer architecture aimed at diagnosing pest infestations and diseases on tea foliage. The design rests on a MobileViT backbone, supplemented by EMA (Efficient Multi-scale Attention), RFBNet (Receptive Field Block Network), and ODConv (Omni-dimensional Dynamic Convolution) modules, all working together to extract nuanced cues from leaf images and improve accuracy.

Nonetheless, while Transformer-driven models can outperform purely CNN-based systems in detection quality, they are hampered by higher parameter counts and a reliance on copious amounts of labeled training data. These factors curb their broader practical use.

Measured against its predecessors, YOLOv11 achieves stronger feature extraction through a polished network design and more faithfully captures elaborate patterns within images. Concurrently, the optimization regimen has been refined, enabling the model to achieve high performance benchmarks with fewer parameters and faster inference. Motivated by these merits, the present study advances a TBD approach anchored in YOLOv11. It directly addresses the inherent challenges of TBD by integrating purpose-built attention and feature-merging components into YOLOv11, yielding appreciable gains in detection efficacy.

Materials and Methods

The comprehensive network layout of the proposed TBD approach is shown in **Figure 1**. The system builds directly on the YOLOv11 foundation, augmenting it with two custom-engineered modules: the Synergistic Object-Spatial Attention (SOSA) mechanism and the Global-local Attention Guided Feature Fusion (GAGFF) strategy, both aimed at improving detection quality. The subsections that follow unpack the construction details of SOSA and GAGFF in turn. In particular, within the Neck segment of the network, SOSA computes an object-centric attention map from the combined low-level representations and subsequently merges it with layer-wise outputs, whereas GAGFF supplants the ordinary concatenation step.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

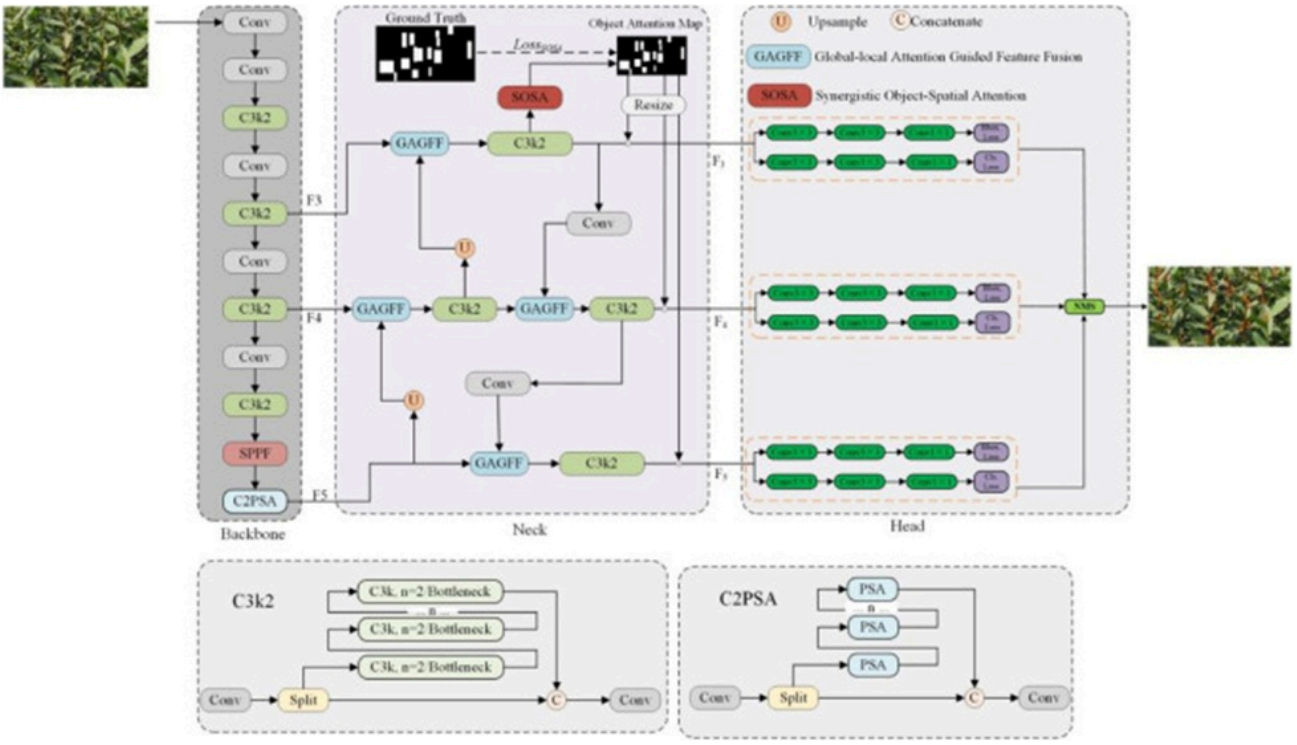


Figure 1. The network architecture of the proposed TBD-Y.

Synergistic object-spatial attention

Attention schemes are modeled after biological visual perception, enabling neural networks to autonomously prioritize the most informative portions of incoming data and operate more efficiently. In the Deep Learning context, such mechanisms assign non-uniform weights to individual features, amplifying those that matter for the given objective and subduing extraneous signals. The resultant boost in performance can be dramatic. As a result, attention has been widely adopted across domains, from natural language processing to computer vision, consistently yielding notable improvements.

In the specific setting of TBD, a core obstacle is the pronounced chromatic overlap between tea buds and surrounding greenery, compounded by their dense packing, which often leads to both false positives and missed targets. To address this, the current paper proposes a Synergistic Object-Spatial Attention (SOSA) mechanism. Its role is to channel the network's focus more rigorously toward tea bud neighborhoods and dampen the influence of background elements, thereby mitigating the twin problems of erroneous and absent detections.

As shown in **Figure 1**, the YOLOv11 design incorporates a Feature Pyramid Network (FPN) that performs top-down aggregation of feature representations extracted from multiple backbone depths, resulting in a set of multi-scale feature maps {Fi, i ∈ (3, 4, 5)}. The SOSA module processes these maps to derive an attention distribution that preferentially highlights tea buds. That attention distribution is then broadcast across the multi-scale maps {Fi, i ∈ (3, 4, 5)}, strengthening signals within bud-bearing zones and weakening those originating from the backdrop.

Figure 2 lays out the internal wiring of the SOSA module. The F3 first passes through the Local Context Attention unit, which accurately pinpoints candidate regions and object centers. Subsequently, a sequence comprising deconvolution, dilated convolution, and a softmax operation is executed to produce the Object Attention Map (OAM). In this map, pixels corresponding to tea bud areas are assigned higher weights, while background pixels receive lower weights. At a high level, SOSA breaks down into two functional blocks: the LCA and the OAM construction procedure, both of which are expounded upon below.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

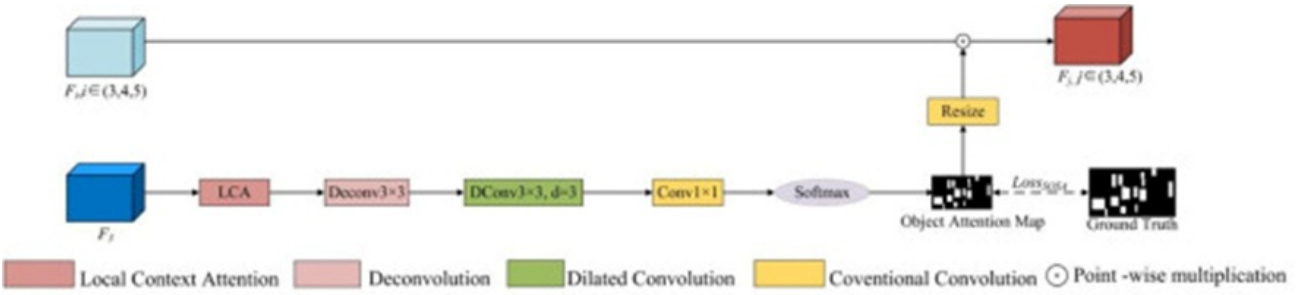


Figure 2. The structure of the SOSA mechanism.

Local context attention

Figure 3 presents the design of the LCA module. The structure comprises two convolutional branches plus a residual pathway. Inside the convolutional branches, the spatial dimension first undergoes strip pooling, which harvests horizontally and vertically oriented features and thereby encodes long-range relationships. A 1 × 1 convolution then shrinks the channel depth. To efficiently gather contextual signals, two depth-wise separable strip convolutions oriented along the horizontal and vertical axes are adopted—this expands the receptive field while maintaining a lean parameter budget and also supports feature extraction from slender, elongated shapes. A subsequent 1 × 1 convolution restores the original channel count. After being processed by Group Normalization (GN) followed by a sigmoid nonlinearity, two spatial position attention maps—one per direction—are obtained. These two maps are ultimately merged with the input feature volume through a residual connection, producing the enriched representation Fm. The computation can be formally written as:

$$F_c^H = \frac{1}{H} \sum_{0 \le i < H} I_c(h, i) \quad (1)$$

$$F_c^W = \frac{1}{W} \sum_{0 \le i < H} I_c(j, w) \quad (2)$$

$$F^H = DWSConv_{1 \times k_b}\left(Conv_{1 \times 1}\left(F_c^H\right)\right) \quad (3)$$

$$F^W = DWSConv_{k_b \times 1}\left(Conv_{1 \times 1}\left(F_c^W\right)\right) \quad (4)$$

$$F_n^H = \sigma\{GN(Conv_{1 \times 1}(F_H))\} \quad (5)$$

$$F_n^W = \sigma\{GN(Conv_{1 \times 1}(F_W))\} \quad (6)$$

$$F_m = I_c(i, j) + F_n^H + F_n^W \quad (7)$$

© 2022 The Author(s). | Published by Agro Intelligence Research Press



Figure 3. The structure of LCA.

Let the incoming feature block be $\mathrm{R}^{H\times W\times C}$. In strip pooling, each channel is independently pooled across the H and W axes. F_c^H denotes the c-th channel's output along the H axis, and F_c^W denotes the c-th channel's output along the W axis. The notations DWSConv1×kb and DWSConvkb×1 refer to depth-wise strip convolutions with kernel dimensions 1 × kb and kb × 1, whereas σ stands for the sigmoid gating function.

Once transformed by LCA, the features equip the network to zero in on regions of interest with high spatial accuracy, all while remaining computationally modest. The design also widens the receptive span and competently absorbs surrounding contextual patterns.

Object attention map generation

The object attention map helps channel the model's attention more strongly to tea bud zones. It is constructed by assigning region-specific weighting coefficients across the image plane. The step-by-step generation protocol is detailed below.

Denote the representations existing in LCA as $Fm\in.\mathrm{R}^{H\times W\times C}$ These undergo a deconvolution operation, yielding feature maps $F'm\in\mathrm{R}^{640\times 640\times C}$ whose spatial extent matches the original input imagery. A 3 × 3 dilated convolution is subsequently applied to broaden the receptive coverage. Thereafter, a 1 × 1 convolution compresses the channel dimension without altering the spatial resolution, resulting in $Fn\in\mathrm{R}^{640\times 640\times 2}$, a two-channel tensor. A softmax activation is then imposed to derive the object attention map $m\in\mathrm{R}^{640\times 640\times 1}$. This map m is next resized into a set of multi-resolution attention maps, denoted {mi, i ∈ (3, 4, 5)}. The feature pyramids {*Fi*,*i*∈(3,4,5)} are multiplied pixel-wise with the correspondingly resized attention maps, yielding the reinforced outputs {*Fj*,*j*∈(3,4,5)}. The overall procedure can be cast in mathematical form as:

$$F_m = LCA\left(F_3\right) \quad (8)$$

$$F'_m = Deconv_{3\times 3}\left(F_m\right) \quad (9)$$

$$F_n = Dconv_{3\times 3}\left(Conv_{3\times 3}\left(F'_m\right)\right) \quad (10)$$

$$F'_n = Conv_{1\times 1}\left(F_n\right) \quad (11)$$

$$m = Softmax(F'_n) \quad (12)$$

© 2022 The Author(s). | Published by Agro Intelligence Research Press

$$m_i = Resize\,(m)\,, i \in (3,4,5) \quad (13)$$

$$F_j = m_i \cdot F_i, i \in (3,4,5) \quad (14)$$

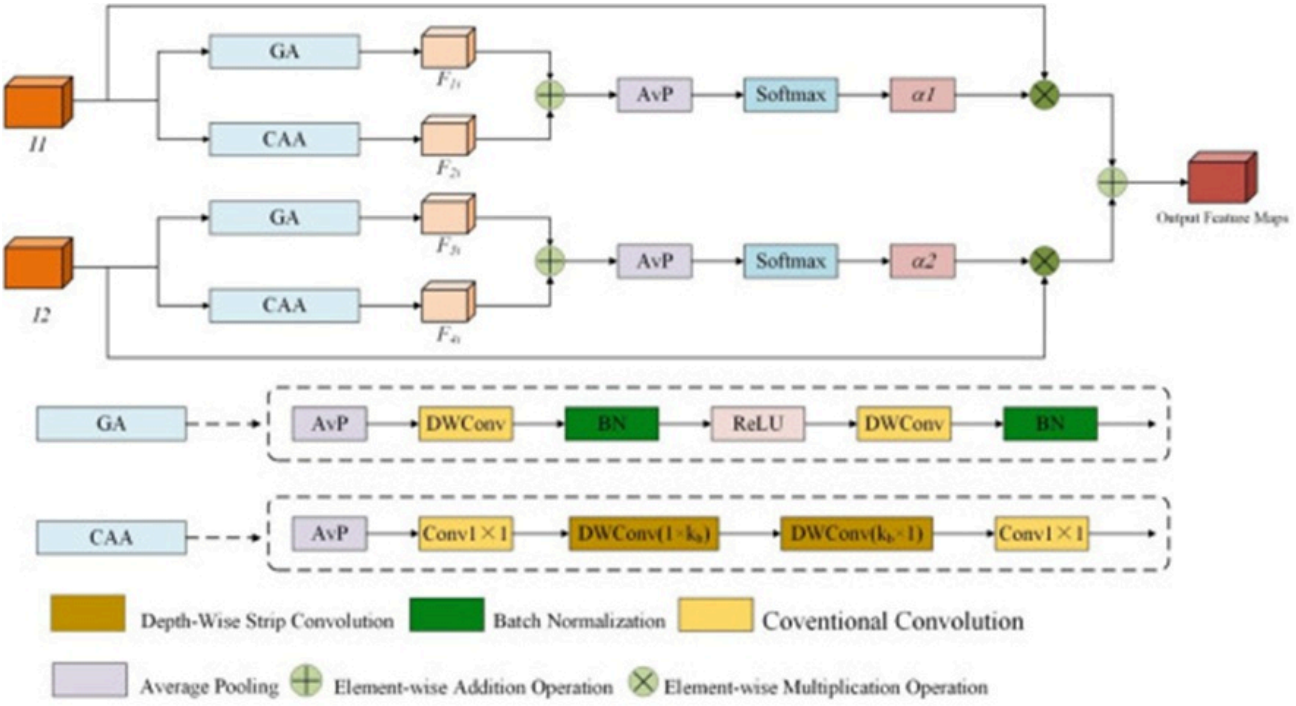
To sharpen SOSA's capacity to amplify tea bud signals selectively, this study employs the Binary Cross-Entropy Loss (BCE Loss) to guide the construction of the OAM. BCE Loss gauges the mismatch between model forecasts and the Ground Truth (GT) via the cross-entropy metric. Its expression is:

$$\begin{aligned} Loss_{SOSA} &= \frac{1}{N}\sum_i L_i \\ &= \frac{1}{N}\sum_i -[y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)] \end{aligned} \quad (15)$$

In this notation, yi denotes the GT class indicator for a sample, restricted to 0 or 1. pi denotes the model's estimated probability, bounded in [0,1]. N signifies the total quantity of samples. The GT utilized for supervising the OAM is identical to the image-level annotation—a manually delineated bounding box produced with the LabelImg utility.

Global-local attention guided feature fusion

Standard Feature Pyramid Networks (FPN) largely rely on concatenation to combine feature maps at different resolutions, thereby pooling multi-scale contextual cues. Yet straightforward concatenation often overlooks the semantic discrepancies that naturally separate distinct feature hierarchies. To remedy this, the present work introduces the GAGFF strategy, which fuses cross-scale representations more judiciously, thereby elevating TBD accuracy. The schematic of the GAGFF strategy appears in **Figure 4**. Unlike plain concatenation, GAGFF unfolds as two parallel arms. Each arm refines the incoming representations from a complementary viewpoint. The results from both arms are then concatenated into a unified feature volume. This composite feature map simultaneously carries abstract semantic knowledge and granular low-level textural cues, thereby strengthening the detector's ability to handle tea buds of widely varying dimensions.



© 2022 The Author(s). | Published by Agro Intelligence Research Press

Figure 4. The structure of GAGFF.

Let I1 and I2 symbolize the two feature tensors slated for fusion, with shapes and . Initially, both I1 and I2 are routed through a Global Attention (GA) unit and a Context Anchor Attention (CAA) unit [20], yielding four feature bundles: F1, F2, F3, and F4. Element-wise summation is then executed on the pairs (F1, F2) and (F3, F4), resulting in Fn1 and Fn2. To discard spatial position dependency and distill channel-wise global statistics, global average pooling is applied to both Fn1 and Fn2, yielding holistic summaries. A softmax operator subsequently computes two scalar weights α1 and α2, one for each arm. The raw input feature maps are multiplied element-wise by α1 and α2, generating the weighted feature tensors F′n1 and F′n2. These two are finally combined via element-wise addition, yielding the output representation O. The precise sequence can be expressed as:

$$F_1, F_3 = GA\left(I1, I2\right) \quad (16)$$

$$F_2, F_4 = CAA\left(I1, I2\right) \quad (17)$$

$$F_{n1} = F_1 + F_2, F_{n2} = F_3 + F_4 \quad (18)$$

$$\alpha 1, \alpha 2 = Softmax(\mathrm{AvP}\left(\mathrm{F_{n1}}\right), \mathrm{AvP}\left(\mathrm{F_{n2}}\right)) \quad (19)$$

$$F'_{n1} = \alpha 1 \times I1 \quad (20)$$

$$F'_{n2} = \alpha 2 \times I2 \quad (21)$$

$$O = F'_{n1} + F'_{n2} \quad (22)$$

In the given equations, GA abbreviates Global Attention, CAA abbreviates Context Anchor Attention, and AvP stands for average pooling. The modules GA and CAA are formulated as follows:

$$GA = \beta\left(DWConv\left(\sigma\left(\beta\left(DWConv\left(AvP\left(I_i\right)\right)\right)\right)\right), i \in (1,2)\right. \quad (23)$$

$$\begin{aligned} &CAA \\ &= AvP(Conv_{1\times1}\left(DWSConv_{k_b\times1}\left(DWSConv_{1\times k_b}\left(\mathrm{Conv_{1\times1}}\left(I_i\right)\right)\right)\right), \\ &i \in (1,2) \end{aligned} \quad (24)$$

Where β indicates batch normalization, σ denotes the sigmoid function, and DWSConv refers to depth-wise strip convolution.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Total loss

The composite loss function adopted throughout this paper is:

$$Loss_{total} = Loss_{SOSA} + Loss_{Cls} + Loss_{Reg} \quad (25)$$

Within this expression, LossSOSA corresponds to the BCE loss term employed inside SOSA, whereas LossCls and LossReg designate the classification and regression loss components, respectively. Their detailed formulations are:

$$\begin{aligned} Loss_{Cls} &= \frac{1}{N}\sum_i L_i \\ &= \frac{1}{N}\sum_i -[y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)] \end{aligned} \quad (26)$$

$$DFL(S_i, S_{i+1}) = -(y_{i+1} - y)\, log(S_i) + (y - y_i)\, log(S_{i+1}) \quad (27)$$

$$CIOU = 1 - IOU + \frac{p^2(b, b^{gt})}{c^2} + av \quad (28)$$

$$Loss_{Reg} = CIOU + DFL(S_i, S_{i+1}) \quad (29)$$

In these relations, y denotes the actual bounding box coordinate, and yi plus yi+1 are its discretized neighboring coordinates, obeying yi ≤y<yi+1. Si and Si+1 represent the distribution probabilities emitted by the model, indicating the chance that y falls within the interval [yi, yi+1). The symbols b and bgt refer to the central points of two rectangular boxes, p measures their Euclidean distance, and c is the diagonal extent of the smallest enclosing rectangle that wraps both boxes. The term v quantifies the agreement in aspect ratio between the two boxes, while a serves as a balancing coefficient.

Results and Discussion

The designed SOSA was integrated into the YOLOv11-S, YOLOv11-M, and YOLOv11-L variants, and the GAGFF module replaced the standard concatenation steps within these backbones. This integration produced three TBD model scales, named TBD-Y-L, TBD-Y-M, and TBD-Y-S. To substantiate the effectiveness of the proposed architectures, empirical evaluations were conducted on two data sources: the custom-built TBD collection and the publicly accessible GWHD_2021 benchmark. The ensuing subsections detail the experimental findings and their interpretation.

Datasets and evaluation criteria

Datasets

Because no off-the-shelf public TBD corpus exists, a bespoke TBD dataset was initially curated to verify the methodology's merits. Additionally, tests were performed on the publicly available GWHD_2021 dataset [21]—a task-adjacent resource—to further probe the approach's generalizability.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Self-built TBD dataset. Photographs in this collection were taken with digital cameras and mobile phones spanning diverse weather settings: overcast skies (596 shots), rainy conditions (758 shots), and sunny weather (1222 shots). Four tea cultivars are covered: Longjing 43 (768 shots), Zhongcha 108 (556 shots), white tea (572 shots), and yellow tea (680 shots). All tea bud instances were annotated manually. Image resolutions include 3840 × 2160, 3120 × 1440, and 1600 × 1200. Altogether, the dataset contains 2576 images, split at random into training and test folds following a 4:1 ratio. A selection of exemplar images is provided in **Figure 5a**. To guarantee labeling consistency and trustworthiness, the following practices were enforced:

(1) Annotation protocol definition: Before labeling commenced, a precise specification of what constitutes a tea bud was documented, including developmental stages and the minimum visible proportion required. Bounding boxes were required to hug the tea bud boundaries tightly, and any instance with occlusion exceeding 50% was disregarded.

(2) Iterative cross-validation: During the annotation campaign, a multi-round strategy was embraced. Initially, 30% of the image pool was annotated, after which the same subset underwent repeated labeling rounds to gauge self-agreement using Intersection over Union (IoU). An average IoU of 0.85 attests to the annotations' stability and reliability. After the remaining images were labeled, a randomly selected 10% slice of the corpus was reinspected to identify and correct potential bounding-box misplacements or category errors.

(3) Tool-aided refinement: The automatic edge-alignment function in LabelImg was used to reduce manual labeling imprecision.



(a) Self-built TBD dataset



(b) GWHD_2021 dataset

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Figure 5. Examples of self-built TBD dataset and GWHD_2021 dataset.

GWHD_2021 dataset. This resource was released in 2021 through the joint efforts of multiple global bodies, among them the French National Institute for Agricultural Research. It comprises 6515 images, partitioned into 3567 training, 1476 validation, and 1382 test samples. A few representative examples are shown in **Figure 5b**.

Evaluation criteria

A battery of standard indicators was used to quantify the detection performance of the presented technique: IoU (Intersection over Union), Precision, Recall, mAP (mean Average Precision), FLOPs (Floating-point Operations Per second), and overall parameter volume. In the object detection literature, a predicted bounding region is conventionally considered a successful localization when the IoU exceeds 0.5. Consistent with that convention, this investigation sets IoU = 0.5 as the acceptance threshold. The metrics are defined through the following relations:

$$P = \frac{TP}{TP+FP} \quad (30)$$

$$R = \frac{TP}{TP+FN} \quad (31)$$

$$AP = \int_0^1 P(R)dR \quad (32)$$

$$mAP = \frac{1}{K}\sum_{i=1}^{K} AP(i) \quad (33)$$

$$IoU = \frac{TP}{TP+FP+FN} \quad (34)$$

Within these expressions, P stands for Precision, R for Recall, and K for the total count of object categories. The symbols TP, FP, TN, and FN correspond to True Positive, False Positive, True Negative, and False Negative detections, respectively.

Parameter settings

The full model architecture was implemented, optimized, and tested using the PyTorch library. Regarding the computational environment, the experiments ran on a machine equipped with an 11th Gen Intel(R) Core(TM) i7–11,700 K processor operating at 3.60 GHz, together with a Quadro RTX 6000 GPU carrying 24 GB of video RAM and 32 GB of host memory.

Across all experimental runs, the YOLOv11 backbone was bootstrapped from parameters pretrained on the COCO corpus. The learning procedure relied on Stochastic Gradient Descent (SGD) as the optimizer, configured with a momentum factor of 0.937, weight decay of 0.0005, and an initial learning rate and a floor learning rate both set to 0.01, processing samples in mini-batches of 8. Every input image was resized to a canonical 640 × 640-pixel grid, and the optimization loop was run for 150 epochs.

The training data were enriched using two complementary pipelines—offline and online. The offline phase applied Rotation (through an angle of 45 degrees) and the injection of Gaussian-distributed noise (with zero mean and a standard

© 2022 The Author(s). | Published by Agro Intelligence Research Press

deviation of 0.1). The online phase, executed dynamically during training, comprised perturbations to the Hue channel (0.015), Saturation (0.7), and Value (0.4), supplemented by Mosaic augmentation (1.0), Horizontal Flipping (0.5), and random Scaling (0.5).

Comparison with the state-of-the-art methods

To empirically verify the strength of the proposed design, the TBD-Y-L, TBD-Y-M, and TBD-Y-S configurations were evaluated alongside 17 well-established detection paradigms drawn from the one-stage, two-stage, and Transformer-centric families, tested on both data sources. The one-stage representatives consisted of SSD [13], RetinaNet [14], YOLOv5 [22], YOLOv6 [9], YOLOv7 [11], YOLOv8 [12], and YOLOv11 [8]. The two-stage set comprised Faster R-CNN [15], Cascade R-CNN [23], Grid R-CNN [24], Dynamic R-CNN [25], and DetectoRS [17]. Finally, the Transformer-oriented category included DETR [26] and the Vision and Detection Transformer (ViDT) [27].

Quantitative Comparison of Detection Accuracy

The numerical head-to-head for both collections is summarized in **Tables 1 and 2**, where the highest and second-highest entries are highlighted in red and blue, respectively. A clear pattern emerges from these tables: CNN-grounded detectors hold a performance edge. Compared with Transformer-based solutions such as DETR and ViDT, the TBD-Y-L variant sets new benchmarks on both testbeds. On the privately assembled TBD corpus, its Precision, Recall, and mAP outpace the formerly dominant YOLOv11 model by 1.35%, 0.17%, and 2.18%, respectively, culminating in 78.64%, 82.87%, and 87.89%. Turning to the GWHD_2021 dataset, the same model leads YOLOv11-L by 0.7% in Precision, 0.7% in Recall, and 1.9% in mAP, reaching 92.4%, 88.2%, and 95.4%, respectively. Transformer architectures, for all their capacity to model extended spatial interactions and panoramic scene structure, generally impose heavier computational tolls and tend to demand voluminous training corpora before reaching peak effectiveness.

Table 1. Comparison results using different object detection methods on the self-built TBD dataset.

Method	Backbone	Year	mAP (%)	R (%)	P (%)
Faster R-CNN	ResNet-50	2016	79.24	82.99	73.66
Cascade R-CNN	ResNet-50	2018	79.01	82.27	74.42
Dynamic R-CNN	ResNet-50	2020	78.78	80.73	74.61
DetectoRS	ResNet-50	2021	80.16	81.66	**79.96**
SSD	ResNet-50	2016	74.28	70.14	75.17
YOLOv6-S	EfficientRep	2022	83.45	79.11	75.45
YOLOv6-M	EfficientRep	2022	84.44	78.89	75.76
YOLOv6-L	EfficientRep	2022	84.76	79.56	76.81
YOLOv7	CSPDarknet-53	2022	83.85	81.01	74.34
YOLOv8-S	Darknet-53	2023	84.20	78.77	76.14
YOLOv8-M	Darknet-53	2023	84.79	80.67	75.59
YOLOv8-L	Darknet-53	2023	85.14	82.15	77.04
YOLOv11-S	Darknet-53	2024	84.91	80.77	78.18
YOLOv11-M	Darknet-53	2024	85.23	81.77	77.54
YOLOv11-L	Darknet-53	2024	85.71	82.70	77.28

© 2022 The Author(s). | Published by Agro Intelligence Research Press

DETR	ResNet-50	2020	84.37	79.07	76.62
ViDT	Swin-Transformer	2022	84.63	82.09	76.82
Ours (TBD-Y-S)	Darknet-53	2024	86.98	81.67	77.92
Ours (TBD-Y-M)	Darknet-53	2024	**87.44**	82.20	78.35
Ours (TBD-Y-L)	Darknet-53	2024	**87.89**	**82.87**	**78.64**

Table 2. Comparison results using different object detection methods on the GWHD_2021 dataset.

Method	Backbone	Year	mAP (%)	R (%)	P (%)
Faster R-CNN	ResNet-50	2016	85.0	57.5	73.2
Cascade R-CNN	ResNet-50	2018	85.8	58.5	74.1
Grid R-CNN	ResNet-50	2019	84.5	61.7	73.8
DetectoRS	ResNet-50	2021	87.5	62.4	77.9
RetinaNet 50	ResNet-50	2017	84.1	57.3	67.3
YOLOv5-L	CSPDarknet-53	2021	89.6	81.6	91.0
YOLOv6-L	EfficientRep	2022	92.7	87.0	**92.3**
YOLOv7	CSPDarknet-53	2022	92.2	87.6	91.2
YOLOv8-S	Darknet-53	2023	91.9	85.5	90.7
YOLOv8-M	Darknet-53	2023	92.6	86.0	91.6
YOLOv8-L	Darknet-53	2023	93.0	86.7	91.9
YOLOv11-S	Darknet-53	2024	92.3	86.2	91.1
YOLOv11-M	Darknet-53	2024	92.7	86.4	91.2
YOLOv11-L	Darknet-53	2024	93.5	87.5	91.7
DETR	ResNet-50	2020	88.6	83.5	86.4
ViDT	Swin-Transformer	2022	90.2	85.8	88.6
Ours (TBD-Y-S)	Darknet-53	2024	94.3	87.2	91.9
Ours (TBD-Y-M)	Darknet-53	2024	**94.9**	**87.9**	91.9
Ours (TBD-Y-L)	Darknet-53	2024	**95.4**	**88.2**	**92.4**

Compared with YOLOv11-L, both YOLOv11-S and YOLOv11-M operate with far leaner parameter sets and lighter computational budgets. Evaluated on the in-house TBD dataset, TBD-Y-M posts a precision of 78.35%, a recall of 82.20%, and an mAP of 87.44%, exceeding YOLOv11-M by increments of 0.81%, 0.43%, and 2.21%. TBD-Y-S achieves Precision, Recall, and mAP scores of 77.92%, 81.67%, and 86.98%—representing a negligible Precision regression of 0.26% yet improvements in Recall and mAP of 0.9% and 2.07% contrasted with YOLOv11-S. On the GWHD_2021 collection, TBD-Y-M delivers Precision of 91.9%, Recall of 87.9%, and mAP of 94.9%, surpassing YOLOv11-M by margins of 0.7%, 1.5%, and 2.2%. Likewise, TBD-Y-S achieves 91.9%, 87.2%, and 94.3%, beating YOLOv11-S by 0.8%, 1.0%, and 2.0%, respectively.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Comparison of Model Parameters and Computational Complexity

A full breakdown of parameter sizes and computational expenditures across the tested methods appears in **Table 3**, with FLOPs serving as the complexity yardstick. The figures reveal that, relative to YOLOv11-L, the proposed TBD-Y-L secures a 2.18 % mAP enhancement at the expense of an extra 1.93 M parameters and 7.9 GFLOPs. In the same vein, TBD-Y-M gains 2.21 % mAP over YOLOv11-M while adding 1.92 M parameters and 2.9 GFLOPs, capturing the second-best accuracy overall. TBD-Y-S, when set against YOLOv11-S, improves mAP by 2.07 % with 1.81 M more parameters and 1.55 GFLOPs of added computation.

Table 3. Comparison of model parameters and computational complexity.

Method	Year	Backbone	FLOPs (G)	Parameters (M)	mAP (%)
Faster R-CNN	2016	ResNet-50	216.30	41.12	79.24
Cascade R-CNN	2018	ResNet-50	244.10	68.93	79.01
DetectoRS	2021	ResNet-50	239.86	123.2	80.16
SSD	2016	ResNet-50	87.72	24.39	74.28
YOLOv5-S	2021	CSPDarknet-53	15.80	7.01	80.01
YOLOv5-M	2021	CSPDarknet-53	47.86	20.85	81.39
YOLOv5-L	2021	CSPDarknet-53	107.60	46.11	81.73
YOLOv6-S	2022	EfficientRep	44.16	17.21	83.45
YOLOv6-M	2022	EfficientRep	82.01	34.23	84.44
YOLOv6-L	2022	EfficientRep	143.79	58.47	84.76
YOLOv7	2022	CSPDarknet-53	103.20	36.48	84.20
YOLOv8-S	2023	Darknet-53	28.44	11.13	84.20
YOLOv8-M	2023	Darknet-53	78.69	25.84	84.79
YOLOv8-L	2023	Darknet-53	164.82	43.61	85.14
YOLOv11-S	2024	Darknet-53	21.3	9.43	84.91
YOLOv11-M	2024	Darknet-53	68.2	20.01	85.23
YOLOv11-L	2024	Darknet-53	86.60	25.28	85.71
DETR	2020	ResNet-50	86.0	41	84.37
ViDT	2022	Swin-Transformer	–	100	84.63
Ours (TBD-Y-S)	2024	Darknet-53	22.85	11.24	86.98
Ours (TBD-Y-M)	2024	Darknet-53	71.10	21.93	**87.44**
Ours (TBD-Y-L)	2024	Darknet-53	94.50	27.21	**87.89**

Detectors founded on Transformer designs routinely pack considerably larger parameter volumes. A case in point is ViDT, whose parameter count is roughly quadruple that of YOLOv11-L, while its mean Average Precision (mAP) trails YOLOv11-L by 1.08 %. In stark contrast, the lightweight TBD-Y-S proposed in this study significantly reduces both the parameter budget and computational load when benchmarked against YOLOv11-L, while preserving competitive

© 2022 The Author(s). | Published by Agro Intelligence Research Press

detection fidelity. Quantitatively, its mAP outstrips YOLOv11-L by 1.27%, yet its parameter footprint is merely 44%, and its FLOPs are only 26% of YOLOv11-L's figures—delivering an outstanding balance between recognition accuracy and practical efficiency.

Visualization Comparison of Detection Results

Figure 6 presents side-by-side visual comparisons of detection outputs from several competing frameworks, featuring the two-stage schemes Faster R-CNN and DetectoRS alongside the one-stage schemes YOLOv8-L and YOLOv11-L. A visual inspection confirms that, despite challenging, heavily textured surroundings, the methodology described in this paper demarcates tea buds with distinctly higher fidelity. The advantage is especially pronounced for diminutive targets, where the proposed solution substantially curtails both phantom predictions and overlooked instances, phenomena that plague alternative detectors.

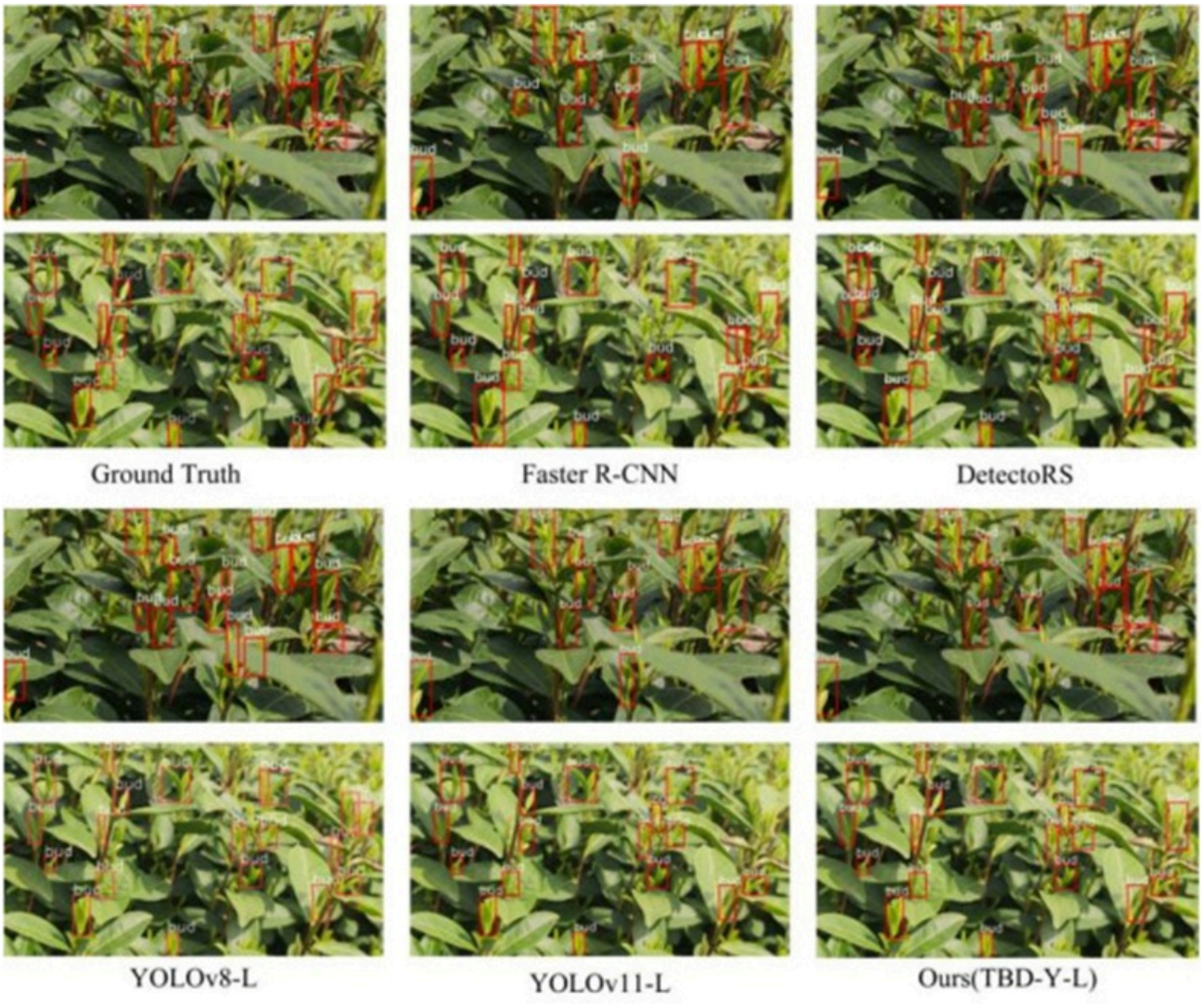


Figure 6. Visualization of detection results using different methods.

Ablation studies

To verify the contribution of each proposed component, a set of ablation experiments was executed on the custom-built TBD dataset. The findings are laid out in the following subsections.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Ablation studies of different components

Table 4 reports the ablation outcomes when the suggested SOSA and GAGFF are independently plugged into the framework. The starting point for all comparisons is the vanilla YOLOv11-L model.

Table 4. Ablation studies on the self-built TBD dataset.

Baseline	GAGFF	SOSA	FLOPs (G)	Parameters (M)	mAP (%)
√			86.60	25.28	85.71
√	√		89.10	27.15	87.26
√		√	94.20	25.38	87.39
√	√	√	94.50	27.21	87.89

From **Table 4**, it can be seen that adding GAGFF alone lifts mAP by 1.55 percentage points, whereas adding SOSA alone increases mAP by 1.68 percentage points. The respective parameter increases are modest, at 1.87 M and 0.1 M. In the case of SOSA, the observed gains can be traced to its redirection of the model's focal attention to tea bud areas while filtering out irrelevant background clutter. Beyond that, the LCA unit in SOSA can efficiently harvest local contextual patterns, yielding even sharper spatial pinpointing of individual buds.

Bringing both GAGFF and SOSA into the architecture simultaneously produces an even larger boost, elevating mAP by 2.18 percentage points over the vanilla baseline. This outcome clearly validates that the two modules work in complementary ways to improve detection quality.

To present a more intuitive, visual account of how these modules reshape the model's internal representations, XGrad-CAM [28] was applied to generate activation heatmaps before and after each module was introduced. These heatmaps are shown in **Figure 7**. The visualizations confirm that both GAGFF and SOSA shift the model's attention distribution toward tea bud regions, consistent with the quantitative improvements observed.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

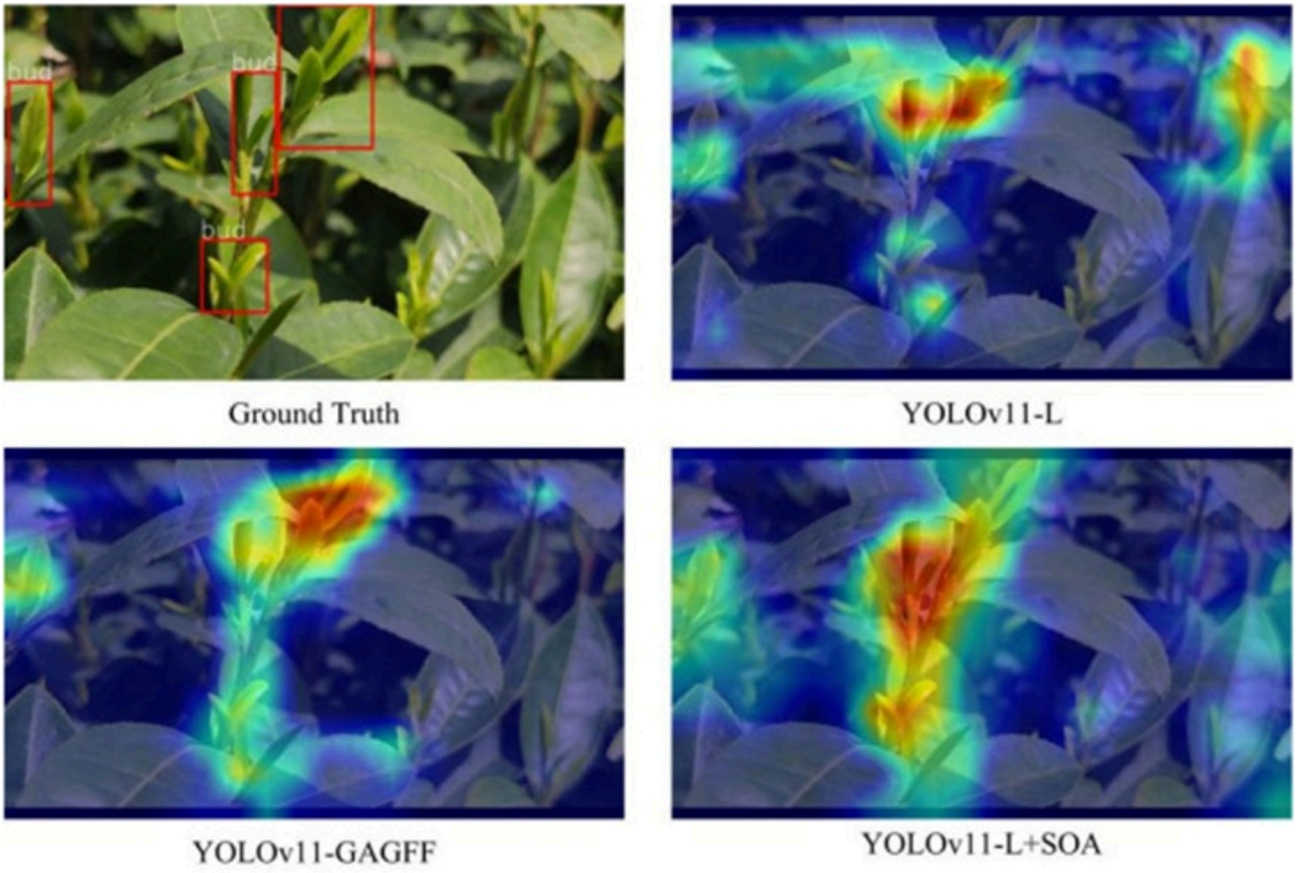


Figure 7. Heatmaps of the GAGFF and SOSA on the self-built TBD dataset.

Effect of LCA on detection performance

The newly designed LCA was benchmarked against several widely used attention modules, specifically the Convolutional Block Attention Module (CBAM) [29], Efficient Channel Attention (ECA) [30], and Shuffle Attention (SA) [31]. The performance figures gathered on the in-house TBD dataset are collated in **Table 5**, once again using YOLOv11-L as the common base.

Table 5. Comparison of LCA with other attention mechanisms.

Attention mechanism	+Parameters-to-Baseline (M)	mAP (%)
Baseline	0	85.71
Baseline + CBAM	0.4566	86.88
Baseline + Shuffle Attention (SA)	0.0275	86.78
Baseline + ECA	0.0185	87.04
Baseline + Our Attention (LCA)	**0.0125**	**87.39**

Inspection of **Table 5** reveals that LCA delivers stronger detection scores than all three competing attention mechanisms. The primary explanation lies in LCA's superior ability to accurately localize the spatial extent of tea bud regions.

Effect of kb value on the performance of LCA and GAGFF

© 2022 The Author(s). | Published by Agro Intelligence Research Press

The LCA module and the GAGFF module both rely on depth-wise separable strip convolutions, where the kernel span kb acts as a critical hyperparameter. To understand how this parameter affects each module's behavior, a series of experiments was conducted to identify the optimal setting.

Effect of kb Value on the Performance of LCA. Five different assignments for kb—namely 3, 5, 7, 9, and 11—were trialed. The resulting performance figures on the proprietary TBD dataset are summarized in **Table 6**. A clear trend emerges: detection quality peaks when kb is set to 11. Hence, throughout this work, kb=11 was adopted as the LCA configuration.

Table 6. Effect of kb value on the performance of LCA.

Empty Cell	mAP (%)
3	87.08
5	86.71
7	86.97
9	87.21
11	**87.39**

Effect of kb Value on the Performance of GAGFF. The identical candidate set of kb values 3, 5, 7, 9, and 11 was examined for the GAGFF module on the same dataset. The recorded outcomes are compiled in **Table 7**. Here too, kb=11 yields the topmost detection performance, which motivated the choice of kb=11 within GAGFF as well.

Table 7. Effect of kb value on the performance of GAGFF.

Empty Cell	mAP (%)
3	87.01
5	87.03
7	86.89
9	87.07
11	**87.26**

Effect of loss functions on the performance of SOSA

The impact of the supervisory loss function used in the SOSA module was also empirically tested. On the custom TBD dataset, several candidate losses—L1, CE, L2, and BCE—were used to constrain the SOSA training signal. The numerical comparison is displayed in **Table 8**.

Table 8. Effect of different loss functions on SOSA performance.

Loss Function	mAP (%)
L1 loss	87.11
L2 loss	86.93
CE loss	86.88

© 2022 The Author(s). | Published by Agro Intelligence Research Press

BCE Loss	**87.39**

The results indicate that BCE loss consistently outperforms L1, L2, and CE losses in terms of final detection accuracy. This finding motivated the exclusive use of BCE loss for network optimization in this paper.

Conclusion

This study has proposed a TBD solution that balances efficiency and high accuracy. Drawing on a comprehensive body of experimental evidence, several conclusions can be stated:

(1) The proposed SOSA mechanism markedly enriches feature representational capacity, attenuates background disturbance, and thereby raises overall detection accuracy.

(2) The newly introduced LCA module effectively amplifies local contextual signals and sharpens the spatial localization of tea bud regions.

(3) The suggested GAGFF strategy successfully fuses semantically rich, high-level information with detail-oriented, low-level cues, reconciling feature maps that differ in both scale and semantic character, thereby strengthening their representational quality.

Experimental validation on both the privately collected dataset and the public benchmark confirms that the TBD-Y-L variant sets a new state-of-the-art performance level relative to previously published methods. Meanwhile, the compact TBD-Y-S variant surpasses YOLOv11-L in detection accuracy while operating with substantially leaner parameters and computational budgets. These properties render the methodology highly suitable for integration into an intelligent tea-harvesting apparatus.

Even with the strong detection fidelity achieved, certain obstacles remain unresolved:

(1) The in-house TBD dataset currently contains only 2576 images despite applying both offline and online augmentation techniques to inflate the training volume artificially; a genuine data shortage persists, which hampers generalization and limits the ceiling on achievable accuracy.

(2) The parameter footprint and computational cost must still be reduced further. In their present form, they may fall short of satisfying the stringent constraints imposed by real-world deployment scenarios.

In upcoming work, plans are in place to acquire and annotate a substantially larger collection of tea bud imagery, thereby significantly scaling up the proprietary TBD dataset. In parallel, efforts will be channeled into engineering a more lightweight network architecture that preserves high detection fidelity while being inherently more amenable to embedding within operational tea-picking hardware.

Acknowledgements

This work in this paper was supported by the Beijing Natural Science Foundation under Grant (L211017), the Technology System of Modern Agricultural Industry in Shandong Province (SDAIT19), the Key R&D Projects in Zhejiang Province (2023C02043), and the Research start-up funds-TRI-SAAS (CXGC2023G33).

Conflict of interest

None

© 2022 The Author(s). | Published by Agro Intelligence Research Press

Financial support

None

Ethics statement

None

Received: 02 Oct 2021 Revised: 07 Dec 2021 Accepted: 24 Dec 2021
Published online: 25 May 2022

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Girolametti F, Annibaldi A, Illuminati S, Damiani E, Carloni P, Truzzi C. Essential and potentially toxic elements (PTEs) content in European tea (Camellia sinensis) leaves: risk assessment for consumers. Molecules. 2023;28.

2. Chen T, Li H, Chen J, Zeng Z, Han C, Wu W. Detection network for multi-size and multi-target tea bud leaves in the field of view via improved YOLOv7. Comput Electron Agric. 2024;218.

3. Chen Y-T, Chen S-F. Localizing plucking points of tea leaves using deep convolutional neural networks. Comput Electron Agric. 2020;171.

4. Li Y, Ma R, Zhang R, Cheng Y, Dong C. A tea buds counting method based on YOLOv5 and Kalman filter tracking algorithm. Plant Phenomics. 2023;2023.

5. Lu J, Luo H, Yu C, Xiao L, Huang J, Wu H, et al. Tea bud DG: A lightweight tea bud detection model based on dynamic detection head and adaptive loss function. Comput Electron Agric. 2024;227.

6. Xu W, Zhao L, Li J, Shang S, Ding X, Wang T. Detection and classification of tea buds based on deep learning. Comput Electron Agric. 2022;192.

7. Bochkovskiy A, Wang C-Y, Liao H-YM. YOLOv4: Optimal speed and Accuracy of Object detection, ArXiv, abs/2004.10934. 2020.

8. Khanam R, Hussain M.YOLOv11: An overview of the key architectural enhancements, arXiv preprint arXiv:2410.17725. 2024.

9. Li C, Li L, Jiang H, Weng K, Geng Y, Li L, et al. YOLOv6: A single-stage object detection framework for industrial applications, ArXiv, abs/2209.02976. 2022.

10. Redmon J, Farhadi A. YOLOv3: An incremental improvement, ArXiv, abs/1804.02767 .2018 .

11. Wang C-Y, Bochkovskiy A, Liao H-YM. YOLOv7: Trainable bag-of-freebies sets new State-of-the-art for real-time object detectors. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2022:7464-75.

12. Varghese R, Sambath M. YOLOv8: A novel object detection algorithm with enhanced performance and robustness. 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). 2024. https://doi.org/10.1109/ADICS58448.2024.10533619.

13. Liu W, Anguelov D, Erhan D, Szegedy C, Reed SE, Fu C-Y, Berg AC. SSD: single shot MultiBox detector. European Conference on Computer Vision. 2015.

14. Lin T-Y, Goyal P, Girshick RB, He K, Dollár P. Focal loss for dense object detection. IEEE Trans Pattern Anal Mach Intell. 2017;42:318-27.

15. Ren S, He K, Girshick RB, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2015;39:1137-49.

16. He K, Gkioxari G, Dollár P, Girshick RB. Mask R-CNN. 2017.

17. Qiao S, Chen L-C, Yuille AL. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2020:10208-19.

18. Li J, Li J, Zhao X, Su X, Wu W. Lightweight detection networks for tea bud on complex agricultural environment via improved YOLO v4. Comput Electron Agric. 2023;211.

19. Chen Z, Zhou H, Lin H, Bai D. TeaViTNet: tea disease and pest detection model based on fused multiscale attention. Agronomy. 2024;14.

20. Cai X, Lai Q, Wang Y, Wang W, Sun Z, Yao Y. Poly Kernel inception Network for remote sensing detection. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024:27706-16.

21. David E, Serouart M, Smith D, Madec S, Velumani K, Liu S, et al. Global Wheat Head Detection 2021: an improved dataset for benchmarking wheat head detection methods. Plant Phenomics. 2021;2021.

22. Jocher G. ultralytics/yolov5: v3.1 - bug fixes and performance improvements. 2020.

23. Cai Z, Vasconcelos N. R-CNN Cascade, Delving into high quality object detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2017:6154-62.

24. Lu X, Li B, Yue Y, Li Q, Yan J. Grid R-CNN. 2018.

25. Zhang H, Chang H, Ma B, Wang N, Chen X. Dynamic R-CNN: towards high quality object detection via Dynamic Training. ArXiv. 2020;abs/2004.06002.

26. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. End-to-End object detection with Transformers, ArXiv, abs/2005.12872. 2020.

27. Song H, Sun D, Chun S, Jampani V, Han D, Heo B, et al. Vidt: an efficient and effective fully transformer-based object detector. arXiv preprint arXiv:2110.03921. 2021.

28. Fu R, Hu Q, Dong X, Guo Y, Gao Y, Li B. Axiom-based Grad-CAM: towards accurate visualization and explanation of CNNs. arXiv preprint arXiv:2008.02312. 2020.

29. Woo S, Park J, Lee J-Y, Kweon I-S. CBAM: convolutional block Attention Module, ArXiv, abs/1807.06521. 2018.

30. Wang Q, Wu B, Zhu PF, Li P, Zuo W, Hu Q. ECA-net: efficient channel attention for deep convolutional neural networks. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2019:11531-9.

© 2022 The Author(s). | Published by Agro Intelligence Research Press

31. Zhang Q-L, Yang Y. SA-Net: shuffle attention for deep convolutional neural networks. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021:2235-9.

© 2022 The Author(s). | Published by Agro Intelligence Research Press