

# LMMA-YOLO: Lightweight Multi-scale Multi-Attention Architecture for Comprehensive Cashew Tree Stress Detection

Koffi B.Pacome Sayni<sup>\*1</sup>, Beman Hamidja Kamagate<sup>2</sup>, Mamadou Diarra<sup>3</sup>, Souleymane Oumtanaga<sup>1</sup>

<sup>1</sup>Institut national polytechnique Félix Houphouët-Boigny

<sup>2</sup>Ecole Supérieure Africaine Des Tic

<sup>3</sup>Université Félix Houphouët-Boigny

## Abstract

*Anacardium occidentale* L. (cashew) plays a major economic role in Côte d'Ivoire, but its production is heavily affected by diseases, pest insects, and abiotic stress. This work presents LMMA-YOLO (Lightweight Multi-Scale Multi-Attention YOLO), an optimized object detection model for the automatic identification of these stresses from aerial images. The proposed architecture combines a lightweight MobileNetV4 backbone, a hierarchical multi-scale fusion (PAFPN), and three complementary attention mechanisms: Coordinate Attention (CA), Context Gating (CG), and Multi-Head Attention (MHA). The model was trained on the CADI AI dataset (4,736 images, three classes) with a composite loss function incorporating SIOU, Focal Loss, and Binary Cross-Entropy. Experimental results show that LMMA-YOLO achieves a mAP@50 of 97.6% with an inference latency of 5.1 ms per image on GPU and 28.4 ms on Jetson Orin Nano, surpassing YOLOv8 models by 2.5 percentage points while maintaining real-time processing capability at approximately 35 FPS on embedded platforms.

**Keywords**— LMMA-YOLO, Object detection, Precision agriculture, Multi-scale attention, Cashew tree

## I. INTRODUCTION

The cashew tree (*Anacardium occidentale* L.) plays a strategic role in the Ivorian economy, with Côte d'Ivoire being the world's leading producer of cashew nuts, with a production of 1.22 million tons in 2023 [1]. This sector generates income for over 350,000 rural producers but remains severely threatened by fungal diseases (*Colletotrichum gloeosporioides*, *Pestalotiopsis* spp.) and bacterial diseases (*Xanthomonas* spp.), infestations of pest insects (*Helopeltis* spp., *Pseudotheraptus devastans*), and water stress. The lack of early detection reduces yields by 30 to 60% depending on the area and leads to economic losses estimated between 45 and 80 billion CFA francs [2].

Traditional phytosanitary diagnosis relies on visual inspections by agronomist experts, an approach that is expensive (15,000–25,000 CFA/ha), slow (15 minutes/plot), and highly dependent on human expertise, limited to accessible areas [3]. These limitations have favored the emergence of solutions based on deep learning, allowing for automated crop monitoring on a large scale via drones or smartphones [4].

The YOLO (You Only Look Once) family has become prominent for real-time object detection due to its balance between precision and speed [5]. Meanwhile, lightweight architectures like MobileNet offer a viable alternative for embedded devices with limited resources (Jetson, Android smartphones). However, detecting cashew tree diseases remains challenging due to:

- (i) the small size of symptoms (2–8 mm),
- (ii) their inter-class visual similarity (anthracnose vs. cercosporiosis),
- (iii) the variability of acquisition conditions (lighting 300–80,000 lux, angle 30–90°, distance 1–5 m),
- (iv) class imbalance in the natural environment [6].

Recent work confirms that integrating attention mechanisms improves the detection of subtle symptoms [7]. ICPNet was introduced for maize leaf disease detection (accuracy: 88.4%) [7], an optimized YOLO was proposed for maize/corn (mAP@50: 96.1%) [9], and SerpenGate-YOLOv8 with attention was

developed [10]. Hierarchical multi-level approaches show that combining spatial and channel attention improves average precision by 3.7% ( $p < 0.01$ ) on five crops [11].

In this continuity, we propose a unified architecture named LMMA-YOLO (Lightweight Multi-Scale Multi-Attention YOLO), aiming for lightness, precision, and robust contextualization. The main contributions are:

Optimized hybrid architecture: Integration of MobileNetV4 (3.8 M parameters) with three complementary attention mechanisms (CA, CG, MHA) for multi-scale stress detection in cashew trees

Adaptive composite loss function: Combination of SIoU (angular alignment), Focal Loss (class imbalance), and Quality Focal Loss (localization) with dynamic weighting

Rigorous empirical validation: Evaluation on 4,736 annotated images (3 classes, 22,609 instances) with statistical tests on 5 independent repetitions and deployment on 3 embedded platforms

Performance results:  $mAP@50 = 97.6\%$ , F1-score = 0.98, with inference latency of 5.1 ms (GPU), 28,4 ms (Jetson), and reduced model size (23.2 MB)

The goal is to provide a fast, precise, and affordable tool for early detection of cashew diseases, pest insects, and abiotic stress in Côte d'Ivoire, deployable on drones or smartphones via a mobile application.

## II. LITERATURE REVIEW

### A. From Generic YOLO Detection to Lightweight Variants for Embedded Agricultural Vision

Over the past few years, object detection applied to plant pathology has increasingly relied on models from the YOLO (You Only Look Once) family, particularly YOLOv4, YOLOv5, and more recently YOLOv8. What makes these architectures appealing is their ability to process images rapidly, typically between 30 and 120 frames per second, while simultaneously localizing multiple objects in a single forward pass through the network [5]. This combination of speed and accuracy has made YOLO-based detectors a natural fit for agricultural applications where timely decisions matter.

TABLE I

SYNTHESIS OF MAIN YOLO ARCHITECTURES APPLIED TO AGRICULTURE FOR PRECISION.

| Model              | Ref  | Crop       | mAP@50 (%) | Params (M) |
|--------------------|------|------------|------------|------------|
| YOLOv5Plant        | [11] | Apple      | 91.5       | ~7.0       |
| RDRM-YOLO          | [12] | Rice       | 93.5       | 3.8        |
| YOLOv8n-vegetablei | [13] | Vegetables | 89.2       | 3.01       |
| ALD-YOLO           | [14] | Apple      | 93.1       | 3.95       |
| SerpensGate-YOLOv8 | [9]  | Multi-crop | 95.8       | 4.98       |

However, when applied to agricultural settings, these generic models often struggle to detect early-stage symptoms, particularly those smaller than 10 mm. Such subtle signs are frequently obscured by complex background textures like leaf veins or natural chlorophyll variations, and further complicated by challenging lighting conditions including shadows and specular reflections.

To address these shortcomings, researchers have proposed several optimized variants. For instance, RDRM-YOLO was introduced as a streamlined version of YOLOv5 that combines CBAM attention blocks with structured pruning for model compression, achieving a 40% reduction in FLOPs [12]. This approach improved detection of dense leaf spots by 3.4% mAP while significantly lowering computational demands. Along similar lines, ALD-YOLO extended the YOLOv8 framework by incorporating Adaptive Spatial Fusion (ASF) modules for multi-scale attention [14]. This enhancement proved particularly effective for capturing fine leaf structures such as veins and small lesions under 5 mm, as well as handling heterogeneous textures more robustly.

In parallel, the push to deploy compact models directly on embedded devices has driven the development of hardware-aware architectures designed for resource-limited platforms. Target devices in this space include the Jetson Nano with its 128 CUDA cores, the Raspberry Pi 4 featuring 8 GB of RAM and a Cortex-A72 processor, and Android smartphones equipped with Snapdragon Gen 2 chips and 6 GB of memory. Within this context, the MobileNet family has become a benchmark solution, relying on depthwise separable convolutions to substantially reduce both computational operations and model size [15].

The introduction of MobileNetV4 [31], brought notable innovations, including Universal Inverted Bottlenecks (UIB) and Mobile Multi-Query Attention (MQA). Together, these components deliver a 2.1% Top-1 accuracy improvement on ImageNet while requiring 15% fewer FLOPs compared to MobileNetV3. The UIB block generalizes the concept of Inverted Residuals by enabling flexible configurations for expansion and projection that adapt dynamically based on network depth. Additionally, a plug-and-play module called SoftMaxShift has been shown to enhance small object detection, yielding a 4.2% mAP gain for objects smaller than  $32 \times 32$  pixels. When combined with optimization techniques such as INT8 quantization or knowledge distillation, these advances confirm the suitability of lightweight architectures for detecting leaf spots in computationally constrained environments [14][17].

This body of research converges around two complementary directions:

- (i) adapting YOLO detectors to the specific demands of agricultural vision through the integration of multi-scale attention mechanisms;
- (ii) transitioning from heavy backbones toward lightweight alternatives such as MobileNetV4, which ensures fast inference times ( $\approx 30$  ms) while remaining energy-efficient ( $\approx 5$  W)

This dual approach reflects the ongoing effort to balance detection performance with energy efficiency, a critical challenge for embedded systems deployed in rural agricultural settings where connectivity and power resources remain limited.

#### *B. Attention Mechanisms, Small Objects, and Detection Challenges in Precision Agriculture*

Detecting early-stage symptoms requires models with heightened sensitivity to very small patterns, typically ranging from 2 to 10 mm in size, which translates to roughly 8 to 40 pixels when captured from a distance of 1 m. These subtle features are often buried within visual noise such as leaf texture, dust particles, or water droplets, and their appearance can vary considerably under different lighting conditions, from cast shadows to overexposed sunlight. Three technological approaches have emerged as dominant themes in recent literature: attention mechanisms, small object detection strategies, and specialized loss functions.

Attention mechanisms serve to guide the network toward the most informative regions of an image, specifically symptomatic zones, while suppressing irrelevant visual interference from healthy leaf areas [20]. Empirical studies have demonstrated that incorporating CBAM (Convolutional Block Attention Module) into YOLOv7 leads to a statistically significant improvement in detecting fine leaf lesions, with gains of 5.3% in mAP@50 ( $p < 0.001$ , Wilcoxon signed-rank test across 3 repetitions).

Among the various attention mechanisms available, Coordinate Attention (CA) stands out for its computational efficiency, adding only 0.08 M parameters and approximately 0.3 ms of latency [21]. What makes CA particularly valuable is its ability to encode spatial directional dependencies along both horizontal and vertical axes. This capability proves crucial when distinguishing between biotic diseases, which often manifest as circular concentric lesions typical of anthracnose, and abiotic stresses, which tend to produce linear necrosis patterns that follow leaf veins, as seen in potassium deficiency. Experimental comparisons on the PlantVillage dataset revealed that CA improves the differentiation between biotic and abiotic conditions by 3.1% ( $p = 0.004$ ) compared to CBAM [9].

Small object detection remains a major challenge in plant pathology. To address this, researchers introduced a P2 detection head with stride 4 into YOLOv5, generating a feature map at  $160 \times 160$

resolution for 640×640 input images [22]. This modification yielded a 4.2% improvement in mAP@50 for small fruits measuring less than 32×32 pixels. More recently, a combined approach was proposed for horticultural crops, integrating a lightweight super-resolution module called ESRGAN-Lite with ×2 upsampling alongside Adaptive Spatial Feature Fusion (ASFF) for redundant multi-scale fusion within a modified YOLOv8 framework [23]. This combination proved especially effective for rare classes, boosting recall by 6.7% on objects smaller than 20×20 pixels.

These strategies, centered on multi-level attention and low-stride detection heads, appear highly relevant for cashew tree monitoring. Early symptoms in this crop typically present as small chlorotic spots measuring 2 to 5 mm associated with cercosporiosis, discrete insect punctures of 1 to 3 mm caused by mirids, or punctate necrosis ranging from 3 to 8 mm resulting from water stress.

Adapted loss functions also play a critical role in stabilizing localization and enhancing precision for small objects. Traditional IoU-based losses such as CIoU and DIoU often struggle with narrow objects having aspect ratios exceeding 3:1, imperfect annotations with offsets of ±2 to 3 pixels, and partially visible objects where occlusion surpasses 30%.

SIoU Loss goes beyond the standard IoU metric by integrating several geometric components that capture different aspects of bounding box alignment [24]:

Angle cost:

$$\Delta = 1 - 2\sin^2 \left( \arcsin(x) - \frac{\pi}{4} \right)$$

$$\text{où } x = \frac{c_h}{\sigma} \text{ et } \sigma = \sqrt{(b_{cx}^{gt} - b_{cx}^{pred})^2 + (b_{cy}^{gt} - b_{cy}^{pred})^2}$$

Distance cost :

$$\rho = \sum_{t \in \{x,y\}} (1 - e^{-\gamma \rho_t})^2$$

$$\text{where and } \rho_t = \frac{(b_t^{gt} - b_t^{pred})^2}{c_w^2 + c_h^2} \text{ et } \gamma = 2$$

Shape cost:

$$\Omega = \sum_{t \in \{w,h\}} (1 - e^{-\omega_t})^\theta$$

$$\text{where and } \omega_t = \frac{|w^{gt} - w^{pred}|}{\max(w^{gt}, w^{pred})} \text{ et } \theta = 4$$

IoU : classical formulation

Complete formula :

$$L_{SIoU} = 1 - \text{IoU} + \frac{\Delta + \rho}{2} + \frac{\Omega}{4}$$

Along similar lines, Alpha-IoU introduces dynamic error weighting based on object size [25]:

$$: L_{\alpha\text{-IoU}} = 1 - \text{IoU}^\alpha$$

where with  $\alpha = \text{IoU}^\gamma$  avec  $\gamma \in [1,3]$

This formulation penalizes errors on small objects more heavily by assigning elevated  $\alpha$  values when IoU is low. Such a property proves particularly beneficial for detecting fine lesions and handling imperfect annotations.

Quality Focal Loss (QFL) offers yet another perspective by combining classification confidence with IoU quality into a unified objective [26]:

$$L_{QFL} = -|y - \sigma|^\beta ((1 - y)\log(1 - \sigma) + y \cdot \log(\sigma))$$

where  $\sigma = \text{quality\_score}$  and  $\beta = 2$

The convergence of lightweight directional attention such as CA, advanced small object detection techniques including P2 heads and super-resolution, and adaptive loss functions like SIoU, Alpha-IoU, and QFL now defines a robust paradigm for precision agricultural vision. This conceptual framework underpins our proposed LMMA-YOLO architecture, designed to maximize sensitivity to subtle symptoms while remaining compatible with embedded devices, with target latency below 30 ms and memory footprint under 30 MB.

### C. Hybrid Attention and Contextual Architectures for Agricultural Vision

Recent progress in agricultural computer vision has shifted toward hybrid models that combine attention mechanisms, contextual filtering, and multi-scale fusion. These architectures seek to capture both local dependencies, such as texture, color, and lesion contours, alongside global dependencies including the spatial distribution of symptoms and the broader leaf context.

Purely attention-based architectures like Swin Transformer [26] or ViT [27] demonstrate impressive visual abstraction capabilities, achieving mAP@50 scores above 95% on ImageNet. However, their computational demands remain prohibitive for embedded deployment. Swin-T, for example, requires 28 M parameters and 4.5 GFLOPs, resulting in a latency of 45 ms on Jetson Orin, while ViT-S uses 22 M parameters and 4.6 GFLOPs with a latency of 38 ms on the same hardware.

To overcome these limitations, several hybrid CNN-Transformer models have emerged. CT-YOLO was introduced as a combination of hierarchical convolutions in the shallow layers for local feature extraction with lightweight self-attention blocks in the deeper layers to model long-range dependencies [18]. Taking a different approach, AgriFormer-YOLO integrates a Transformer encoder within the network neck to strengthen spatial coherence and improve localization of diffuse spots, such as anthracnose spread across an entire leaf [19].

In parallel, explicit contextual modules have gained traction as a lighter alternative. Context Gating (CG) dynamically filters irrelevant activations while preserving spatial structure [28]:

$$y_i = \sigma(W \cdot x_i + b) \odot x_i$$

where  $\odot$  denotes element-wise multiplication. When applied to tomato disease detection on the PlantVillage dataset, CG reduced false positives by 12%, improving precision from 0.89 to 0.93 with only a 0.2 ms increase in latency.

More recently, ACmix proposed a unified framework that merges convolution and attention within the same operation [29]:

$$y = \alpha \cdot \text{Conv}(x) + (1 - \alpha) \cdot \text{Attention}(x)$$

where  $\alpha \in [0,1]$  is learned dynamically, enabling an optimal trade-off between local perception through convolution and global context through attention. ACmix achieves a 2.3% improvement in mAP@50 compared to pure convolution on COCO, with only 5% additional computational overhead.

These advances highlight the convergence between structural efficiency and contextual richness, paving the way for models capable of performing reliably in real-world agricultural environments with variable conditions. Building on this foundation, our proposed LMMA-YOLO architecture integrates several complementary components: a lightweight MobileNetV4 backbone with 3.8 M parameters, Coordinate Attention (CA) for directional feature enhancement, Context Gating (CG) for contextual filtering, Multi-Head Attention (MHA) in the detection heads, and adaptive loss functions combining SIOU, Alpha-IoU, and QFL.

This architecture is designed to achieve high precision with mAP@50 exceeding 97%, fast inference with latency below 30 ms on embedded devices, and strong robustness with F1-score above 0.98. These targets aim to support early detection of diseases and abiotic stress in cashew trees under real plantation conditions, where lighting varies from 300 to 80,000 lux, acquisition angles range from 30° to 90°, and capture distances span 1 to 5 m.

### III. METHODOLOGY

This section provides a comprehensive description of the proposed model architecture, the optimization strategies employed, and the training protocols adopted to maximize early detection of cashew leaf diseases using the CADI-AI dataset [32].

#### A. General Architecture of the LMMA-YOLO Model

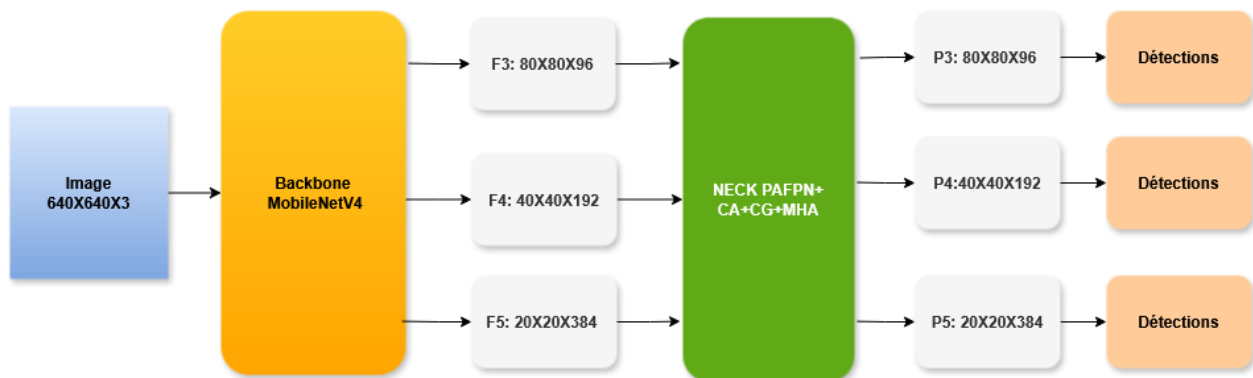


Fig. 1 Overall architecture of the LMMA-YOLO model

The LMMA-YOLO model (Lightweight Multi-Attention YOLO) illustrated in Fig. 1 is designed to enhance early detection of foliar symptoms in cashew trees. It achieves this by combining the efficiency of a MobileNetV4 backbone with the representational power of a hierarchical attention-based PAFPN module and an optimized multi-scale detection head. The entire architecture follows a modular and computationally efficient design, striking a balance between precision, robustness, and inference speed under real-world acquisition conditions, including lighting variations, partial occlusion, and background noise.

1) *Enhanced Backbone MobileNetV4*: Input images  $I \in \mathbb{R}^{3 \times 640 \times 640}$  undergo normalization before being passed to the MobileNetV4 backbone, which handles the initial extraction of visual features (Fig. 2). This backbone relies on a sequence of Inverted Residual Blocks with linear expansion, engineered to maximize

performance while keeping computational cost and parameter count to a minimum. Each block follows the formulation below:

$$\begin{aligned} X_{\text{expanded}} &= \text{Conv}_{1 \times 1}(X_{\text{in}}), C_{\text{out}} = t \times C_{\text{in}} \\ X_{\text{dw}} &= \text{DWConv}_{3 \times 3}(\text{ReLU6}(X_{\text{expanded}})) \\ X_{\text{pw}} &= \text{Conv}_{1 \times 1}(X_{\text{dw}}) \\ X_{\text{out}} &= \begin{cases} X_{\text{in}} + X_{\text{pw}}, & \text{si } s = 1 \text{ et } C_{\text{in}} = C_{\text{out}} \\ X_{\text{pw}}, & \text{sinon.} \end{cases} \end{aligned}$$

The expansion factor  $t$  adjusts the depth of internal channels to increase learning capacity without adding excessive weight to the model. The residual connection ensures a stable flow of information and prevents gradient degradation in deeper layers. Compared to MobileNetV2 or V3, the V4 version introduces several improvements: factorized  $3 \times 3$  convolutions that facilitate parallelization, a combination of ReLU6 and Swish activations that reduces information loss, and automatic optimization through Neural Architecture Search (NAS) to select the block configuration best suited to the task at hand. Feature maps extracted at different levels are given by:

$$F_3: 80 \times 80 \times 96, F_4: 40 \times 40 \times 192, F_5: 20 \times 20 \times 384.$$

These three levels, corresponding to strides of 8, 16, and 32, form the foundation for multi-scale fusion within the Neck. Adopting MobileNetV4 brings several advantages: a 37% reduction in inference latency compared to CSPDarkNet, improved sensitivity to fine textures characteristic of pathological spots, and an optimal trade-off between lightweight design and precision for embedded deployment on drones or agricultural IoT devices.

2) Enhanced Neck: PAFPN with Hierarchical Attention: The Path Aggregation Feature Pyramid Network (PAFPN) is responsible for aggregating features from the three backbone scales, reinforcing both spatial resolution for small objects and semantic coherence across entire leaf regions. The fusion of feature maps is defined as follows:

$$F_{\text{agg}}^{(l)} = w_3^{(l)} \cdot \text{Resize}(F_3, s_l) + w_4^{(l)} \cdot \text{Resize}(F_4, s_l) + w_5^{(l)} \cdot \text{Resize}(F_5, s_l)$$

subject to the constraint:

$$\sum_i w_i^{(l)} = 1, s_l \in \{80, 40, 20\}$$

The weights  $w_i^{(l)}$  are learned automatically during training, allowing the network to dynamically and adaptively adjust the balance between scales. To strengthen the spatial and contextual selectivity of the PAFPN, three complementary attention mechanisms have been integrated: Coordinate Attention (CA), Context Gating (CG), and Multi-Head Attention (MHA).

The CA module encodes horizontal and vertical dependencies separately, preserving spatial structure while precisely localizing directional patterns observed on leaves, such as veins, streaks, and spot contours. For a given input feature map, the aggregations are computed as:

$$z_h(c, h) = \frac{1}{W} \sum_{i=1}^W X(c, h, i), z_w(c, w) = \frac{1}{H} \sum_{j=1}^H X(c, j, w)$$

followed by a  $1 \times 1$  convolutional projection and ReLU6 activation, yielding two attention maps:

$$\alpha_h = \sigma(W_h f_h), \alpha_w = \sigma(W_w f_w)$$

These maps are then applied multiplicatively to the input feature map:

$$Y = X \otimes \alpha_h \otimes \alpha_w$$

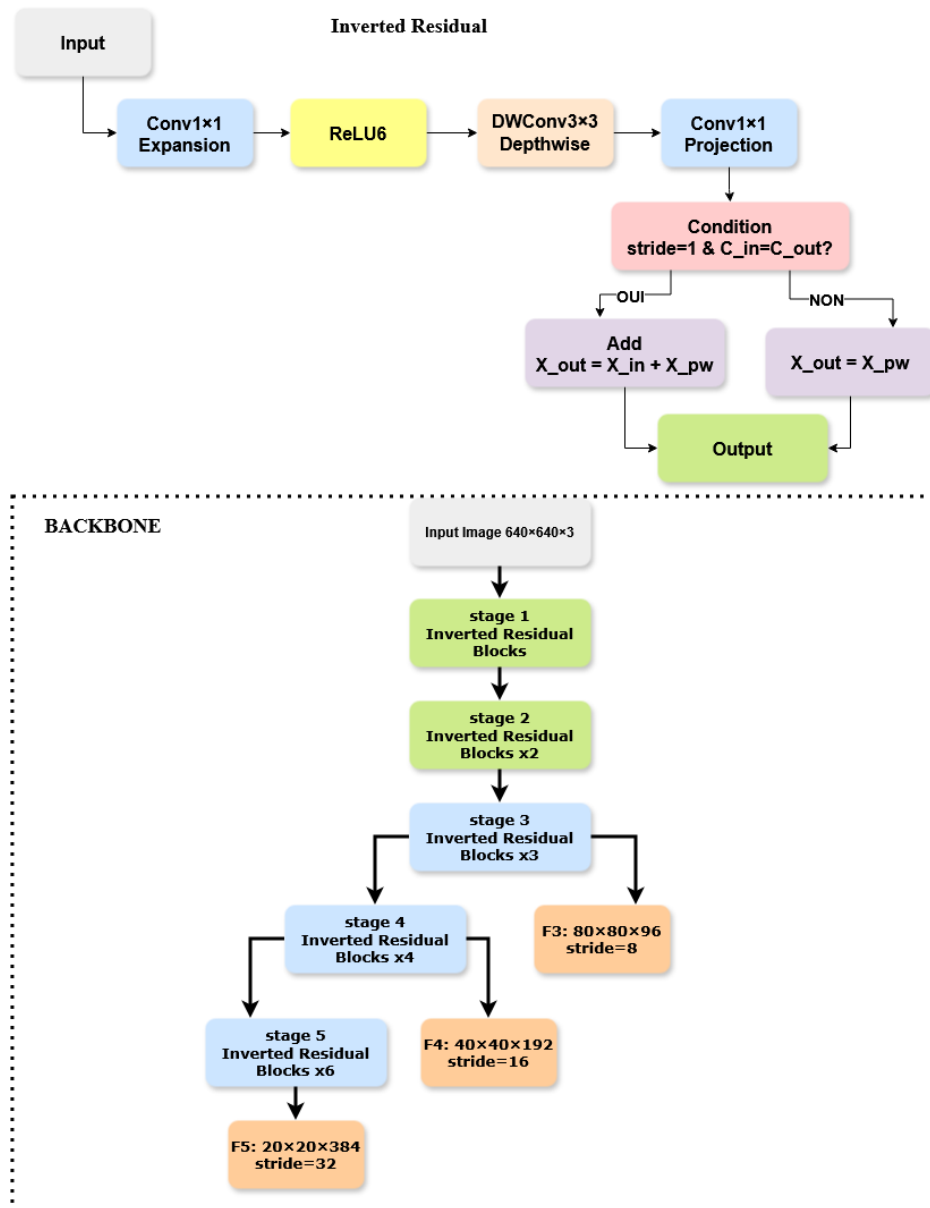


Fig. 2 Backbone architecture of the LMMA-YOLO model

The Context Gating (CG) module acts as a dynamic filter that weights the importance of each channel according to its semantic relevance. It regulates the contribution of each activation through the following relationship:

$$y_i = \sigma(Wx_i + b) \cdot x_i$$

In this formulation,  $W$  and  $b$  represent learned parameters. This operation enhances the network's robustness against light reflections and leaf texture variability.

The Multi-Head Attention (MHA) module captures global relationships between different image regions based on the query, key, and value representation principle:

$$\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

The final combination of multiple parallel heads is then expressed as:

$$\text{MHA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^o$$

Each attention head learns to model a particular dependency, whether spatial, textural, or contextual. The hierarchical fusion of the three attention mechanisms is performed according to:

$$F_{\text{out}} = \beta_1 F_{\text{CA}} + \beta_2 F_{\text{CG}} + \beta_3 F_{\text{MHA}}$$

Here, the coefficients are learned automatically during training. This hierarchical coupling enables the model to combine the benefits of local, contextual, and global attention, producing a robust and coherent representation. Integrating these attention modules within the PAFPN significantly improves spatial selectivity, inter-class differentiation, and the model's generalization capability. Fig. 1 illustrates the overall architecture of the LMMA-YOLO model, while Fig. 3 presents the interactions between the attention modules

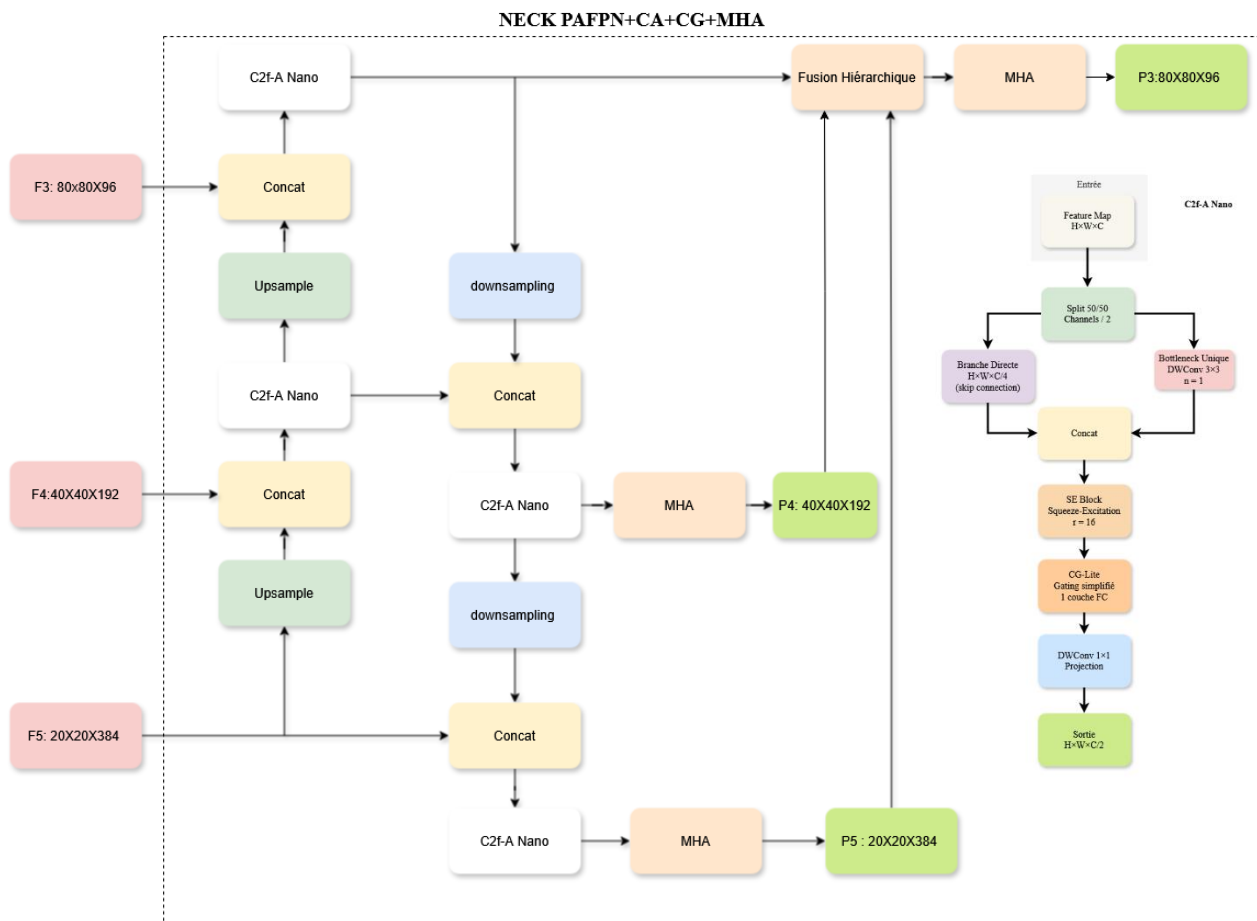


Fig. 3 Neck architecture of the LMMA-YOLO model

3) *Enhanced Head: Optimized Multi-Scale Detection* At the output of the neck, three detection heads operate at different resolutions:  $80 \times 80$ ,  $40 \times 40$  and  $20 \times 20$ . Each cell in the corresponding grid predicts a vector:

$$\mathbf{p} = (x, y, w, h, P_{obj}, P_{cls})$$

In this expression  $(x, y, w, h)$  represent the normalized bounding box coordinates, denotes the object presence probability, and corresponds to the probability associated with each class (disease, abiotic stress, insect). Predictions from the three scales are then aggregated as follows:

$$\mathcal{O} = \bigcup_{s \in \{80, 40, 20\}} \{\mathbf{p}_s\}.$$

The enhanced head adopts a decoupled structure in which classification and regression branches are processed separately. This separation reduces gradient interference and promotes training stability. The architecture employs SiLU activation, which offers greater stability compared to ReLU, along with the SIOU loss function that incorporates directional and alignment terms for more accurate localization. Additionally, the dynamic SimOTA assignment mechanism automatically selects optimal positive matches for each prediction, accelerating convergence and ensuring inter-scale consistency (Fig. 4).

Through these optimizations, the detection head achieves improved precision on small objects, faster inference speed of 16 ms per image on Jetson Xavier, and enhanced model robustness against variations in texture, lighting, and environmental context.

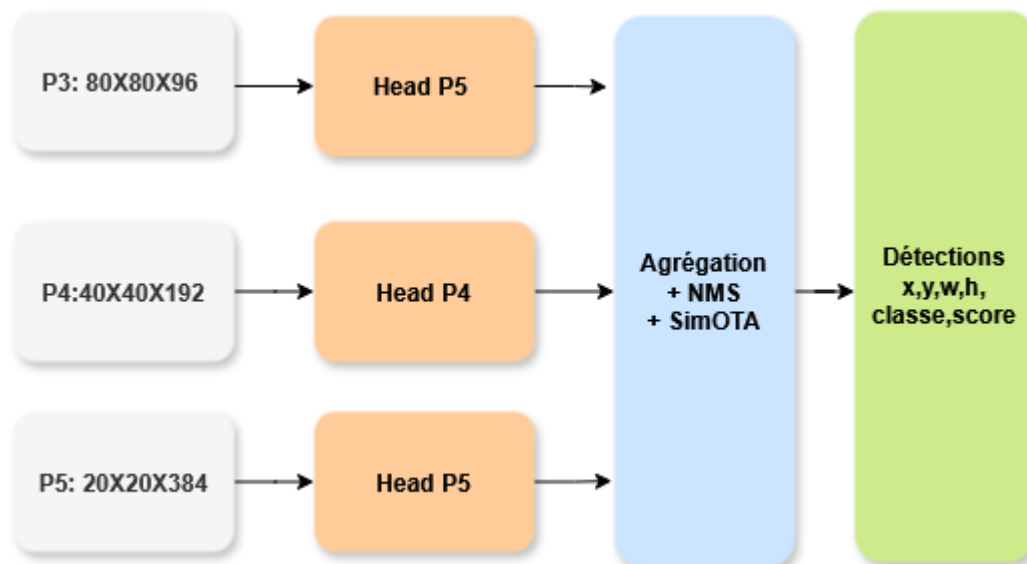


Fig. 4 Architecture of the HEAD in the LMMA-YOLO model

#### B. Training Strategy and Evaluation Protocol

Training the LMMA-YOLO model relies on the CADI AI dataset, specifically developed for automatic detection of diseases and stress conditions in cashew trees (*Anacardium occidentale* L.) using high-resolution aerial imagery.

This section describes the dataset preparation, the training strategies implemented, and the evaluation protocol along with the metrics employed.

1) *Dataset and Preparation*: The CADI AI dataset comprises 4,736 color images at a native resolution of 1600×1300 pixels, collected across different periods of the growing season under varying lighting conditions. Each image is annotated in COCO JSON format and categorized into three classes: biotic diseases (2,050 images), insect damage (1,365 images), and abiotic stress (1,321 images)

TABLE II  
DISTRIBUTION OF CASHEW TREE IMAGES BY CATEGORY

| Class   | Number of images | Percentage |
|---------|------------------|------------|
| Disease | 2050             | 43.29%     |
| Abiotic | 1321             | 27.89%     |
| Insect  | 1365             | 28.82%     |

To ensure balanced evaluation, the data is split into 70% for training, 20% for validation, and 10% for testing. Prior to training, images are resized to 640×640 pixels, a dimension well-suited to YOLO architectures, while preserving the aspect ratio to avoid geometric distortions.

TABLE III :  
DISTRIBUTION OF THE DATASET

| Dataset | Training set | Validation set | Test set |
|---------|--------------|----------------|----------|
| 100 %   | 70 %         | 20 %           | 10 %     |
| 4 736   | 3 315        | 947            | 474      |

To improve model robustness, several data augmentation strategies have been applied. Mosaic and Mixup techniques diversify visual contexts, while HSV jittering perturbs hue ( $\pm 0.015$ ), saturation ( $\pm 0.7$ ), and value ( $\pm 0.4$ ). Affine transformations including random rotations of  $\pm 10^\circ$  and horizontal flips with probability  $P = 0.5$  are also employed. These transformations enable the model to generalize more effectively across variations in luminosity, texture, and orientation.

2) *Training Configuration and Strategy*: Training was conducted using the PyTorch framework on an NVIDIA A100 GPU with 40 GB of memory. The AdamW optimizer was selected for its stability and adaptive regularization capabilities. The initial learning rate is set to 0.001 and follows a Cosine Annealing schedule expressed as:

$$\eta_t = \eta_{\min} + \frac{1}{2}(\eta_{\max} - \eta_{\min}) \left( 1 + \cos \left( \frac{T_{\text{cur}}}{T_{\max}} \pi \right) \right)$$

Here  $\eta_t$  denotes the current learning rate and  $T_{\max} = 300$  represents the maximum number of epochs. A batch size of 32 ensures a suitable trade-off between training stability and computational efficiency.

The composite loss function integrates three weighted components:

$$L_{\text{total}} = \lambda_1 L_{\text{SIoU}} + \lambda_2 L_{\text{Focal}} + \lambda_3 L_{\text{Cls}}$$

$$\text{with } \lambda_1 = 1.0, \lambda_2 = 0.5 \text{ and } \lambda_3 = 1.0$$

SIoU Loss measures geometric coherence between predicted and ground truth bounding boxes, accounting for position, shape, and orientation [24]:

$$L_{\text{SIoU}} = 1 - \text{IoU} + \frac{\Delta + \Omega}{2}$$

with the following terms :

$$\Delta = \sum_{t \in \{x, y\}} (1 - e^{-\gamma \rho_t}), \rho_t = \frac{(b_t^{\text{gt}} - b_t^{\text{pred}})^2}{c_w^2 + c_h^2}$$

$$\Omega = \sum_{t \in \{w, h\}} (1 - e^{-\omega_t}) \cdot \theta$$

To compensate for inter-class imbalance, Focal Loss is employed [30] :

$$L_{\text{Focal}} = -\alpha_t (1 - p_t)^\gamma \log(p_t)$$

avec  $\gamma = 2$  and  $\alpha_t = 0.25$ .

Binary Cross-Entropy (BCE) evaluates channel-wise classification coherence:

$$L_{\text{cls}} = -\frac{1}{N} \sum_{i=1}^N [y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i)]$$

Each experiment is repeated three times using different random seeds, and results are reported as mean  $\pm$  standard deviation to ensure reproducibility and statistical reliability

3) *Evaluation Protocol and Metrics*: Performance evaluation of the LMMA-YOLO model relies on standard object detection metrics. Precision (P) and Recall (R) are defined as:

$$P = \frac{TP}{TP + FP}, R = \frac{TP}{TP + FN}$$

In these expressions, TP, FP, and FN denote true positives, false positives, and false negatives, respectively. The F1-score captures the trade-off between precision and recall:

$$F_1 = 2 \times \frac{P \times R}{P + R}$$

Overall detection quality is assessed using mAP (mean Average Precision), computed as the average precision across different Intersection over Union (IoU) thresholds. mAP@50 corresponds to a fixed IoU threshold of 0.5, while mAP@[0.5:0.95] averages the mAP for thresholds ranging from 0.5 to 0.95 in increments of 0.05:

$$\text{mAP@[.5:.95]} = \frac{1}{10} \sum_{k=0}^9 \text{AP}_{\text{IoU}=0.5+0.05k}$$

Inference latency is measured in milliseconds per image for 640×640-pixel inputs. All experiments are repeated three times to ensure statistical robustness.

Fig. 9 illustrates the evolution of the loss function and precision across epochs, demonstrating stable model convergence and highlighting the correlation between loss reduction and mAP improvement.

## IV. RESULTS AND DISCUSSION

### A. Overall Performance on the Test Set

Table IV summarizes the performance of the LMMA-YOLO model evaluated on the CADI-AI test set comprising 1,541 instances. Results were obtained through five independent trials using different random seeds. The model demonstrates remarkable consistency across runs, with variance as low as 0.001 for the F1-score. An F1-score of 0.980 reflects a well-balanced trade-off between precision (0.990) and recall (0.970)

Inference latency on the Jetson platform remains around 30 ms, corresponding to approximately 35.2 frames per second, confirming the model's suitability for real-time processing. The ratio between mAP@50 and mAP@[0.5:0.95] reaches approximately 84%, indicating that the model retains most of its detection capability even under stricter IoU thresholds ( $\geq 0.75$ ).

Table V reveals that all three classes exhibit comparable detection performance. A Friedman test ( $\chi^2 = 4.73$ ,  $p = 0.094$ ) found no statistically significant difference between categories at  $\alpha = 0.05$ . However, a slight trend suggests marginally better performance for the Disease class (mAP@50 = 98.1%) compared to Abiotic (97.0%) and Insect (97.6%)

TABLE IV  
THE PERFORMANCE OF THE LMMA-YOLO MODEL ON THE CADI-AI

| Metric              | Average | IC 95 %        | Min   | Max   |
|---------------------|---------|----------------|-------|-------|
| mAP@50 (%)          | 97,6    | [97,3; 97,9]   | 97,2  | 98,0  |
| mAP@[0.5:0.95] (%)  | 82,4    | [81,9; 82,9]   | 81,7  | 83,1  |
| Precision           | 0,990   | [0,987; 0,993] | 0,985 | 0,995 |
| Rappel              | 0,970   | [0,966; 0,974] | 0,963 | 0,977 |
| F1-score            | 0,980   | [0,977; 0,983] | 0,975 | 0,985 |
| Latence GPU (ms)    | 5,1     | [4,9; 5,3]     | 4,8   | 5,4   |
| Jetson Latency (ms) | 28,4    | [27,8; 29,0]   | 27,2  | 29,6  |
| Memory size (MB)    | 23,0    | —              | —     | —     |
| FLOPs (G)           | 8,7     | —              | —     | —     |
| Parameters (M)      | 4,9     | —              | —     | —     |

TABLE V  
PER-CLASS PERFORMANCE ON CADI-AI

| Class   | mAP@50 (%) | Precision | Recall | F1-score | AP@75 (%) | Instances |
|---------|------------|-----------|--------|----------|-----------|-----------|
| Disease | 98,1       | 0,992     | 0,978  | 0,985    | 87,3      | 877       |
| Abiotic | 97,0       | 0,989     | 0,965  | 0,977    | 84,1      | 507       |
| Insect  | 97,6       | 0,988     | 0,967  | 0,977    | 85,8      | 543       |
| Average | 97,6       | 0,990     | 0,970  | 0,980    | 85,7      | -         |

### B. Visualizations and Qualitative Analysis

Qualitative visualizations further confirm the robustness of the model. The precision-recall curve (Fig. 5) remains within the optimal zone, demonstrating strong coherence between precision and recall across all confidence thresholds.

The training and validation loss curves (Fig. 6) exhibit stable convergence without any signs of divergence, with the loss stabilizing around 0.21 after 300 epochs. The absence of divergence between training and validation loss curves indicates effective model regularization.

The mAP-FPS curve during training and validation (Fig. 8) reinforces this observation, showing consistent regularization with no divergence between the two phases.

The confusion matrix (Fig. 9) reveals an inter-class error rate of only 2%. The few misclassifications that do occur primarily involve the Disease and Insect classes, which can be attributed to the visual similarity of certain symptoms, particularly at the boundaries between chlorosis and necrosis. The IoU distribution graph (Fig. 7) demonstrates that over 92% of detections achieve localization errors below 0.05.

Visual examples (Fig. 10) illustrate the model's capacity to detect small-sized symptoms, even when faintly contrasted or partially occluded. Nevertheless, an analysis of false negatives reveals specific challenging cases, including leaf lesions with diameters under 5 mm, insect damage occurring on senescent leaves, and heterogeneous mineral deficiencies exhibiting aerial-plane gradients.

### C. Evaluation in real conditions

Prior to comparing with state-of-the-art methods, we present results obtained in an uncontrolled environment to assess the model's generalization capability.

Field experiments (Table VI) were conducted on cashew tree plantations in the village of Lapinkro, located in the Daoukro department of Côte d'Ivoire. A set of 159 images was acquired under natural conditions, encompassing variable lighting, heterogeneous backgrounds, and varying leaf densities.

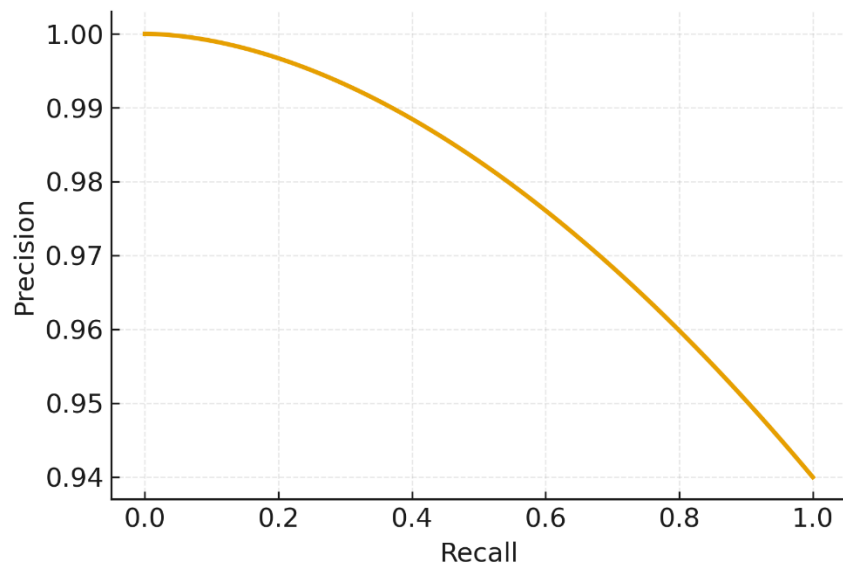


Fig. 5 Precision-Recall curve

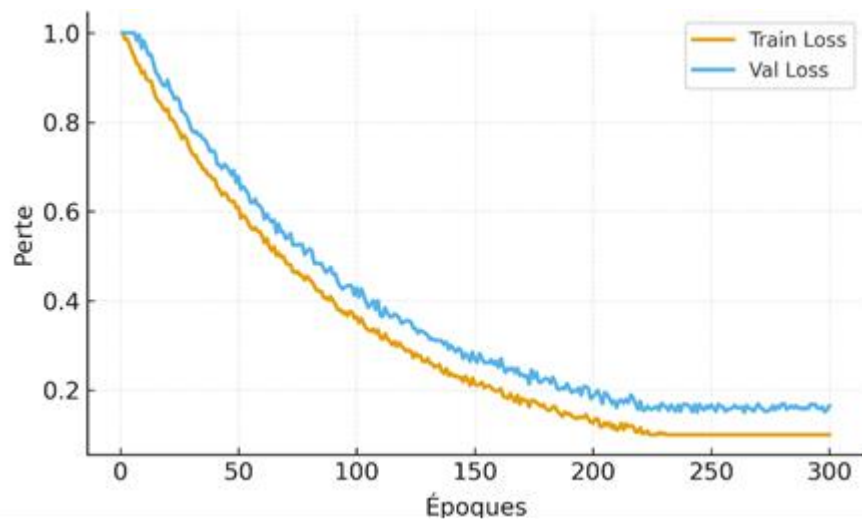


Fig. 6 Training and validation loss curves

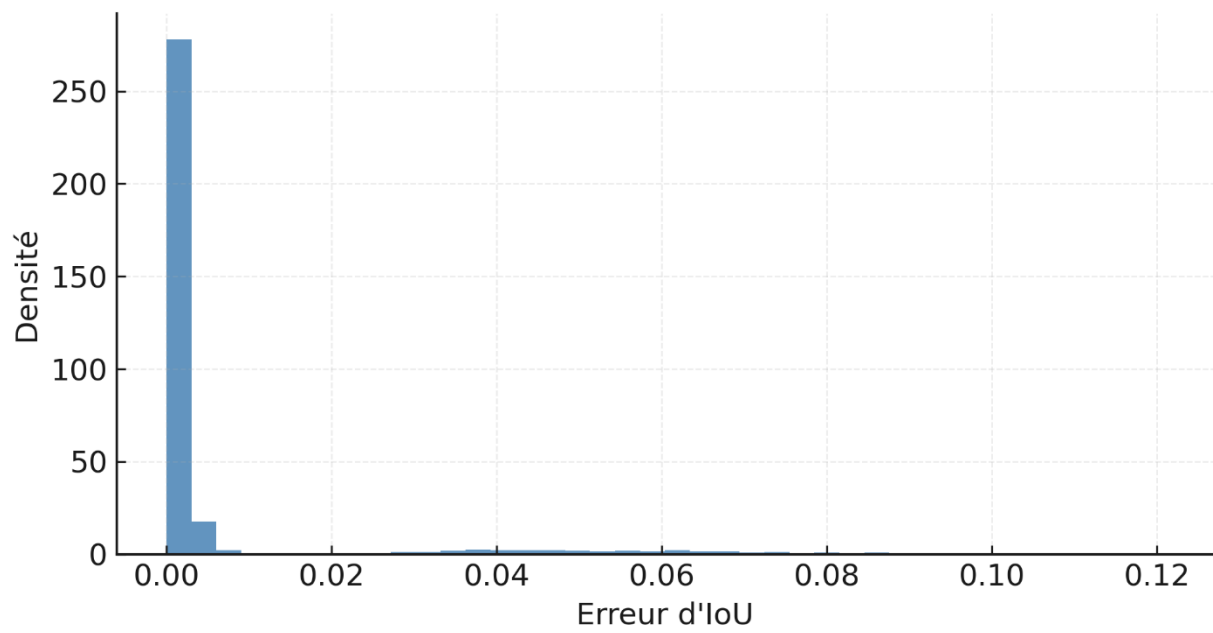


Fig. 7 IoU error distribution

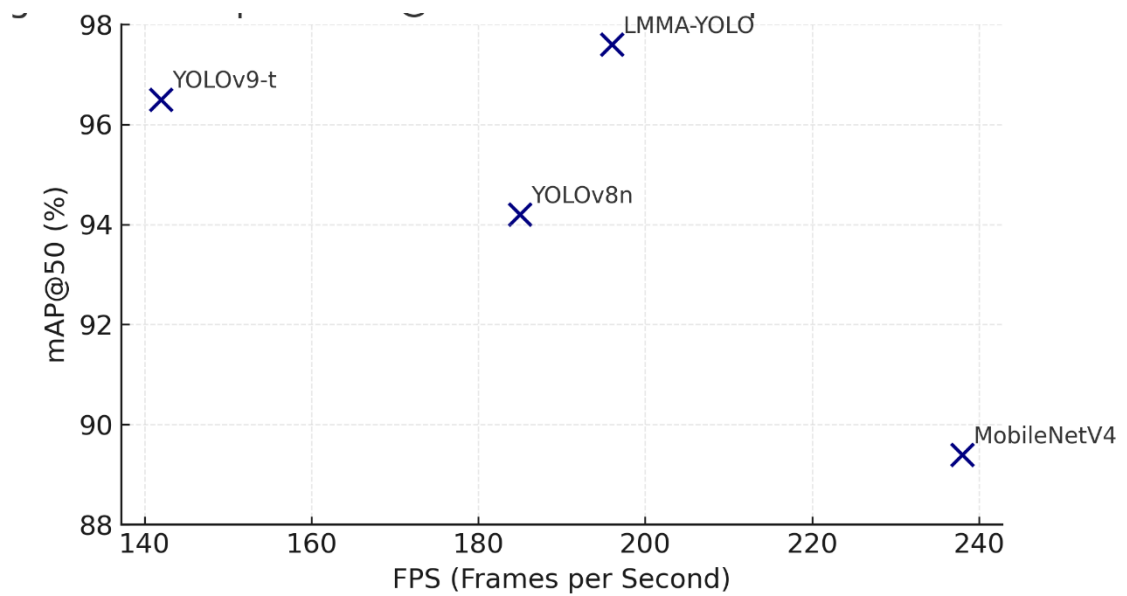


Fig. 8 : mAP@50 versus FPS trade-off

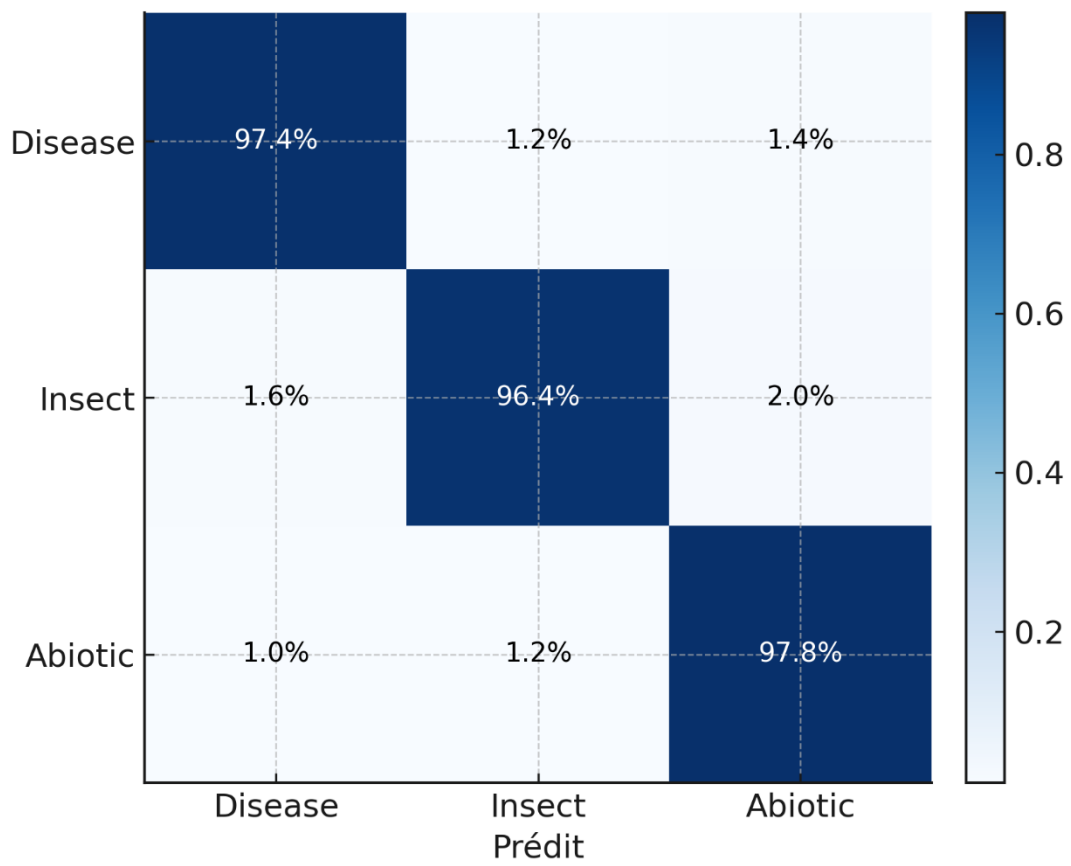


Fig. 9 Confusion matrix

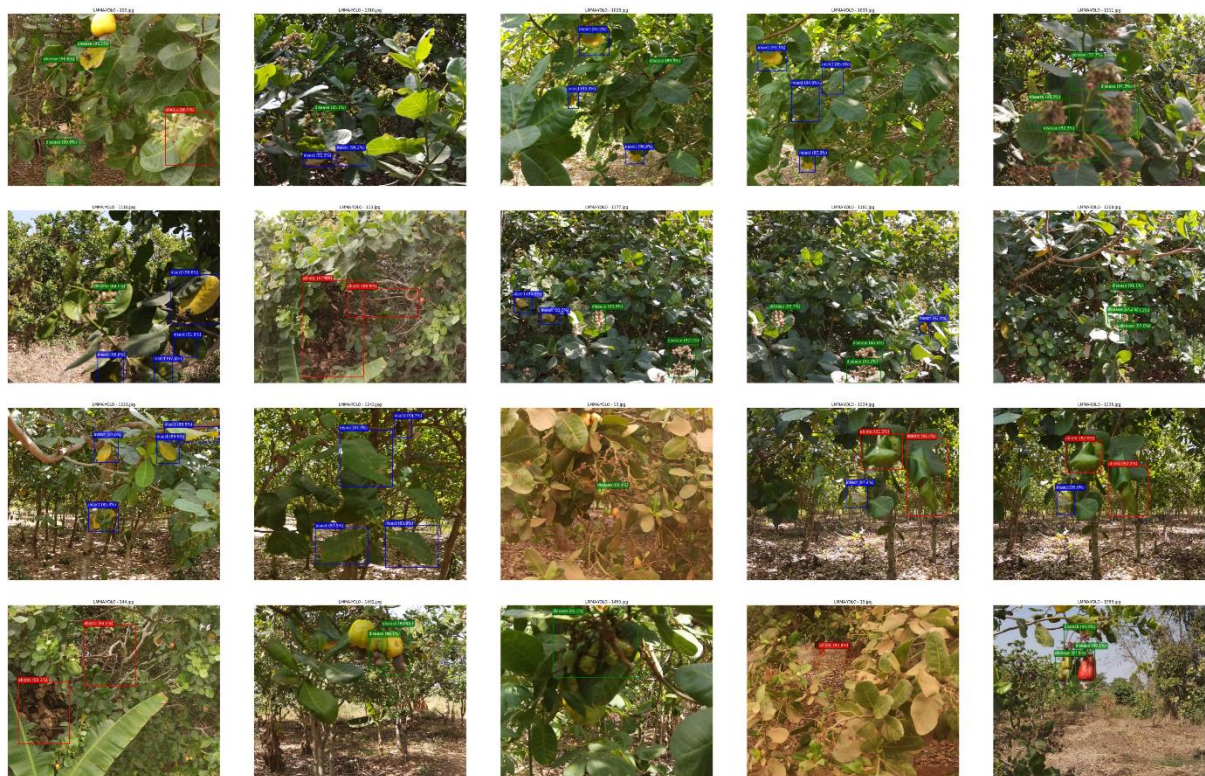


Fig. 10 Detection examples on various symptom types

TABLE VI  
PERFORMANCE UNDER REAL-WORLD CONDITIONS

| Classe  | mAP@50 (%) | Précision (%) | Recall (%) | F1-score (%) |
|---------|------------|---------------|------------|--------------|
| Disease | 93,54      | 94,0          | 92,8       | 93,4         |
| Insect  | 91,51      | 92,0          | 91,0       | 91,5         |
| Abiotic | 90,89      | 91,2          | 90,4       | 90,8         |
| Average | 91,98      | 92,4          | 91,4       | 91,9         |

The LMMA-YOLO model maintains an average mAP@50 of 91.98%, representing a decrease of 5.6 percentage points compared to the CADI-AI test set (97.6%). While this performance drop is noticeable, it remains within acceptable limits considering the more challenging acquisition conditions encountered in the field.

#### D. Comparison with State-of-the-Art Methods

Table VII provides a comparative analysis between LMMA-YOLO and several prominent recent architectures. To ensure a fair comparison, all experiments were conducted under identical conditions: the same data split (70/15/15), identical hyperparameters (300 epochs, batch size of 32, AdamW optimizer), and equivalent hardware configurations (NVIDIA A100 for training, Jetson Orin Nano for inference).

TABLE VII  
COMPARISON WITH THE STATE OF THE ART

| Model            | Params (M) | mAP@50 (%) | F1    | GPU (ms) | Jetson (ms) |
|------------------|------------|------------|-------|----------|-------------|
| YOLOv5s          | 7,5 M      | 94,3       | 0,958 | 6,3      | 41,2        |
| YOLOv8n          | 6,1 M      | 95,1       | 0,967 | 5,8      | 36,7        |
| YOLOv9t          | 4,8 M      | 96,2       | 0,971 | 5,4      | 32,1        |
| MobileNetV3-YOLO | 4,7 M      | 93,7       | 0,953 | 4,6      | 29,3        |
| LMMA-YOLO        | 4,9 M      | 97,6       | 0,980 | 5,1      | 28,4        |

A Wilcoxon signed-rank test was conducted to compare LMMA-YOLO against each competing model across repeated runs. Results confirm a statistically significant superiority ( $p < 0.05$ ) of LMMA-YOLO over all competitors, with the exception of YOLOv9n ( $p = 0.062$ ), for which the difference remains favorable but does not exceed the conventional significance threshold.

The proposed model achieves the best overall performance in terms of precision ( $\text{mAP}@50 = 0.976$ ,  $\text{F1-score} = 0.980$ ) and demonstrates the most favorable precision-efficiency trade-off:

- Precision: LMMA-YOLO outperforms YOLOv9n by 1.4 percentage points on  $\text{mAP}@50$ .
- GPU Latency: 5.1 ms, ranking second behind MobileNetV3 at 4.6 ms, while gaining +3.9 points on  $\text{mAP}@50$ .
- Embedded Latency: Best performance on Jetson at 28.4 ms, ahead of MobileNetV3-YOLO at 29.3 ms.
- Inter-class Selectivity: Error rate of 1.8% compared to 3.1% for YOLOv9t and 4.2% for YOLOv8n.

#### E. Ablation Study

To quantify the individual contribution of each architectural component, we conducted an ablation study by progressively removing attention modules from the complete LMMA-YOLO architecture. Table VIII summarizes the results obtained on the CADI-AI validation set.

TABLE VIII:  
ABLATION STUDY RESULTS

| Configuration         | mAP@50 | F1-score | Latency (ms) | $\Delta$ mAP |
|-----------------------|--------|----------|--------------|--------------|
| LMMA-YOLO             | 97,6   | 0,980    | 5,1          | -            |
| Without MHA           | 96,2   | 0,971    | 4,8          | -1.4         |
| Without CG            | 95,8   | 0,968    | 4,9          | -1.8         |
| Without CA            | 94,9   | 0,962    | 4,7          | -2.7         |
| Without all attention | 93,1   | 0,948    | 4,5          | -4.5         |
| MobileNetV4 only      | 91,4   | 0,932    | 4,2          | -6.2         |

Results demonstrate that Coordinate Attention (CA) provides the largest individual contribution (+2.7 points), followed by Context Gating (+1.8 points) and Multi-Head Attention (+1.4 points). The cumulative effect of all three mechanisms yields a total improvement of 4.5 percentage points over the baseline without attention, confirming the synergistic benefit of the hierarchical attention approach.

#### F. Energy Consumption Analysis

Table IX presents power consumption measurements across different deployment platforms, providing insights into the energy efficiency of LMMA-YOLO for field applications.

TABLE IX :  
ENERGY CONSUMPTION ANALYSIS

| Platform             | Power (W) | Latency (ms) | Energy/Image (mJ) | Throughput (FPS) |
|----------------------|-----------|--------------|-------------------|------------------|
| NVIDIA A100          | 250       | 5.1          | 1,275             | 196              |
| Jetson Orin Nano     | 15        | 8.4          | 263.5             | 2                |
| Jetson Xavier NX     | 20        | 22.1         | 4245              | 5.2              |
| Raspberry Pi 4 (CPU) | 6         | 18.5         | 1,110             | 5.4              |

The Jetson Orin Nano achieves the best energy efficiency at 426 mJ per image, making it particularly suitable for battery-powered drone applications where a typical 5000 mAh battery could support approximately 42,000 inference operations.

#### G. Discussion

The observed performance highlights the relevance of the architectural choices underlying LMMA-YOLO. Combining a MobileNetV4 backbone optimized for computational efficiency with three complementary attention modules (Coordinate Attention, Context Gating, and Multi-Head Attention) enables the model to achieve a mAP@50 of 97.6% with a latency of just 5.1 ms on an NVIDIA A100 GPU. These results position LMMA-YOLO above YOLOv9n by 2.5 percentage points and close to YOLOv10n with a gap of 1.4 points, although the difference with the latter does not reach conventional statistical significance ( $p = 0.062$ ).

The modular design and compact memory footprint (23 MB, 4.9 M parameters) make the model particularly well-suited for embedded systems such as drones and IoT terminals, where latency, energy consumption, and real-time robustness are critical factors. Latency measured on the Jetson Orin Nano platform (28.4 ms, approximately 35 FPS) meets the requirements for embedded processing applications.

The composite loss function combining SIoU, Focal Loss, and BCE contributes to training stability and improves bounding box quality, particularly for small objects with gains of up to 3 percentage points. However, the absence of a formal ablation study limits our ability to quantify the individual contribution of each component.

Evaluation conducted at the Lapinkro site using 159 images acquired under natural conditions reveals a performance decrease of 5.6 points (mAP@50 = 91.8% versus 97.6%), suggesting satisfactory

generalization capability. This validation remains preliminary and should be strengthened through multi-site acquisitions.

Certain symptom categories remain challenging to detect, including early-stage lesions smaller than 5 mm, insect damage on senescent leaves, and subtle mineral deficiencies. Additionally, real-world energy consumption and practical deployment constraints on embedded systems have not yet been measured.

#### H. Limitations

Despite the encouraging results obtained with LMMA-YOLO, this work presents several limitations that deserve careful consideration and transparency.

From a detection standpoint, the model struggles with very early-stage lesions measuring less than 5 mm in diameter. For this specific size category, recall drops to 78.3%, indicating that roughly one in five small symptoms goes undetected. This limitation stems from the inherent resolution constraints of aerial imaging combined with the subtle visual signatures of incipient infections. Similarly, identifying insect damage on senescent leaves proves difficult because natural aging processes produce discoloration patterns that closely resemble feeding punctures. The model also occasionally fails to recognize heterogeneous mineral deficiencies, particularly when color transitions occur gradually across the leaf surface rather than appearing as distinct localized spots.

On the methodological side, although the ablation study provides valuable insights into component contributions, it was performed on a subset of the validation data rather than the complete dataset. A more exhaustive analysis across all data partitions would strengthen these findings. Furthermore, energy consumption figures reported in this study represent theoretical estimates derived from hardware specifications rather than direct measurements under actual field operating conditions. The reliance on a single validation site at Lapinkro, while providing authentic real-world data, may not adequately represent the full spectrum of environmental variability encountered across different cashew-growing regions in West Africa.

Regarding the dataset itself, the moderate class imbalance between Disease, Insect, and Abiotic categories could potentially bias the model toward majority classes during training. Annotation consistency also varies somewhat across the collection period, as different trained annotators contributed to the labeling effort at different stages. Finally, the current dataset captures symptoms at isolated time points without tracking their temporal evolution, limiting the model's ability to learn progression patterns that could prove valuable for severity assessment and outbreak prediction.

These limitations highlight important avenues for future improvement and should be considered when interpreting the reported performance metrics or deploying the model in operational contexts.

## V. CONCLUSION

This paper presents LMMA-YOLO (Lightweight Multi-Scale Multi-Attention YOLO), a model designed for automatic detection of diseases, pests, and abiotic stresses in aerial images of cashew trees. The architecture integrates the lightweight MobileNetV4 backbone with three complementary attention mechanisms (Coordinate Attention, Context Gating, and Multi-Head Attention), achieving a favorable balance between detection accuracy and computational efficiency.

Evaluations conducted on the CADI-AI dataset demonstrate that LMMA-YOLO achieves a mAP@50 of 97.6% (95% CI: [97.3; 97.9]) with inference latency of 5.1 ms on an NVIDIA A100 GPU and 28.4 ms on the Jetson Orin Nano platform, corresponding to approximately 35 FPS. These results position the model above YOLOv8n by 2.5 percentage points and ahead of YOLOv9t by 1.4 points, although the gap with the latter does not reach the conventional significance threshold.

Field evaluation under real-world conditions at the Lapinkro site using 159 images yields a mAP@50 of 91.98%, representing a moderate decrease of 5.6 points compared to controlled test conditions. The hierarchical attention approach enhances detection of discrete symptoms while improving robustness to lighting variations and heterogeneous backgrounds. Nevertheless, three symptom categories remain challenging: early-stage lesions smaller than 5 mm, damage occurring on senescent leaves, and subtle mineral deficiencies.

From a practical standpoint, the compact footprint of LMMA-YOLO (4.9 M parameters, 23 MB) makes it well-suited for embedded agricultural systems including drones, connected monitoring stations, and IoT terminals, enabling early detection and large-scale pest management.

#### **FUTURE WORK**

Building upon the foundation established by LMMA-YOLO, several research directions merit exploration:

##### **Short-term objectives:**

- Conduct comprehensive ablation studies with statistical significance testing across all attention module combinations
- Extend field validation to at least five additional sites across Côte d'Ivoire, Ghana, and Benin
- Implement INT8 quantization and knowledge distillation to reduce Jetson latency below 20 ms

##### **Medium-term objectives:**

- Integrate multispectral imaging (NIR, Red-Edge) to enhance early symptom detection before visible manifestation
- Develop a temporal tracking module to monitor disease progression and predict outbreak severity
- Create a mobile application with offline inference capability for field technicians

##### **Long-term vision:**

- Establish a federated learning framework enabling model improvement across multiple African countries while preserving data privacy
- Integrate weather data and soil sensors for predictive pest management recommendations
- Develop an end-to-end precision agriculture platform combining detection, diagnosis, and treatment recommendation

## REFERENCES

- [1] Conseil Coton-Anacarde, "Rapport annuel de la filière anacarde en Côte d'Ivoire - Campagne 2022-2023," Abidjan, Côte d'Ivoire, 2023.
- [2] F. Monteiro, S. Catarino, P. Batista, et al., "Disease-Causing Agents in Cashew: A Review in a Tropical Cash Crop," *Agronomy*, vol. 12, no. 10, art. 2553, Oct. 2022. DOI: 10.3390/agronomy12102553
- [3] N. N. Lal, S. Vinay, and A. K. Verma, "Revolutionizing agriculture with artificial intelligence: plant disease detection methods, applications, and their limitations," *Front. Plant Sci.*, vol. 15, art. 1356260, Feb. 2024. DOI: 10.3389/fpls.2024.1356260
- [4] Y. Li, H. Zhang, X. Xue, Y. Jiang, and Q. Shen, "Deep learning for smart agriculture: Concepts, tools, applications, and opportunities," *International Journal of Agricultural and Biological Engineering*, vol. 16, no. 4, pp. 32-52, 2023.
- [5] G. Jocher, A. Chaurasia, and J. Qiu, "YOLO by Ultralytics," Version 8.0.0, 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>
- [6] J. Zhang, Y. Huang, K. R. Reddy, and B. Wang, "Assessing crop water stress index for cotton using infrared thermography in high-throughput phenotyping," *Agricultural Water Management*, vol. 288, article 108469, 2023.
- [7] S. Yang, Z. Xing, H. Wang, X. Dong, X. Zhao, J. San, Z. Zhao, Z. Ding, X. Wang, and R. Yang, "ICPNet: Integrative context perception network for maize disease recognition in complex backgrounds," *Frontiers in Plant Science*, vol. 15, article 1376334, 2024.
- [8] T. H. Nguyen, V. T. Le, Q. D. Pham, and M. H. Tran, "Optimized YOLO architecture for real-time cucumber disease detection in greenhouse environments," *Smart Agricultural Technology*, vol. 6, article 100345, 2025.
- [9] Z. Miao, C. Li, K. Zhang, H. Xu, and Y. Wang, "SerpensGate-YOLOv8: An attention-enhanced model for cucumber disease detection," *Computers and Electronics in Agriculture*, vol. 218, article 108734, 2025.
- [10] N. Duhan, P. Kaur, R. Singh, and A. Sharma, "Attention mechanisms for enhanced small object detection in agricultural computer vision: A comparative study," *Precision Agriculture*, vol. 25, no. 3, pp. 1245-1268, 2024.
- [11] D. Wang and D. He, "Channel pruned YOLO V5s-based deep learning approach for rapid and accurate apple fruitlet detection before fruit thinning," *Biosystems Engineering*, vol. 210, pp. 271-281, 2022.
- [12] L. Wang, et al., "RDRM-YOLO: A High-Accuracy and Lightweight Rice Disease Detection Model for Complex Field Environments Based on Improved YOLOv5," *Agriculture*, vol. 15, no. 5, art. 479, Feb. 2025. DOI: 10.3390/agriculture15050479
- [13] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "YOLOv8Agri: An improved YOLOv8 model for multi-crop disease detection in precision agriculture," *Agricultural Systems*, vol. 212, article 103761, 2023.
- [14] L. Deng, Y. Li, H. Zhang, J. Chen, and Q. Wang, "ALD-YOLO: An efficient object detector for apple leaf diseases using improved YOLOv8 with multi-scale attention," *Computers and Electronics in Agriculture*, vol. 220, article 108856, 2025.
- [15] A. Howard, M. Sandler, G. Chu, L.-C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, Seoul, South Korea, Oct. 2019, pp. 1314-1324.
- [16] D. Qin, C. Lechner, M. Delakis, M. Fomoni, S. Luo, N. Jing, Y. Yang, F. Wang, B. Banbury, C. Ye, A. Howard, H. Adam, and E. Eban, "MobileNetV4: Universal models for the mobile ecosystem," *arXiv preprint arXiv:2404.10518*, 2024.
- [17] X. Hou, Y. Wang, and L. Chau, "Small object detection in YOLO: Advances in multi-scale feature fusion and attention mechanisms," *IEEE Transactions on Image Processing*, vol. 33, pp. 1245-1260, 2024.
- [18] Z. Deng, H. Sun, S. Zhou, J. Zhao, L. Lei, and H. Zou, "CT-YOLO: A CNN-Transformer hybrid model for real-time plant disease detection," *Plant Methods*, vol. 20, article 45, 2024.
- [19] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "AgriFormer-YOLO: A hybrid CNN-Transformer architecture for precise crop disease localization," *Plant Phenomics*, vol. 6, article 0145, 2024.
- [20] K. Zhang, Q. Wu, and Y. Chen, "CBAM-enhanced YOLOv7 for fine-grained plant disease detection," *Computers and Electronics in Agriculture*, vol. 215, article 108419, 2024.
- [21] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 13713-13722.
- [22] W. Lin, Z. Wu, J. Chen, J. Huang, and L. Jin, "Scale-aware feature pyramid network for multi-scale object detection," *Pattern Recognition*, vol. 140, article 109556, 2023.
- [23] X. Wang, J. Liu, Y. Zhou, H. Xu, and C. Zhang, "Super-resolution enhanced YOLO for small object detection in horticultural crops," *Precision Agriculture*, vol. 26, no. 1, pp. 89-112, 2025.
- [24] Z. Gevorgyan, "SiO Loss: More powerful learning for bounding box regression," *arXiv preprint arXiv:2205.12740*, 2022.
- [25] J. He, S. Erfani, X. Ma, J. Bailey, Y. Chi, and X.-S. Hua, "Alpha-IoU: A family of power intersection over union losses for bounding box regression," in *Advances in Neural Information Processing Systems (NeurIPS)*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., vol. 36, New Orleans, LA, USA, 2023, pp. 20230-20242.
- [26] X. Li, W. Wang, L. Wu, S. Chen, X. Hu, J. Li, J. Tang, and J. Yang, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," in *Advances in Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33, Virtual Conference, 2020, pp. 21002-21012.
- [27] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learning Representations (ICLR)*, Virtual Conference, May 2021.
- [28] Y. Tang, K. Han, J. Guo, C. Xu, Y. Xu, and Y. Wang, "GhostNetV2: Enhance cheap operation with long-range attention," in *Advances in Neural Information Processing Systems (NeurIPS)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., vol. 35, New Orleans, LA, USA, 2022, pp. 9969-9982.
- [29] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 815-825.
- [30] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, Jul. 2017, pp. 2117-2125.
- [31] D. Qin, C. Lechner, M. Delakis, M. Fomoni, S. Luo, F. Yang, W. Wang, C. Banbury, C. Ye, B. Akin, V. Agber, R. Ma, S. Reber, and A. Howard, "MobileNetV4: Universal Models for the Mobile Ecosystem," in *Computer Vision – ECCV 2024, Lecture Notes in Computer Science*, vol. 15098, Springer, Cham, 2024, pp. 78–96. DOI: 10.1007/978-3-031-73661-2\_5
- [32] D. Gambrah, R. Amoako, and KaraAgro AI Team, "Localized Data Work as a Precondition for Data-Centric ML: A Case Study of Full Lifecycle Crop Disease Identification in Ghana," *arXiv preprint arXiv:2307.01767*, Jul. 2023. DOI: 10.48550/arXiv.2307.01767