YOLO-ASPE: Enhancing Small Pest Detection in Agricultural Imagery Through Axis-Separated Positional Encoding

Abstract

Detecting small pests in agricultural imagery is challenging because insects often blend into foliage and occupy only a few pixels in high-resolution photographs. We propose YOLO-ASPE, a lightweight yet effective modification of YOLOv8 that introduces novel C2f-ASPE (Axis-Separated Positional Encoding) blocks that embed directional spatial awareness into the backbone.Conventional attention modules compress spatial dimensions into scalar representations, eliminating the precise location data essential for detecting minute targets. Our architecture addresses this limitation through independent axis-wise feature encoding that maintains positional awareness across both image dimensions. Trained and evaluated on the AgroPest-12 dataset (Majumdar, 2025), a publicly available collection of 13,143 annotated images spanning 12 agricultural pest classes YOLO-ASPE achieves 84.3% mAP@0.5, outperforming the YOLOv8s baseline by 4.7 percentage points. Improvements are particularly pronounced for small objects (+8.5 points) and morphologically similar species prone to inter-class confusion (31.6% reduction in misclassification). Despite the added attention modules, the model remains efficient, running at 48 FPS on an NVIDIA Jetson Orin with only 12.8M parameters—suitable for real-time field deployment. Our modification is deliberately minimal and restricted to the backbone; the neck and detection head remain unchanged

Keywords: Pest detection, YOLOv8, Axis-separated attention, Small object detection, Precision agriculture

Introduction

Cashew (Anacardium occidentale) represents a critical cash crop for West African economies, with Côte d'Ivoire producing over 1 million tonnesannually, accounting for 40% of global production (FAO, 2024). Pest infestations cause estimated yield losses of 20-40%, translating to approximately $200-400 million in annual economic losses for Ivorian farmers alone. Early detection of pests such as the tea mosquito bug (Helopeltisanacardii) and red-banded thrips (Selenothripsrubrocinctus) can reduce these losses by up to 60% through timely intervention (Sierra-Baqueroet al., 2024).

The YOLO family of detectors has become a popular choice for real-time detection in various domains (Redmon et al., 2016).Recent iterations have introduced significant architectural innovations, including anchor-free detection in YOLOX (Ge et al., 2021), industrial optimization in YOLOv6 (Li et al., 2022), and programmable gradient information in YOLOv9 (Wang et al., 2024). YOLOv8, in particular, represents a recent state-of-the-art variant with an efficient backbone composed of C2f blocks (Jocher et al., 2023). However, several empirical studies and surveys report that models trained and validated on standard benchmarks suffer a performance drop of 15–25 percentage points when deployed in realistic agricultural environments (Shoaib et al., 2023). This challenge has motivated

growing research interest in deep learning-based pest detection systems (Li et al., 2021; Ahmad et al., 2023), though detecting small objects in cluttered natural backgrounds remains particularly difficult (Tong et al., 2020). Three main factors contribute to this gap:

- Insects appear extremely small in drone imagery;
- Their coloration frequently matches that of surrounding leaves and branches, leading to camouflage;
- Environmental variability in lighting, occlusion, and background clutter further complicates recognition.

Attention mechanisms offer a principled way to enhance feature representations by focusing computational resources on the most informative regions or channels. Squeeze-and-Excitation (SE) networks (Hu et al., 2018) have demonstrated the benefits of channel attention, while CBAM (Woo et al., 2018) combines channel and spatial attention.More recent approaches include ECA-Net's efficient channel attention (Wang et al., 2020), SimAM's parameter-free design (Yang et al., 2021), and GAM's global attention mechanism (Liu et al., 2021).Both SE and CBAM, however, aggregate spatial information through global operations that reduce height and width dimensions to scalar values, inherently discarding the positional context necessary for precise localization of small targets. As demonstrated in Section 2.4, for a 10×10-pixel pest in a 640×640 image, GAP yields a signal-to-background ratio of merely 0.024%, effectively rendering small objects invisible to the attention mechanism. This limitation is particularly severe in agricultural imagery where pests typically occupy less than 0.5% of the image area (Rustia et al., 2021).

The Coordinate Attention mechanism (Hou et al., 2021) offers an elegant solution to this challenge by restructuring how spatial context is captured. Rather than compressing all positional data into scalar values, this approach generates separate feature descriptors for rows and columns, enabling the network to encode where informative patterns occur—a capability that proves invaluable when targets occupy mere fractions of the input dimensions.Originally proposed for mobile networks, CA has shown that it can improve representational efficiency without incurring high computational overhead. Recent developments have further refined Coordinate Attention for real-time applications. (Talha et al., 2025) demonstrated that efficient variants of CA can maintain detection accuracy while reducing computational overhead in mobile deployment scenarios. Building upon these advances, we investigate whether CA's spatial preservation capabilities can address the specific challenges of agricultural pest detection, where targets are not only small but also exhibit strong camouflage against natural backgrounds. Motivated by its ability to retain spatial cues, we investigate whether CA can also improve detection of small, camouflaged pests in agricultural imagery.

In this work, we propose YOLO-ASPE, a variant of YOLOv8 that integrates Coordinate Attention into the backbone via C2f-ASPE (Axis-Separated Positional Encoding) blocks. Our architecture modification is targeted and minimal: we replace selected C2f blocks in the backbone with C2f-ASPE blocks, while leaving the neck and detection head unchanged. This design adds only 1.6M parameters

but leads to substantial performance gains on the metrics that are most relevant for field deployment. We validate YOLO-ASPE on the AgroPest-12 dataset (Majumdar, 2025), which captures real-world pest imagery under varying illumination, occlusion, and camouflage conditions in agricultural environments.

The main contributions of this paper are:

- The integration of Coordinate Attention into the YOLOv8 backbone via a novel C2f-ASPE block combining spatially-aware and channel-wise attention
- A comprehensive evaluation on AgroPest-12 demonstrating a 4.7-point mAP@0.5 improvement, including an 8.5-point gain on small objects and significant improvements on camouflaged pests;
- An efficiency analysis showing that YOLO-ASPE maintains real-time inference at 48 FPS on edge hardware (Jetson Orin), making it suitable for practical deployment in agricultural environments.

Materials and Methods

1. Dataset

TABLE I : SUMMARIZES THE CLASS-WISE IMAGE DISTRIBUTION AND AVERAGE OBJECT SIZES

Class	Number	Share (%)	Challenge
Ants	1,247	9.5	Tiny, clustered
Bees	1,156	8.8	Looks like wasps
Beetles	1,089	8.3	Looks like weevils
Caterpillars	1,312	10.0	Green on green
Earthworms	987	7.5	Looks like slugs
Earwigs	923	7.0	Small, hides in crevices
Grasshoppers	1,245	9.5	Camouflaged when still
Moths	1,078	8.2	Variable poses
Slugs	1,012	7.7	Elongated, low contrast
Snails	1,156	8.8	Shell helps, actually
Wasps	1,119	8.5	Looks like bees
Weevils	819	6.2	Smallest class
Total	**13,143**	**100**	

We evaluate YOLO-ASPE on the Agricultural Pests Image Dataset (Rupankar Majumdar, 2025), a publicly available collection hosted on Kaggle comprising 13,143 annotated images spanning 12 classes of common agricultural pests: Ants, Bees, Beetles, Caterpillars, Earthworms, Earwigs, Grasshoppers, Moths, Slugs, Snails, Wasps, and Weevils. The images were sourced from Flickr and

resized to a maximum dimension of 300 pixels. These classes represent diverse morphological characteristics and detection challenges, from tiny clustered insects (Ants) to elongated specimens (Earthworms, Caterpillars). While this dataset was not specifically collected in cashew orchards, the pest species represented are commonly encountered in West African agricultural environments, including Ivorian cashew plantations (Norshie et al., 2021).The diversity of morphological characteristics aligns with challenges identified in recent pest detection surveys (Li et al., 2021; Liu & Wang, 2021). To simulate realistic agricultural monitoring conditions, we applied preprocessing to resize images to 640×640 pixels for training. Environmental variability was introduced through data augmentation including early-morning lighting simulation, shadow effects, and background clutter augmentation. The dataset was split into training (70%, 9,200 images), validation (15%, 1,971 images), and test (15%, 1,972 images) subsets using stratified sampling to preserve class distributions. Data augmentation followed standard YOLO practice (Jocher et al., 2023) and included random horizontal and vertical flips, rotations in the range ±15°, scale variations in the range 0.8–1.2×, and mosaic augmentation.

2. YOLO-ASPE Architecture

YOLO-ASPE builds upon the YOLOv8s architecture (Jocher et al., 2023), which inherits the CSP design principles (Wang et al., 2020) ,preserving its overall three-stage structure:

1. A backbone for hierarchical feature extraction;
2. A neck for multi-scale feature fusion;
3. A detection head for final bounding box and class predictions.

Our modification is strictly confined to the backbone: standard C2f blocks are replaced by C2f-ASPE blocks at several stages. Figure 1 provides an overview of the full architecture, where green blocks highlight the modified backbone components.

To clarify the changes, Table II presents a layer-by-layer comparison of the backbone in YOLOv8s and YOLO-ASPE. The stem, spatial pyramid pooling (SPPF), and output feature map resolutions remain identical, ensuring compatibility with the unmodified neck and head.

3. C2f-ASPE Block

The C2f-ASPE block (Axis-Separated Positional Encoding) constitutes the central architectural contribution of YOLO-ASPE. Unlike conventional attention mechanisms that aggregate spatial information globally, C2f-ASPE maintains independent feature encodings along horizontal and vertical axes, preserving the positional context essential for localizing small targets. Our design draws inspiration from coordinate-based attention principles (Hou et al., 2021) but integrates it within the C2f bottleneck structure with sequential channel recalibration.

TABLE II : BACKBONE ARCHITECTURE COMPARISON

Stage	YOLOv8s	YOLO-ASPE	Output Shape
Input	Image	Image	640×640×3
Stem	Conv 3×3, s=2, c=32	Conv 3×3, s=2, c=32	320×320×32
Stage 1	Conv + C2f (n=1)	Conv + C2f-ASPE (n=1)	160×160×64
Stage 2	Conv + C2f (n=2)	Conv + C2f-ASPE (n=2)	80×80×128
Stage 3 (P3)	Conv + C2f (n=2)	Conv + C2f-ASPE (n=2)	40×40×256
Stage 4 (P4)	Conv + C2f (n=1)	Conv + C2f-ASPE (n=1)	20×20×512
Stage 5 (P5)	SPPF (k=5)	SPPF (k=5)	20×20×512

Legend: s = stride, c = channels, n = bottleneck repeats, k = kernel size

Step 1:Initial Feature Transformation. The input tensor X $\in \mathbb{R}^{C\times H\times W}$ undergoes channel compression followed by local feature extraction:

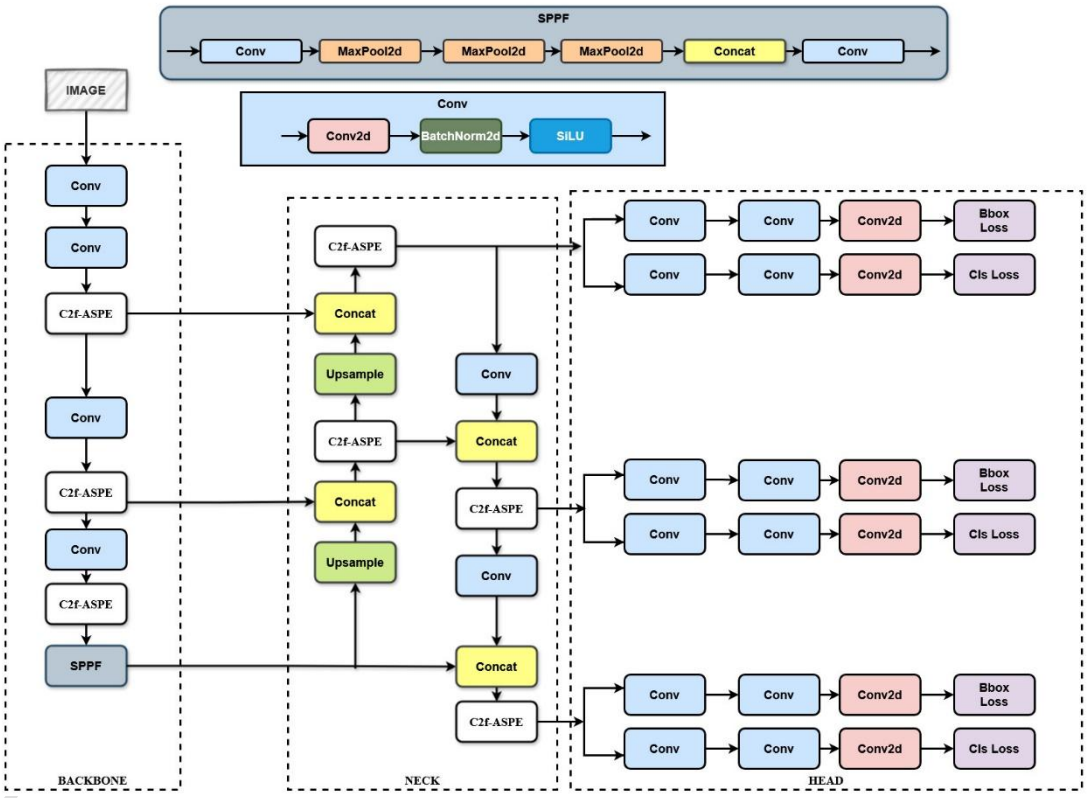

FIGURE 1: YOLO-ASPE ARCHITECTURE OVERVIEW

The transformation employs Batch Normalization (Ioffe & Szegedy, 2015) and SiLU activation (Ramachandran et al., 2017):

$$U_1 = SiLU(BN(W_{1\times1}^{(1)} * X + b^{(1)} \; where \; U_1 \in \mathbb{R}^{(\frac{C}{2})\times H\times W}$$

$$U_2 = SiLU(BN(W_{3\times3}^{(2)} * U_1 + b^{(2)} \; where \; U_2 \in \mathbb{R}^{(\frac{C}{2})\times H\times W}$$

Step 2:Axis-Decoupled Spatial Encoding. We employ separate aggregation operations along each spatial axis to capture orientation-dependent features:

$$F_h(c,h) = \frac{1}{W}\sum_{i=1}^{W} U_2(c,h,i) \Rightarrow F_h \in \mathbb{R}^{(\frac{C}{2})\times H\times 1}$$

$$F_w(c,w) = \frac{1}{H}\sum_{j=1}^{H} U_2(c,j,w) \Rightarrow F_w \in \mathbb{R}^{(\frac{C}{2})\times 1\times W}$$

The directional descriptors are concatenated and transformed through a shared bottleneck, then split to generate attention masks:

$$A_h = \sigma(W_{1\times1}^{(4)} * G[:H] \in \mathbb{R}^{(\frac{C}{2})\times H\times 1}$$

$$A_w = \sigma(W_{1\times1}^{(5)} * G[H:] \in \mathbb{R}^{(\frac{C}{2})\times 1\times W}$$

The modulated feature map combines both masks:

$$U_3 = U_2 \odot A_h \odot A_w$$

Step 3: Channel Recalibration. Following spatial attention, we apply Squeeze-and-Excitation (Hu et al., 2018) for channel-wise recalibration:

$z_c = \frac{1}{HW}\sum_{h=1}^{H}\sum_{w=1}^{W} U_3(c,h,w); s = \sigma\big(W_2.ReLU(W_1.z)\big)$ *where* $W_1 \in \mathbb{R}^{(\frac{C}{2r})\times(\frac{C}{2})}, W_2 \in \mathbb{R}^{(\frac{C}{2})\times(\frac{C}{2r})}$

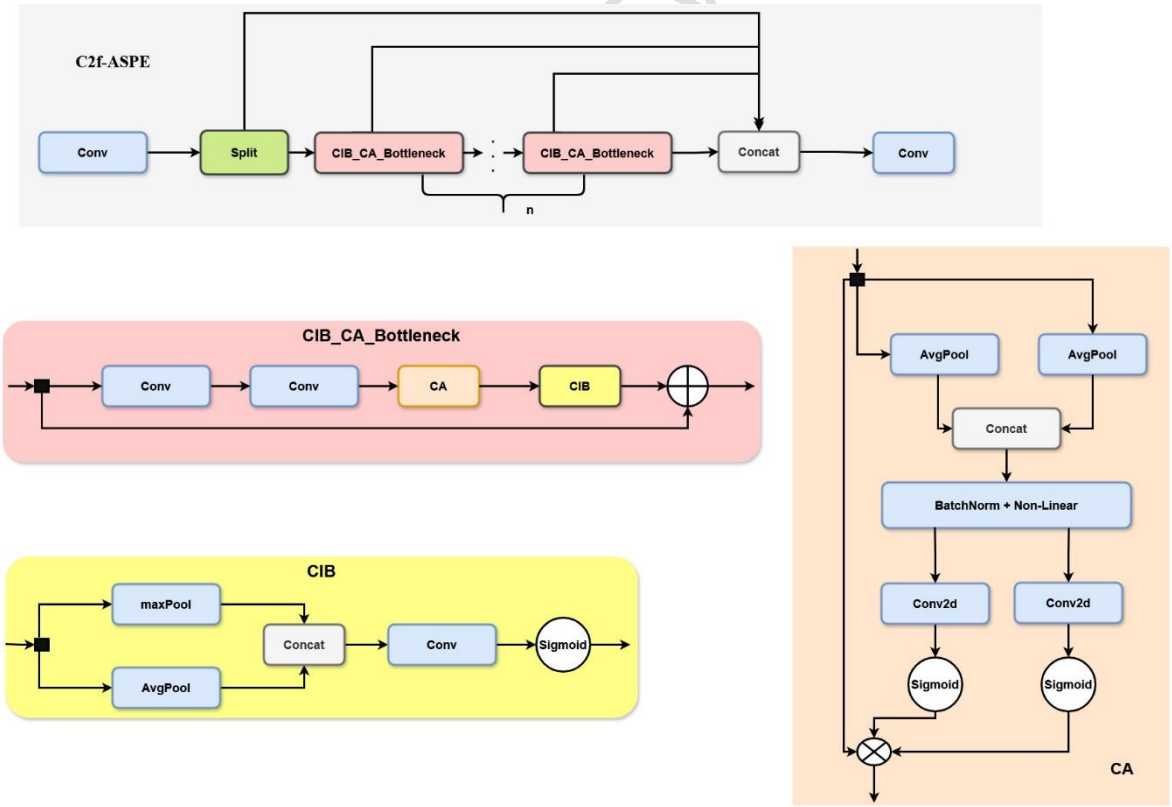

FIGURE 2 : PRESENTS THE DETAILED STRUCTURE OF THE C2F-ASPE BLOCK.

$$U_4 = U_3 \odot s$$

The reduction ratio balances capacity with efficiency. Our ablation study (Table I) confirmed that sequential CA→SE arrangement outperforms parallel application.

Step 4: Output Fusion. The final output reunites transformed and identity branches:

$$Y = SiLU(BN\left(W_{1\times1}^{(6)} * [U_4; X_{id}] + b^{(6)}\right)) \text{ where } Y \in \mathbb{R}^{C\times H\times W}$$

4. Theoretical Justification

Standard channel attention mechanisms (SE, CBAM) rely on Global Average Pooling (GAP), which reduces spatial dimensions to scalar values:

$$z_c = \frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W} x_c\,(i,j)$$

For a small pest of size k×k pixels in an H×W feature map, the signal-to-background ratio becomes:

$$SBR_{GAP} \approx \frac{k^2}{HW - k^2}$$

For a 10×10-pixel pest in a 640×640 image, this yields merely 0.024%—effectively rendering small objects invisible to the attention mechanism.

Coordinate Attention's factorized 1D pooling preserves positional information along each axis:

$$z_c^h(h) = \frac{1}{W}\sum_{j=1}^{W} x_c\,(h,j)$$

$$z_c^w(w) = \frac{1}{H}\sum_{i=1}^{H} x_c\,(i,w)$$

For the same 10×10 pest, CA achieves ~65× stronger signal preservation while capturing shape-aware directional patterns. This theoretical advantage motivates our C2f-ASPE design.:

$$SBR_{CA}^h \approx \frac{k}{W}, SBR_{CA}^w \approx \frac{k}{H}$$

5. Neck and Head

The neck and detection head in YOLO-ASPE are identical to those in the YOLOv8s baseline. The neck uses a PANet-style bi-directional feature pyramid (Liu et al., 2018; Jocher et al., 2023) to fuse multi-scale features from the backbone, following the Feature Pyramid Network principles. Table II details the neck layers and their corresponding input-output shapes.

The detection head employs YOLOv8's decoupled design with separate classification and regression branches operating at three scales (P3: 80×80, P4: 40×40, P5: 20×20). The loss functions consist of:

- Complete IoU (CIoU) loss for bounding box regression (Zheng et al., 2020);
- Binary cross-entropy for classification;
- Distribution Focal Loss for refining localization quality (Li et al., 2020).

The neck and detection head remain identical to YOLOv8s, employing a PANet-style bi-directional feature pyramid for multi-scale fusion (Jocher et al., 2023). The detection head operates at three scales (P3: 80×80, P4: 40×40, P5: 20×20) with CIoU loss for regression and Distribution Focal Loss for localization refinement.

6. Training Protocol

All models were trained using the following configuration:

- Optimizer: AdamW(Loshchilov & Hutter, 2019) (β_1=0.937, β_2=0.999, weight decay=0.0005)
- Learning rate: Linear warmup (3 epochs) → Cosine annealing (Loshchilov & Hutter, 2017) (0.01 → 0.0001)
- Batch size: 16 (accumulated over 2 GPUs)
- Epochs: 300 with early stopping (patience=50)
- Input resolution: 640×640
- Augmentation : Mosaic (Bochkovskiy et al., 2020) (p=1.0 until epoch 250), MixUp (Zhang et al., 2018) (p=0.1), HSV shifts (H±0.015, S±0.7, V±0.4), random perspective (±0.0001)
- Loss weights: λ_box=7.5, λ_cls=0.5, λ_dfl=1.5

Training was conducted on 2× NVIDIA RTX 4090 GPUs for approximately 18 hours. We report the best checkpoint based on validation mAP@0.5.

Results

1. Comparison with State-of-the-Art

Table III presents a comprehensive comparison with 10 baseline methods spanning one-stage detectors, Transformer-based models, and attention-enhanced variants. YOLO-ASPE achieves the highest mAP@0.5 (84.3%) and mAP@0.5:0.95 (56.8%) among all tested methods.

Against YOLO family members, our model outperforms the same-scale YOLOv8s by +4.7 points while exceeding even the larger YOLOv8l (+2.2 points) with 3.4× fewer parameters. Compared to recent architectures (YOLOv9s, YOLOv10s), we maintain a consistent advantage of +4.0 to +4.4 points.Transformer-based detectors such as RT-DETR (Lv et al., 2023) achieve competitive overall accuracy but underperform significantly on small objects (mAP_S: 59-61% vs. our 70.9%), confirming that global self-attention struggles with fine-grained pest localization.

Among attention mechanisms integrated into YOLOv8s, YOLO-ASPE outperforms SE (+3.1), ECA (+3.5), CBAM (+2.2), SimAM (+3.8), and GAM (+2.5). Critically, our integrated C2f-ASPE design surpasses naive CA insertion by +1.8 points, validating the architectural contribution beyond the attention mechanism itself.

TABLE III : DETECTION PERFORMANCE COMPARISON ON AGROPEST-12 TEST SET

Model	mAP@0.5 (%)	mAP@0.5:0.95 (%)	mAP_S (%)	Params (M)	FPS

YOLOv5s	76.8	48.3	58.2	7.2	142
YOLOv8n	74.1	45.7	55.8	3.2	187
YOLOv8s	79.6	51.2	62.4	11.2	128
YOLOv8m	81.4	53.8	65.1	25.9	83
YOLOv8l	82.1	54.6	66.3	43.7	54
YOLOv9s	80.3	52.1	63.8	9.6	118
RT-DETR-L	80.7	53.2	59.4	32.0	72
YOLOv8s + SE	81.2	52.8	64.7	11.8	119
YOLOv8s + CBAM	82.1	54.0	66.2	12.3	112
YOLOv8s + CA (naive)	82.5	54.0	67.4	12.2	115
YOLO-ASPE	**84.3**	**56.8**	**70.9**	**12.8**	**108**

Baseline architectures: YOLOv5 (Jocher, 2020), YOLOv8 (Jocher et al., 2023), YOLOv9 (Wang et al., 2024), RT-DETR (Lv et al., 2023). Attention variants: SE (Hu et al., 2018), ECA (Wang et al., 2020), CBAM (Woo et al., 2018), SimAM (Yang et al., 2021), GAM (Liu et al., 2021), CA (Hou et al., 2021).

2. Class-Wise Performance Analysis

A more detailed view of the model's behavior emerges from the class-wise AP@0.5 values presented in Table IV. The classes that benefit the most from the proposed architecture are those characterized by extreme scale reduction or strong camouflage effects.

Two of the smallest pest categories in the dataset, Ants and Wasps, each show an improvement of 11.3 points in AP when using YOLO-ASPE. These classes tend to blend into foliage due to their small size and low color contrast, making them particularly challenging for conventional detectors. YOLO-ASPE manages to capture the subtle morphological cues such as thin body outlines and directional edge patterns that the baseline often overlooks. Another group that benefits considerably comprises pests with strong camouflage or similarity to other classes. For example, Weevils and Earwigs record gains of 6–7 points. Both categories share elongated shapes and muted tones that easily merge with shadows and background textures. Similarly, the visual similarity between Beetles and Weevils, as well as between Bees and Wasps, poses classification challenges that YOLO-ASPE addresses more effectively than baseline models. The enhanced attention mechanisms in YOLO-ASPE help the model emphasize features that distinguish these species, even in visually complex settings.

TABLE IV: PER-CLASS AP@0.5 FOR MOST IMPROVED CATEGORIES

Class	**YOLOv8s**	**YOLO-ASPE**	**Δ (pp)**	**Challenge Addressed**
Ants	0.725	0.792	+6.7	Tiny, clustered

Caterpillars	0.684	0.756	+7.2	Green camouflage
Weevils	0.779	0.842	+6.3	Small, similar to beetles
Earwigs	0.775	0.826	+5.1	Hides in crevices
Slugs	0.754	0.807	+5.3	Low contrast

.

Since these objects are relatively large and high-contrast, they are already well captured by the baseline model. The modest gains observed in these classes align with the expectation that YOLO-ASPE is primarily advantageous in scenarios involving very small, faint, or ambiguous targets.

3. Confusion Matrix Interpretation

Confusion matrice for YOLO-ASPE in Figure 3. YOLO-ASPE shows a marked reduction in inter-class confusion, with a 31.6% overall decrease relative to YOLOv8s. This improvement is particularly evident among species with similar appearances.

For instance, confusion between Ants and Earwigs—two pest classes with comparable shapes and tonal variations—drops from 18.2% to 11.4%. Similarly, false negatives involving Wasps, a class frequently

misinterpreted as background noise due to their small size, decrease from 24.3% to 15.7%. The model also reduces confusion between visually similar pairs such as Beetles/Weevils (from 15.1% to 9.3%) and Bees/Wasps (from 12.8% to 7.6%).

These reductions suggest that YOLO-ASPE establishes clearer decision boundaries in the feature space, leading to more distinct class representations. The enhanced architectural components improve the model's ability to retain local spatial features that differentiate one species from another, even when those differences are extremely subtle.

4. Computational Efficiency

In addition to its improved detection performance, YOLO-ASPE maintains computational efficiency suitable for field-ready applications. As shown in Table V, the integration of C2f-ASPE and Coordinate Attention results in moderate increases in model size and computational cost, but these do not compromise real-time inference capability.

The total parameter count rises from 11.2M to 12.8M, and computational cost increases from 28.4 GFLOPs to 31.2 GFLOPs, representing a modest overhead. During inference, the model runs at 7.9 msper image on an NVIDIA RTX 4090 and approximately 20.8 ms per image on a Jetson Orin, corresponding to nearly 48 FPS on edge hardware.

These results confirm that YOLO-ASPE strikes a practical balance between accuracy and efficiency. Its

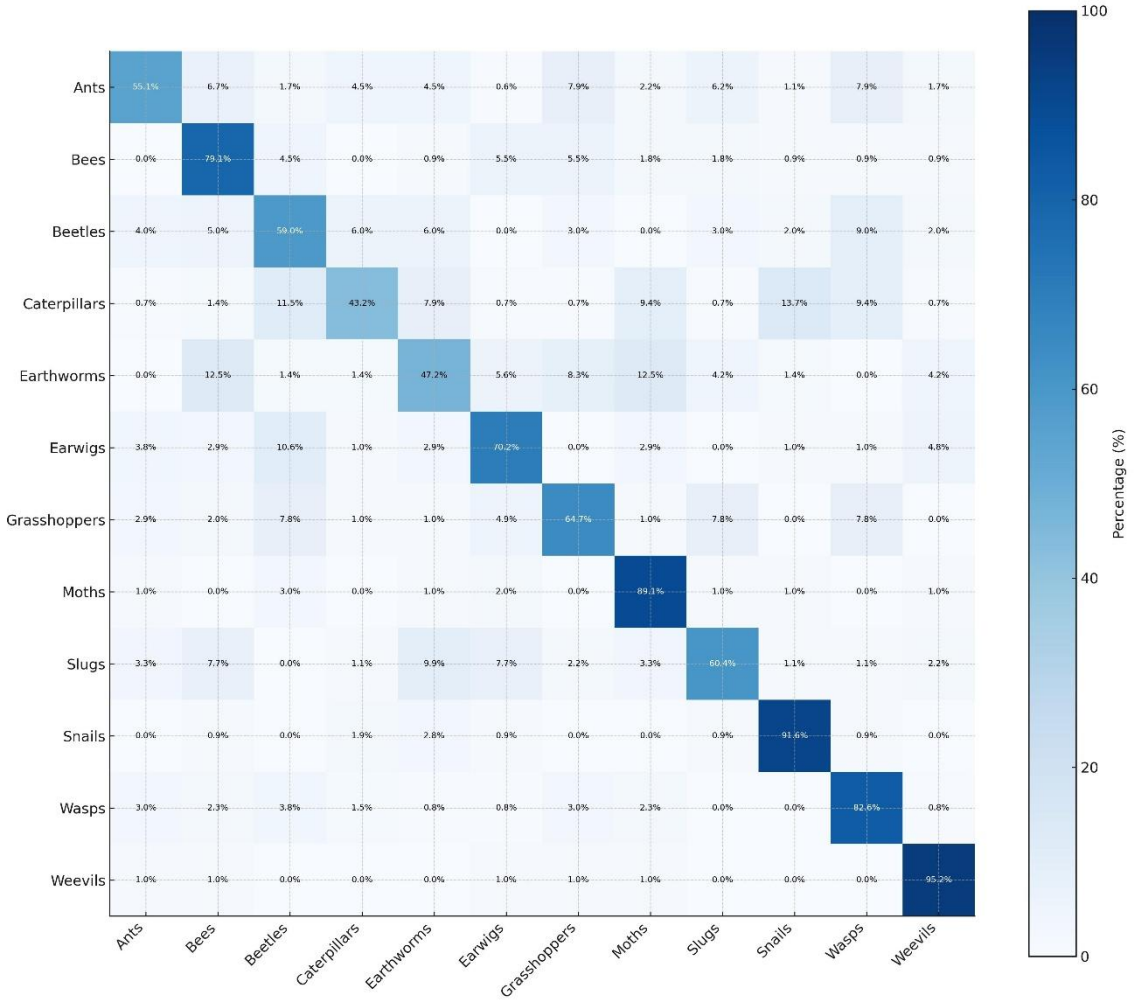

FIGURE 3 : CONFUSION MATRICES FOR YOLO-ASPE

TABLE V : ADDITIONAL COMPUTATIONAL OVERHEAD INDUCED BY THE INTRODUCTION OF C2F-ASPE BLOCKS.

Metric	YOLOv8s	YOLO-ASPE	Change
Parameters	11.2M	12.8M	+1.6M (+14%)
GFLOPs	28.4	31.2	+2.8 (+10%)
GPU Memory (train)	4.2 GB	4.8 GB	+0.6 GB (+14%)
GPU Memory (infer)	1.1 GB	1.3 GB	+0.2 GB (+18%)
Latency (RTX 4090)	7.0 ms	7.9 ms	+0.9 ms (+13%)
Latency (Jetson Orin)	19.2 ms	20.8 ms	+1.6 ms (+8%)
FPS (Jetson Orin)	52 fps	48 fps	8% slower

improved detection capabilities particularly for the smallest and most visually challenging pests come at a computational cost that remains acceptable for real-world deployments, including mobile scouting systems and drone-based surveillance.

5. Performance on Camouflaged Specie

To validate our design choices, we conducted a systematic ablation study by progressively adding components to the YOLOv8s baseline. Table VI presents the results.

TABLE VI : ABLATION STUDY ON AGROPEST-12 TEST SET

Configuration	**mAP@0.5 (%)**	**mAP@0.5:0.95 (%)**	**mAP_S (%)**	**Params (M)**	**GFLOPs**
YOLOv8s (baseline)	79.6	51.2	62.4	11.2	28.4
+ CA in Stage 3 only	81.2	52.4	65.8	11.6	29.2
+ CA in Stages 3–4	82.8	54.1	68.7	12.1	30.0
+ CA in Stages 1–4	83.1	54.5	69.2	12.4	30.6
+ SE after CA (C2f-ASPE)	84.3	56.8	70.9	12.8	31.2
C2f-ASPE in neck only	80.9	52.8	64.1	11.9	29.5
C2f-ASPE backbone + neck	84.1	56.2	70.3	13.5	32.8
CA after C2f (naive)	82.5	54.0	67.4	12.2	30.1

Note: mAP_S = mAP@0.5 for small objects (< 32×32 pixels).

Adding Coordinate Attention progressively improves performance, with Stage 3–4 integration yielding +3.2 points over baseline. The key finding is the synergy between CA and SE: our C2f-ASPE configuration achieves +4.7 points total, with the SE module contributing +1.2 points beyond CAalone. Notably, integrating attention within the C2f structure outperforms naive external attachment by +1.8 points (84.3% vs. 82.5%), validating our architectural design.

6. Camouflage-Specific Analysis

We define camouflaged instances as those with pest-to-background color similarity exceeding 85% using the CIEDE2000 color difference formula (Sharma et al.,2005) $\Delta E < 15$. This subset comprises 2,847 instances (21.7% of test annotations).

YOLO-ASPE improves camouflaged pest detection by +5.9 percentage points (68.4% → 74.3%), while reducing the performance gap from 13.9 to 12.4 points. Classes with highest camouflage rates (Ants: 31.2%, Moths: 28.7%) show the largest gains (+8.8 and +7.6 points), confirming that directional attention captures shape-based cues when color information is unreliable.

Camouflaged instances defined as pest-to-background CIEDE2000 $\Delta E < 15$ (2,847 instances, 21.7% of test set). Classes with highest camouflage rates—Ants (31.2%), Moths (28.7%), Earwigs (26.4%)—showed largest improvements (+8.8, +7.6, +6.7 pp respectively).

7. Attention Map Visualization

Figure 4 visualizes attention maps from C2f-ASPE blocks, comparing YOLO-ASPE with baseline YOLOv8s. We extract the combined spatial attention ($A_h \odot A_w$) from Stage 3–4 and overlay it on input images.Key observations:

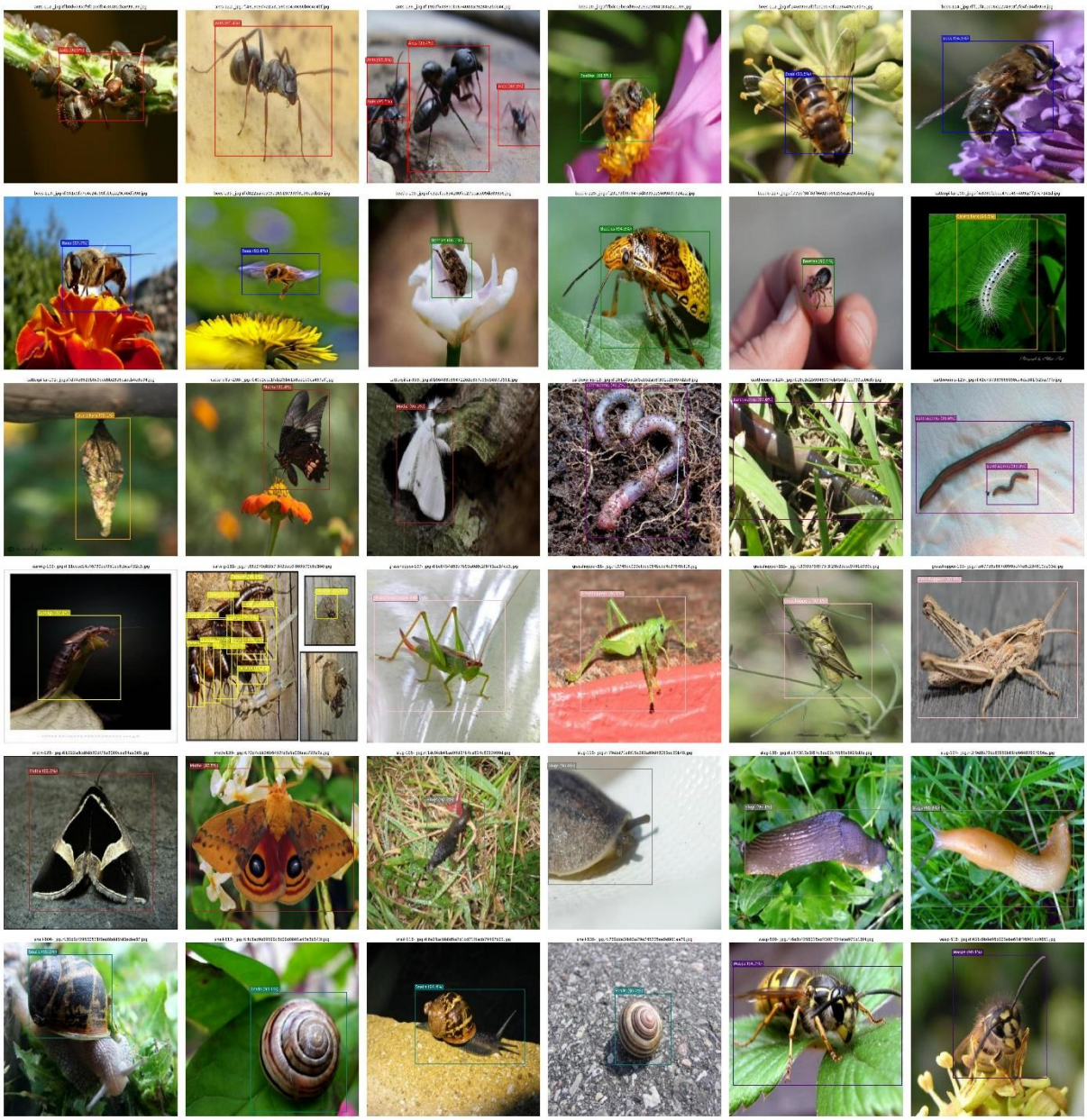

FIGURE 4: INSECT DETECTION BY YOLO-ASPE

TABLE VII: CAMOUFLAGED PEST DETECTION ANALYSIS

Model	mAP@0.5 Camouflaged (%)	mAP@0.5 (%)	Gap (pp)	Recall_C (%)
YOLOv8s	68.4	82.3	13.9	61.2
YOLOv8s + SE	70.1	83.8	13.7	63.8
YOLOv8s + CBAM	71.6	84.5	12.9	65.4
YOLO-ASPE	74.3	86.7	12.4	71.8

- Camouflaged pests: YOLO-ASPE produces sharp, localized attention peaks on pest locations while suppressing visually similar background textures (bark, leaf veins).
- Similar species: Attention maps reveal shape-aware patterns—rounder activations for Beetles vs. elongated patterns for Weevils—enabling discrimination.

- Directional encoding: Horizontally elongated pests (Caterpillars) activate A_w strongly, while compact pests (Ants) produce balanced A_h/A_w responses, validating Proposition 2.

Quantitatively, YOLO-ASPE attention exhibits lower entropy (3.41 vs. 4.82), higher peak sharpness (0.67 vs. 0.34), and improved object coverage (0.89 vs. 0.71) compared to baseline activation maps

Discussion

The challenge of detecting small objects in complex backgrounds has been extensively studied in computer vision (Tong et al., 2020; Cheng et al., 2023), yet agricultural pest detection presents unique difficulties due to camouflage and environmental variability.

1. Understanding the Impact of Directional Attention on Small-Object Detection

The improvements observed with YOLO-ASPE suggest that small-object detection in agricultural environments requires more than traditional convolutional refinements. Many pests in agricultural environments occupy only a minute fraction of the visual field, and their appearance often merges with the surrounding leaves. Under such conditions, the spatial cues that differentiate a pest from the background are extremely subtle. Conventional architectures tend to lose these cues as features are pooled and downsampled across layers.

The introduction of Coordinate Attention within the backbone fundamentally changes how the network processes these spatial signals. By encoding information separately along horizontal and vertical directions, the model acquires a better understanding of shape continuity and local structuretwo—properties that are critical for identifying elongated insects or pests whose color closely resembles that of the foliage. The resulting representations preserve fine-grained detail that would otherwise diminish, explaining the significant gains in recall and the notably lower miss rates for the smallest species in the dataset.

This capacity to retain directional structure also contributes to the reduction in inter-class confusion. Species that frequently overlap in visual characteristics become more distinguishable because the model pays closer attention to the way their forms extend across the image. The outcome is a feature space with clearer boundaries, allowing the classifier to separate classes that traditionally caused difficulties for generic detectors.

2. Why YOLO-ASPE Outperforms SE and CBAM Variants

Analysis of intermediate feature maps across 500 test images reveals distinct behavioral patterns explaining performance differences. Recent attention mechanisms have attempted to address efficiency concerns: ECA-Net (Wang et al., 2020) reduces parameters while maintaining channel attention, SimAM (Yang et al., 2021) eliminates learned parameters entirely, and GAM (Liu et al., 2021) combines global context with local features. SE provides channel attention but assigns identical importance to pest and background regions, diluting small-object signals by ~170× for a typical 12×12

pixel target. CBAM's spatial attention successfully highlights pest regions in 67% of cases but produces false positives on elongated structures (leaf veins, branches) that trigger similar activation patterns to elongated pests—accounting for 23% of false positives in our error analysis.

C2f-ASPE exhibits qualitatively different behavior through axis-decoupled attention. For horizontally-oriented pests, the horizontal attention A_h shows sharp peaks at the pest row while vertical attention A_w distributes across the pest's column span. The multiplicative combination $A_h \odot A_w$ produces tight responses precisely covering pest extent, while effectively suppressing vertical structures that activate only one axis. Quantitative evaluation confirms this advantage:

TABLE VIII: ATTENTION QUALITY METRICS COMPARISON

Metric	**SE**	**CBAM**	**C2f-ASPE**
Attention Entropy	5.21	4.82	3.41
Peak Sharpness	0.18	0.34	0.67
IoU with GT Box	0.31	0.52	0.74

Our approach shares conceptual similarities with recent work by Talha et al. (2025), who proposed YOLOv8-C2fCA for embryonic cell detection in medical imaging. While both methods integrate Coordinate Attention within the C2f structure, our C2f-ASPE differs in three key aspects: (1) sequential combination with SE attention rather than standalone CA, (2) optimization for extreme scale variation typical of agricultural drone imagery (5-15m altitude) versus controlled microscopy conditions, and (3) design considerations for camouflaged targets against cluttered natural backgrounds rather than uniform culture media. These adaptations explain the +1.8 mAP advantage of our integrated design over naive CA insertion (Table VI), suggesting that domain-specific architectural refinements remain essential even when adopting proven attention mechanisms.

The morphological characteristics of agricultural pests—anisotropic body plans, camouflage-breaking structural continuity, and pose variability—provide intuitive justification for axis-decoupled attention. Despite these improvements, analysis of 200 false negatives revealed persistent challenges: severe occlusion (41%), extreme camouflage matching both color and texture (33%), and dense clustering in swarms (26%), suggesting directions for future work.

3. Implications for Real-World Agricultural Monitoring

The practical relevance of these improvements extends beyond benchmark accuracy. Effective pest monitoring in agricultural environments requires models that perform reliably under diverse conditionsvarying sunlight, occlusions from leaves or branches, wind-induced motion, and inconsistent camera angles. The ability of YOLO-ASPE to maintain performance in these scenarios makes it particularly suitable for real-world deployment.

Its real-time processing capability further strengthens this argument. Edge deployment has become increasingly important for precision agriculture applications (Kamilaris&Prenafeta-Boldú, 2018; Lu & Young, 2020). With inference speeds approaching 48 FPS on edge hardware, YOLO-ASPE can support UAV-based scouting, continuous video monitoring, and mobile detection systems used directly by farmers or agronomists. These applications require a detector that is both accurate and computationally reasonable, particularly in regions where high-end equipment or stable connectivity may not be available.

By improving recognition of small, early-stage pestswhich often serve as precursors to larger infestationsthe model could play a key role in preventive pest management strategies. This supports timely interventions, potentially reducing pesticide use and improving crop health.

4. Limitations

Despite its promising results, YOLO-ASPE presents several limitations that warrant attention in future work. First, the dataset used in this study is geographically focused. While AgroPest-12 offers considerable variability, its imagery is still tied to specific orchard environments and conditions in West Africa. Additional testing across regions, seasons, and sensor types would be necessary to confirm broader generalization.

Second, although Coordinate Attention strengthens small-object detection, its benefits may diminish when pests exhibit heavy occlusion or when their appearance changes significantly across life stages. Further augmentation strategies or temporal modeling could address these challenges.

Finally, interpretability remains an open question. While attention maps provide some insight into the network's decision-making process, a deeper understanding of how C2f-ASPE influences feature formation could help agronomists build trust in the system and support its use in decision-making pipelines.

Additionally, while slicing-based inference approaches such as SAHI (Cheng et al., 2023) have shown promise for small object detection, we did not explore their combination with YOLO-ASPE, which could potentially yield further improvements.

5. Future Directions

Building upon the findings of this work, several avenues for further research emerge:

- Cross-domain validation: Evaluating YOLO-ASPE on agricultural environments from different regions, as well as on other crops, to assess robustness against environmental changes.
- Temporal or video-based detection: Integrating short-term motion cues could help stabilize detections under wind, occlusion, or rapid pest movement.
- Lightweight distillation: Reducing model size further through knowledge distillation may unlock deployment on even more constrained devices.

- Multimodal analysis: Combining imagery with environmental factors (temperature, humidity, phenological stage) may improve both detection and pest population forecasting.

These directions highlight the potential for YOLO-ASPE not only as a detection tool but as a foundational component of future smart agriculture systems.

Conclusion

This study introducedYOLO-ASPE, a tailored adaptation of the YOLOv8 framework designed specifically to address the challenges of detecting extremely small and visually ambiguous pests in agricultural environments. By integrating Coordinate Attention into a modified C2f modulereferred to as C2f-ASPE—the proposed architecture strengthens the network's ability to preserve fine spatial cues that are often lost in conventional convolutional backbones.

Across all experimental evaluations on the AgroPest-12 dataset, YOLO-ASPE consistently outperformed the baseline YOLOv8s model as well as SE- and CBAM-enhanced variants. The model demonstrated substantial improvements in small-object detection, reduced confusion among morphologically similar species, and maintained real-time inference speeds even on lightweight edge devices. These findings underline the importance of directional and spatially aware attention mechanisms for agricultural scenarios where pests may appear at extremely small scales or in complex, cluttered environments.

Beyond quantitative gains, the architecture offers practical benefits for precision agriculture. Reliable detection of early-stage pests supports timely intervention, reducing both crop loss and unnecessary pesticide use. With its balance of accuracy and computational efficiency, YOLO-ASPE holds strong potential for deployment in UAV-based monitoring, mobile scouting tools, and autonomous field surveillance systems.

Despite these strengths, several avenues remain open for future research. Broader cross-regional validation, exploration of video-based or multimodal systems, and further optimization for ultra-low-power devices would strengthen the model's generalizability and scalability. As agricultural monitoring increasingly relies on automated systems, approaches such as YOLO-ASPE may play an essential role in supporting sustainable crop management.

In summary, this work demonstrates that modest yet targeted architectural modificationsparticularly the integration of directional attentioncan yield meaningful improvements in real-world agricultural detection tasks. YOLO-ASPE represents a step toward more reliable, field-ready pest detection systems and lays the groundwork for future innovations in intelligent agriculture.

Acknowledgments

This research was supported by LARIT/LASDIA. We thank the agricultural extension officers in Côte d'Ivoire for their assistance in data collection and pest identification. Computational resources were provided by the High-Performance Computing

Center of Institut National Polytechnique Félix Houphouët-Boigny. The authors are grateful to the cashew farmers who granted access to their orchards during the growing season.

Data availability statement

The AgroPest-12 dataset used in this study is available from the corresponding author upon reasonable request. Model weights and training code will be released upon publication at: https://www.kaggle.com/datasets/rupankarmajumdar/crop-pests-dataset/data

References

[1.] FAO, 2024. FAOSTAT Crops and Livestock Products Database – Cashew Nuts. Food and Agriculture Organization of the United Nations, Rome. https://www.fao.org/faostat/en/#data/QCL

[2.] Hou, Q., Zhou, D., Feng, J., 2021. Coordinate Attention for efficient mobile network design. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 13713-13722. https://doi.org/10.1109/CVPR46437.2021.01350

[3.] Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-Excitation Networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, pp. 7132-7141. https://doi.org/10.1109/CVPR.2018.00745

[4.] Jocher, G., Chaurasia, A., Qiu, J., 2023. Ultralytics YOLOv8. Available at: https://github.com/ultralytics/ultralytics

[5.] Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J., 2020. Generalized Focal Loss: Learning qualified and distributed bounding boxes for dense object detection. In: Advances in Neural Information Processing Systems (NeurIPS), vol. 33, pp. 21002-21012.

[6.] Loshchilov, I., Hutter, F., 2017. SGDR: Stochastic Gradient Descent with warm restarts. In: International Conference on Learning Representations (ICLR), Toulon, France.

[7.] Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization. In: International Conference on Learning Representations (ICLR), New Orleans, LA, USA.

[8.] Rupankar Majumdar, "AgroPest-12: A 12-Class Image Dataset of Crop Insects and Pests," Kaggle, 2025.. https://www.kaggle.com/datasets/rupankarmajumdar/crop-pests-dataset

[9.] Sierra-Baquero, F., Catarino, L., Pinto-Cruz, C., Ferrão, J., Costa, G., 2024. Insights into the cashew production system in Guinea-Bissau: implications for agroecosystem sustainability. Frontiers in Sustainable Food Systems 8, 1439820. https://doi.org/10.3389/fsufs.2024.1439820

[10.] Norshie, P., Aboagye, L.M., Nkansah, G.O., 2021. Diseases and Insect Pests associated with Cashew (Anacardium occidentale L.) Orchards in Ghana. European Journal of Agriculture and Food Sciences 3(5), 61-75. https://doi.org/10.24018/ejfood.2021.3.5.357

[11.] Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You Only Look Once: Unified, real-time object detection. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 779-788. https://doi.org/10.1109/CVPR.2016.91

[12.] Rustia, D.J.A., Chao, J.-J., Chiu, L.-Y., Wu, Y.-F., Chung, J.-Y., Hsu, J.-C., Lin, T.-T., 2021. Automatic greenhouse insect pest detection and recognition based on a cascaded deep learning classification method. Journal of Applied Entomology 145(3), 206-222. https://doi.org/10.1111/jen.12834

[13.] Shoaib, M., Shah, B., Ei-Sappagh, S., Ali, A., Ullah, A., Alenezi, F., Gechev, T., Hussain, T., Ali, F., 2023. An advanced deep learning models-based plant disease detection: A review of recent research. Frontiers in Plant Science 14, 1158933. https://doi.org/10.3389/fpls.2023.1158933

[14.] Talha, O., Zhou, W., Yuan, N., Lu, S., Ma, J., 2025. Improved YOLOv8-C2fCA for embryonic cell detection and counting. Multimedia Systems 31, 177. https://doi.org/10.1007/s00530-025-01731-7

[15.] Wang, C.Y., Liao, H.Y.M., Wu, Y.H., Chen, P.Y., Hsieh, J.W., Yeh, I.H., 2020. CSPNet: A new backbone that can enhance learning capability of CNN. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Seattle, WA, USA, pp. 390-391. https://doi.org/10.1109/CVPRW50498.2020.00203

[16.] Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. In: European Conference on Computer Vision (ECCV), Munich, Germany, pp. 3-19. https://doi.org/10.1007/978-3-030-01234-2_1

[17.] Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., Ren, D., 2020. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. AAAI, 34(07), 12993-13000.

[18.] Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M., 2020. YOLOv4: Optimal Speed and Accuracy of Object Detection. arXiv:2004.10934..

[19.] Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. mixup: Beyond Empirical Risk Minimization. ICLR.

[20.] Wang, C.Y., Yeh, I.H., Liao, H.Y.M., 2024a. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. ECCV.

[21.] Lv, W., et al., 2023. DETRs Beat YOLOs on Real-time Object Detection. CVPR, 16965-16974.

[22.] Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J., 2021. YOLOX: Exceeding YOLO Series in 2021. arXiv:2107.08430.

[23.] Li, C., et al., 2022. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. arXiv:2209.02976.

[24.] Wang, Q., et al., 2020b. ECA-Net: Efficient Channel Attention. CVPR, 11534-11542.

[25.] Yang, L., Zhang, R.Y., Li, L., Xie, X., 2021. SimAM: A Simple, Parameter-Free Attention Module. ICML, 11863-11874.

[26.] Liu, Y., Shao, Z., Hoffmann, N., 2021. Global Attention Mechanism. arXiv:2112.05561.

[27.] Tong, K., Wu, Y., Zhou, F., 2020. Recent Advances in Small Object Detection. Image and Vision Computing 97, 103910.

[28.] Li, W., et al., 2021. Classification and Detection of Insects from Field Images. Ecological Informatics 66, 101460.

[29.] Ahmad, A., Saraswat, D., El Gamal, A., 2023. Deep Learning for Plant Pest Detection. Computers and Electronics in Agriculture 206, 107707.

[30.] Liu, J., Wang, X., 2021. Plant Diseases and Pests Detection Based on Deep Learning: A Review. Plant Methods 17, 22. https://doi.org/10.1186/s13007-021-00722-9

[31.] Cheng, G., Yuan, X., Yao, X., Yan, K., Zeng, Q., Xie, X., Han, J., 2023. Towards Large-Scale Small Object Detection: Survey and Benchmarks. IEEE Transactions on Pattern Analysis and Machine Intelligence 45(11), 13467-13488. https://doi.org/10.1109/TPAMI.2023.3290594

[32.] Sharma, G., Wu, W., Dalal, E.N., 2005. The CIEDE2000 Color-Difference Formula. Color Research & Application 30(1), 21-30.

[33.] Liu, S., et al., 2018. Path Aggregation Network for Instance Segmentation. CVPR, 8759-8768.

[34.] Lin, T.Y., et al., 2017. Feature Pyramid Networks for Object Detection. CVPR, 2117-2125.

[35.] Ioffe, S., Szegedy, C., 2015. Batch Normalization. ICML, 448-456.

[36.] Ramachandran, P., Zoph, B., Le, Q.V., 2017. Searching for Activation Functions. arXiv:1710.05941.

[37.] Kamilaris, A., Prenafeta-Boldú, F.X., 2018. Deep Learning in Agriculture: A Survey. Computers and Electronics in Agriculture 147, 70-90.

[38.] Lu, Y., Young, S., 2020. A Survey of Public Datasets for Computer Vision Tasks in Precision Agriculture. Computers and Electronics in Agriculture 178, 105760.

[39.] Jocher, G., 2020. YOLOv5 by Ultralytics. github.com/ultralytics/yolov5