Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(6), 209-235
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

CASA-YOLO: A Unified Framework for Small and Camouflaged Object Detection in Agricultural Pest Imagery

Koffi Bernadin-Pacome Sayni[1*], Apo Chimène Monsan[2], Mamadou Diarra[3], Beman Hamidja Kamagaté[4], Souleymane Oumtanaga[1]

[1]Institut National Polytechnique Félix Houphouët-Boigny, Abidjan, Côte d'Ivoire
[2]Université Virtuelle de Côte d'Ivoire, Abidjan, Côte d'Ivoire
[3]Université Félix Houphouët-Boigny, Abidjan, Côte d'Ivoire
[4]Ecole Supérieure Africaine Des Tic, Abidjan, Côte d'Ivoire
Email: *saynikoffi2022@gmail.com, chmonsan@gmail.com, patoudiarra@gmail.com, beman2017@gmail.com, oumtana@gmail.comn

How to cite this paper: Sayni, K. B.-P., Monsan, A. C., Diarra, M., Kamagaté, B. H., & Oumtanaga, S. (2026). CASA-YOLO: A Unified Framework for Small and Camouflaged Object Detection in Agricultural Pest Imagery. *Open Journal of Social Sciences, 14,* 209-235.
https://doi.org/10.4236/jss.2026.146012

Received: January 23, 2026
Accepted: June 14, 2026
Published: June 17, 2026

Abstract

Small object detection (SOD) and camouflaged object detection (COD) are critical challenges in agricultural computer vision, where pests exhibit both spatial compactness and visual similarity to their surroundings. Existing approaches address these problems in isolation, failing to exploit their shared characteristic: extracting weak visual signals from low signal-to-noise environments. This paper introduces CASA-YOLO (Context-Aware Sparse Attention YOLO), a unified framework addressing SOD and COD through three innovations: 1) Dual-Axis Sparse Attention (DASA), which decomposes global attention into axis-wise operations with adaptive sparse sampling, reducing complexity from O(N^2) to O($N\sqrt{N}/s$); 2) Adaptive Context Gating (ACG), a three-pathway module dynamically balancing local texture, global semantics, and boundary cues; and 3) HFPN-Nano, a hierarchical feature pyramid enabling stride-4 detection of objects as small as 8 × 8 pixels. On the AgroPest-12 benchmark, CASA-YOLO achieves 89.6% mAP@50 and 58.3% mAP@50 - 95, surpassing YOLOv11s (+5.9% mAP@50) and RT-DETR-R18 (+3.3%) at FP32 precision, while maintaining real-time inference (118 FPS with TensorRT INT8 quantization). Field validation on cashew plantations across three regions in Côte d'Ivoire (895 images, 8 sites) confirms practical applicability. Camouflage-stratified analysis further shows that ACG provides significant gains on high-camouflage instances, validating the unified SOD-COD design philosophy for agricultural pest detection.

Keywords

Object Detection, Small Object Detection, Camouflaged Object Detection, Attention Mechanism, Agricultural Computer Vision, Deep Learning, Pest Detection, Cashew Tree

1. Introduction

The detection of small and visually inconspicuous objects constitutes a fundamental challenge in computer vision with far-reaching implications for precision agriculture, autonomous systems, and medical imaging. In agricultural contexts, early detection of crop pests and diseases is paramount: the Food and Agriculture Organization estimates that plant pests and diseases cause annual economic losses exceeding $220 billion globally, with 20% - 40% of crop production lost to these threats (FAO, 2019). This challenge is compounded by the inherent visual characteristics of agricultural threats: many pests measure merely 2 - 5 mm in length, while fungal infections often manifest as subtle discolorations that blend seamlessly with healthy foliage.

Two distinct research communities have emerged to address related aspects of this challenge. Small Object Detection (SOD) focuses on identifying targets occupying minimal image area, typically defined as objects smaller than 32×32 pixels according to COCO terminology (Lin et al., 2014). The primary difficulties include limited discriminative features, sensitivity to localization errors, and severe class imbalance during training. Conversely, Camouflaged Object Detection (COD) addresses objects that deliberately or naturally conceal themselves within their environment through texture similarity, boundary diffusion, or pattern mimicry (Fan, Ji, Sun, et al., 2020). COD challenges arise from semantic ambiguity between foreground and background, rather than from spatial limitations.

Despite their distinct origins, we observe that SOD and COD share a fundamental characteristic: both require extracting weak visual signals from environments where the signal-to-noise ratio is inherently low. In SOD, the signal is spatially compressed; in COD, it is semantically obscured. This observation motivates our central hypothesis: attention mechanisms designed for positional precision (addressing SOD) can synergize with gating mechanisms for foreground-background separation (addressing COD), thereby enabling a unified detection framework that benefits from both SOD and COD design principles. While our evaluation focuses on agricultural pest detection (which inherently combines SOD and COD challenges), dedicated evaluation on standard COD segmentation benchmarks remains a direction for future work.

Current state-of-the-art detectors exhibit significant limitations when confronted with agricultural imagery. Transformer-based architectures such as RT-DETR (Zhao et al., 2024) achieve impressive accuracy but demand substantial computational resources, making them incompatible with edge deployment on

agricultural drones. YOLO variants (Wang, Yeh, & Liao, 2024; Ultralytics, 2024) offer real-time performance but rely on attention mechanisms that either lack sufficient granularity for small objects or impose prohibitive O(N^2) complexity on high-resolution feature maps. Dedicated COD methods (Fan, Ji, Cheng, & Shao, 2022; Mei et al., 2021a) achieve remarkable performance on benchmark datasets but are designed for segmentation rather than detection and lack the efficiency required for real-time applications.

This paper makes the following contributions: we propose CASA-YOLO, a novel object detection architecture that unifies small object detection and camouflaged object detection through principled attention design, representing, to the best of our knowledge, the first real-time detection framework explicitly designed to address both SOD and COD challenges simultaneously within a unified architecture. Our technical innovations include Dual-Axis Sparse Attention (DASA), which reduces attention complexity from O(N^2) to O($N\sqrt{N}$) through sequential axis-wise decomposition. Adaptive sparse sampling with learned stride s further reduces the effective complexity to O($N\sqrt{N}/s$), enabling efficient processing of high-resolution feature maps critical for small object detection; Adaptive Context Gating (ACG) a three-pathway module incorporating local texture analysis, global semantic encoding, and boundary enhancement with learned competitive gating; and HFPN-Nano, an efficient hierarchical feature pyramid with stride-4 detection capabilities adding only 26% computational overhead.

We validate our approach through comprehensive experiments on the AgroPest-12 dataset (Majumdar, 2025) and multi-site field experiments on cashew plantations across three regions in Côte d'Ivoire (Lapinkro, Touba, and Kotobi), demonstrating state-of-the-art performance with practical applicability under diverse real-world conditions.

The remainder of this paper is organized as follows: Section 2 reviews related work in SOD, COD, and attention mechanisms. Section 3 presents the proposed CASA-YOLO architecture in detail. Section 4 describes our experimental methodology and datasets. Section 5 presents comprehensive results, ablation studies, and field validation experiments. Section 6 concludes with future research directions.

2. Related Work

2.1. Small Object Detection

Small object detection has evolved along three principal paradigms: multi-scale feature learning, data augmentation strategies, and specialized network architectures. The Feature Pyramid Network (FPN) (Lin et al., 2017) established the foundation for multi-scale detection by constructing a top-down pathway that propagates semantic information to higher-resolution features. Subsequent works, including Liu et al. (2018) and BiFPN (Tan et al., 2020), enhanced feature fusion through bidirectional connections and weighted aggregation, respectively.

Data-centric approaches address the statistical challenges of small object detection. Copy-Paste augmentation (Ghiasi et al., 2021) increases small object instance

density by compositing objects onto varied backgrounds. SNIP (Singh & Davis, 2018) and SNIPER (Singh, Najibi, & Davis, 2018) introduced scale-specific training that selectively backpropagates gradients based on object size, preventing gradient domination by larger objects. Data-centric strategies can yield significant gains: (Kisantal et al., 2019) demonstrated that oversampling combined with copy-paste augmentation improves small object detection AP by up to 9.7% without architectural modifications, while Bochkovskiy et al. (2020) showed that mosaic augmentation compositing four training images into one further enriches small object context during training.

Architectural innovations specifically targeting small objects include QueryDet (Yang et al., 2022), which employs cascade sparse queries to progressively refine small object proposals, achieving significant improvements on VisDrone while maintaining efficiency. RFLA (Xu et al., 2022) introduces receptive field adaptation that dynamically adjusts convolutional kernels based on target scale. Wang et al. (2025) propose a dedicated small object detection head with deformable attention operating on P2-level features. Despite these advances, existing SOD methods do not account for camouflage scenarios in which small objects additionally exhibit visual similarity to their backgrounds.

2.2. Camouflaged Object Detection

Camouflaged object detection has experienced rapid progress following the introduction of large-scale benchmarks. Fan et al. (2020) released COD10K with 10,000 images spanning 78 categories and proposed SINet, establishing the search-and-identify paradigm where coarse localization precedes fine segmentation. PFNet (Mei et al., 2021b) extended this approach with positioning and focus modules that progressively refine camouflaged boundaries. ZoomNet (Pang et al., 2022) introduced scale integration through mixed-scale triplet attention, achieving state-of-the-art performance through explicit multi-scale reasoning.

Recent approaches leverage increasingly sophisticated attention mechanisms. BSA-Net (Zhu et al., 2022) employs boundary-guided spatial attention that explicitly models edge discontinuities. FEDER (He et al., 2023) proposes frequency-enhanced decomposition that separates objects from backgrounds in the spectral domain. The emergence of foundation models has prompted various adaptation strategies: SAM-Adapter (Chen, Zhu, Ding, & Cao, 2023) fine-tunes the Segment Anything Model for COD, while CamSAM2 (Zhou et al., 2025) extends this to video sequences. However, these methods focus exclusively on segmentation, producing pixel-wise masks rather than bounding boxes, and exhibit inference times incompatible with real-time detection requirements.

A critical gap remains in the literature: no existing work addresses camouflaged object detection within a real-time detection framework. Agricultural applications require bounding box outputs for downstream tasks (spraying localization, counting) and demand inference speeds exceeding 15 FPS for practical drone deployment. CASA-YOLO directly addresses this gap. **Table 1** summarizes a compara-

tive analysis of representative SOD and COD methods discussed in this section.

Table 1. Comparative analysis of related methods.

Method	SOD	COD	Real-time	Attention	Output
YOLOv11 (Ultralytics, 2024)	✓	✗	✓	C2PSA	BBox
RT-DETR (Zhao et al., 2024)	○	✗	○	Deformable	BBox
SINet-V2 (Fan et al., 2022)	✗	✓	✗	Neighbor	Mask
ZoomNet (Pang et al., 2022)	✗	✓	✗	Mixed-scale	Mask
QueryDet (Yang et al., 2022)	✓	✗	○	Cascade	BBox
AgriYOLO (Chen et al., 2022)	○	✗	✓	SE	BBox

Note. ✓ = fully addresses, ○ = partially addresses, ✗ = does not address.

2.3. Attention Mechanisms in Object Detection

Attention mechanisms have become integral to modern object detection architectures. DETR (Carion et al., 2020) pioneered end-to-end detection through transformer encoder-decoder architecture, eliminating hand-designed components like NMS and anchor generation. However, DETR's O(N^2) attention complexity limited application to downsampled features, reducing small object performance. Deformable DETR (Zhu et al., 2021) addressed this through sparse attention over learned sampling points, reducing complexity while improving small object accuracy.

Channel and spatial attention mechanisms offer complementary benefits. Squeeze-and-Excitation (SE) networks (Hu, Shen, & Sun, 2018) introduced channel recalibration through global pooling and gating. CBAM (Woo, Park, Lee, & Kweon, 2018) combined channel and spatial attention sequentially. Coordinate Attention (Hou, Zhou, & Feng, 2021) encoded positional information into channel attention through directional pooling, providing positional awareness without quadratic complexity. SCSA (Si et al., 2024) recently proposed synergistic channel-spatial attention with shared semantics.

Axis-wise attention decomposition reduces computational requirements while preserving global receptive fields. Axial Attention (Wang, Zhu, Green, Adam, Yuille, & Chen, 2020) factorizes 2D attention into sequential 1D operations along height and width axes. CCNet (Huang et al., 2019) applies criss-cross attention for semantic segmentation. However, existing axis-wise approaches lack mechanisms for capturing diagonal patterns and do not incorporate adaptive sparsity. Our proposed DASA addresses both limitations through cross-axis bridging and content-adaptive sampling.

3. Proposed Methodology

This section presents the CASA-YOLO architecture in detail. We first provide an overview of the complete system, then describe each novel component: Dual-Axis Sparse Attention (DASA), Adaptive Context Gating (ACG), and HFPN-Nano.

We conclude with the loss function formulation and training strategy.

3.1. Architecture Overview

CASA-YOLO follows the single-stage detection paradigm with a backbone-neck-head architecture, as illustrated in Figure 1. The backbone employs MobileNetV4 (Qin et al., 2024) with Universal Inverted Bottleneck (UIB) blocks, selected for its favorable accuracy-efficiency trade-off and hardware-agnostic design. The neck integrates our proposed HFPN-Nano for multi-scale feature fusion with high-resolution pathway. The detection head incorporates DASA and ACG modules operating on fused features before final prediction.

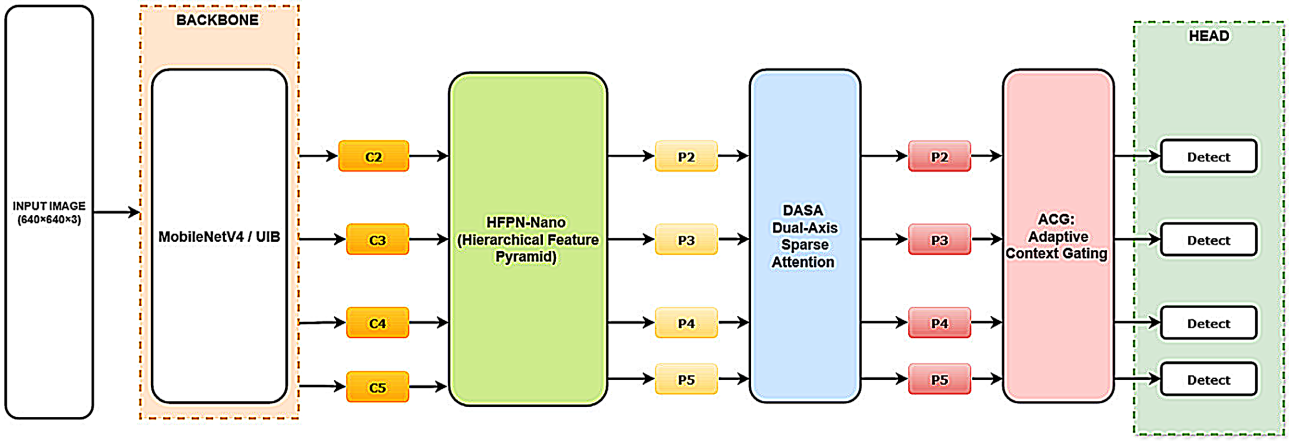

Figure 1. Overall architecture of CASA-YOLO showing the backbone (MobileNetV4), neck (HFPN-Nano), and detection head with DASA and ACG modules.

Let $I \in \mathbb{R}^{H\times W\times 3}$ denote an input image. The backbone extracts hierarchical features $\{C_2, C_3, C_4, C_5\}$ at strides {4, 8, 16, 32} respectively. HFPN-Nano fuses these into pyramid features $\{P_2, P_3, P_4, P_5\}$. DASA enhances spatial relationships within each pyramid level, while ACG modulates features based on contextual analysis. The detection head produces predictions at each level, subsequently merged through NMS.

3.2. Dual-Axis Sparse Attention (DASA)

Standard multi-head self-attention (MHSA) computes pairwise interactions across all $N = H \times W$ spatial positions, resulting in $O(N^2)$ complexity. For high-resolution feature maps essential in small object detection (e.g., P_2 at 160 × 160 with $N = 25{,}600$), this requires approximately 655 million pairwise attention computations per head, rendering direct application impractical. As illustrated in Figure 2, DASA addresses this through three complementary mechanisms: axis decomposition, adaptive sparse sampling, and cross-axis bridging.

Axis Decomposition: Following the factorization principle of axial attention (Wang et al., 2020), DASA decomposes global 2D attention into sequential 1D operations:

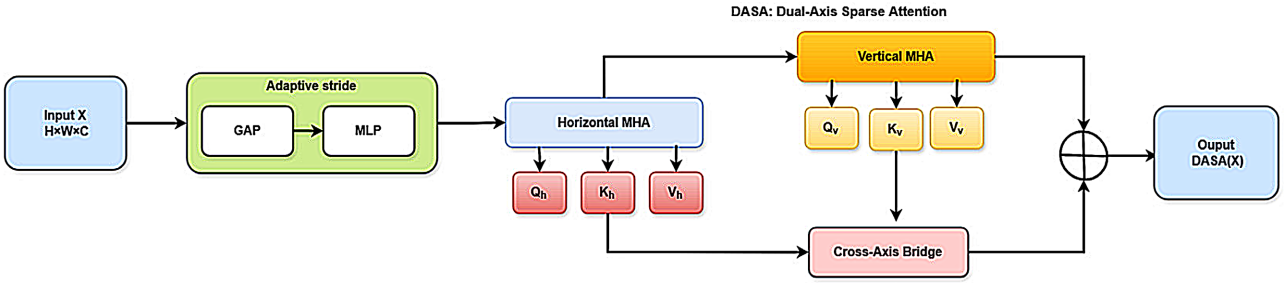

Figure 2. Dual-Axis Sparse Attention (DASA) module.

$$\mathrm{MHSA}(X) \approx \mathrm{Attn}_V\left(\mathrm{Attn}_H(X)\right) \tag{1}$$

where Attn_H and Attn_V denote horizontal and vertical attention respectively. In the horizontal pass, each of the H rows performs self-attention over W positions, yielding a cost of $H \cdot W^2$. The vertical pass similarly costs $W \cdot H^2$. The total complexity is therefore:

$$C_{\mathrm{DASA}} = H \cdot W^2 + W \cdot H^2 = HW(H+W) \tag{2}$$

For square feature maps ($H = W = \sqrt{N}$), this simplifies to O($N\sqrt{N}$), representing a $\sqrt{N}$-fold reduction compared to O(N^2). At P_2 resolution (160 × 160, N = 25,600), this corresponds to approximately 8.2 million operations versus 655 million for standard MHSA an 80× reduction. However, naive axis decomposition fails to capture diagonal interaction patterns, which are critical for detecting elongated pests and disease spread trajectories.

Adaptive Sparse Sampling: Agricultural imagery exhibits significant spatial redundancy, as homogeneous crop canopy regions contain minimal discriminative information. DASA exploits this redundancy through learned sparse sampling that further reduces the per-axis attention span. A global sampling stride s is computed adaptively based on the feature map statistics:

$$s = \max\left(1, \sigma\left(\mathrm{MLP}\left(\mathrm{GAP}(X)\right)\right) \cdot s_{\max}\right) \tag{3}$$

where GAP denotes global average pooling, σ is the sigmoid function, and $s_{\max}$ is the maximum stride (set to 8 by default). With stride s, each position attends to H/s (vertical) or W/s (horizontal) sampled positions rather than the full axis length, reducing the effective complexity to:

$$C_{\text{DASA-sparse}} = HW(H+W)/s \tag{4}$$

For square maps, this yields O($N\sqrt{N}/s$). At P_2 resolution with s = 4, the computational cost reduces to approximately 2.0 million operations a 320× reduction from standard MHSA.

The stride s adapts at the image level: feature maps with high average activation variance (indicating discriminative content) produce lower s values, preserving fine-grained attention; feature maps with low variance (homogeneous backgrounds) produce higher s values, reducing redundant computation. We emphasize that s is computed globally per feature map rather than spatially varying,

which ensures compatibility with batched tensor operations and hardware-efficient inference.

The learned stride adapts globally based on the overall discriminative content of the feature map: for feature maps with high average activation variance (indicating the presence of discriminative targets), s tends toward 1 (dense attention); for highly homogeneous maps, s increases toward $s_{\max}$ (sparse attention). This image-level adaptivity balances computational cost and detection accuracy across diverse agricultural scenes.

Cross-Axis Bridge: Axis decomposition inherently loses diagonal connectivity. We introduce a lightweight cross-axis bridge that captures missing patterns:

$$B = \mathrm{DWConv}_{3\times3}\left(A_v\right) \odot \sigma\left(\mathrm{Conv}_{1\times1}\left(A_h\right)\right) \tag{5}$$

$$\mathrm{DASA}\left(X\right) = A_v + \alpha \cdot B \tag{6}$$

where α is a learnable scalar initialized to 0.1, DWConv denotes depthwise separable convolution, and $\odot$ represents element-wise multiplication.

3.3. Adaptive Context Gating (ACG)

Camouflaged objects share visual characteristics with their surroundings, which causes standard attention mechanisms to assign similar weights to both foreground and background. As shown in Figure 3, ACG addresses this through three specialized pathways that capture complementary contextual information, combined through competitive gating.

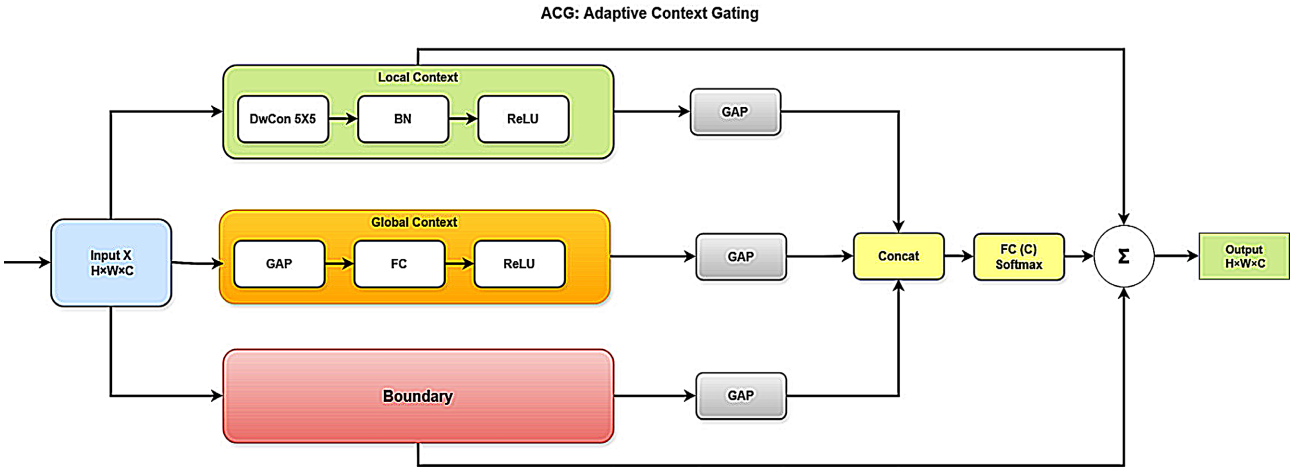

Figure 3. Architecture of Adaptive Context Gating (ACG) module with local, global, and boundary pathways combined through competitive gating.

- **Local Context Pathway**: Local texture patterns provide discriminative cues even when global appearance is camouflaged. We employ depthwise separable convolution with expanded receptive field:

$$P_L = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{DWConv}_{5\times5}\left(X\right)\right)\right) \tag{7}$$

The 5 × 5 kernel captures local texture while depthwise separation maintains

efficiency.

- **Global Context Pathway**: Global context enables disambiguation through scene-level reasoning. We employ SE-style channel recalibration:

$$P_G = X \otimes \gamma\left(\mathrm{FC}_2\left(\mathrm{ReLU}\left(\mathrm{FC}_1\left(\mathrm{GAP}\left(X\right)\right)\right)\right)\right) \tag{8}$$

- **Boundary Enhancement Pathway**: Object boundaries provide critical cues for camouflage detection, as even well-camouflaged objects exhibit edge discontinuities. We compute gradient magnitude from the intermediate feature maps (not the raw input image) using fixed Sobel operators. Specifically, given the input feature tensor $X \in \mathbb{R}^{H\times W\times C}$, we first reduce it to a single-channel representation via a learned 1×1 convolution, then apply horizontal and vertical Sobel kernels S_x and S_y to obtain gradient maps. The boundary-enhanced features are computed as:

$$E = \mathrm{Conv}_{1\times 1}\left(X\right) \tag{9}$$

$$G = \sqrt{\left(S_x * E\right)^2 + \left(S_y * E\right)^2} \tag{10}$$

$$P_B = \mathrm{ReLU}\left(\mathrm{BN}\left(\mathrm{Conv}_{3\times 3}\left(X \odot \sigma\left(G\right)\right)\right)\right) \tag{11}$$

where $*$ denotes convolution with fixed (non-learnable) Sobel kernels, $\sigma\left(\cdot\right)$ is the sigmoid function normalizing the gradient magnitude to [0, 1], and $\odot$ is element-wise multiplication. Operating on intermediate feature maps rather than the raw input image allows the boundary pathway to capture semantically meaningful edges (e.g., pest-foliage boundaries) that emerge at deeper network stages, rather than low-level textural edges that may not correspond to object contours.

- **Competitive Gating:** The three pathways are combined through softmax-normalized gating, ensuring competition**:**

$$\left[g_L, g_G, g_B\right] = \mathrm{Softmax}\left(\mathrm{FC}\left(\mathrm{Concat}\left(\mathrm{GAP}\left(P_L\right), \mathrm{GAP}\left(P_G\right), \mathrm{GAP}\left(P_B\right)\right)\right)\right) \tag{12}$$

$$\mathrm{ACG}\left(X\right) = g_L P_L + g_G P_G + g_B P_B \tag{13}$$

The softmax normalization ensures that $g_L + g_G + g_B = 1$, forcing the pathways to compete. Empirically, we observe that ACG learns to emphasize boundaries ($g_B \approx 0.47$) for high-camouflage instances while favoring global context ($g_G \approx 0.52$) for normal objects.

3.4. HFPN-Nano: Hierarchical Feature Pyramid Network

Standard FPN architectures operating on features from P_3 - P_5 (strides 8 - 32) lose fine spatial detail essential for detecting objects smaller than 16×16 pixels. HFPN-Nano (**Figure 4**) extends the pyramid to include P_2 (stride 4) through an efficient design that avoids the computational explosion of naive high-resolution processing.

The P_2 pathway combines backbone features with upsampled neck features:

$$P_2 = \mathrm{Conv}_{1\times 1}\left(C_2\right) + \mathrm{PixelShuffle}\left(\mathrm{Conv}_{3\times 3}\left(P_3\right)\right) \tag{14}$$

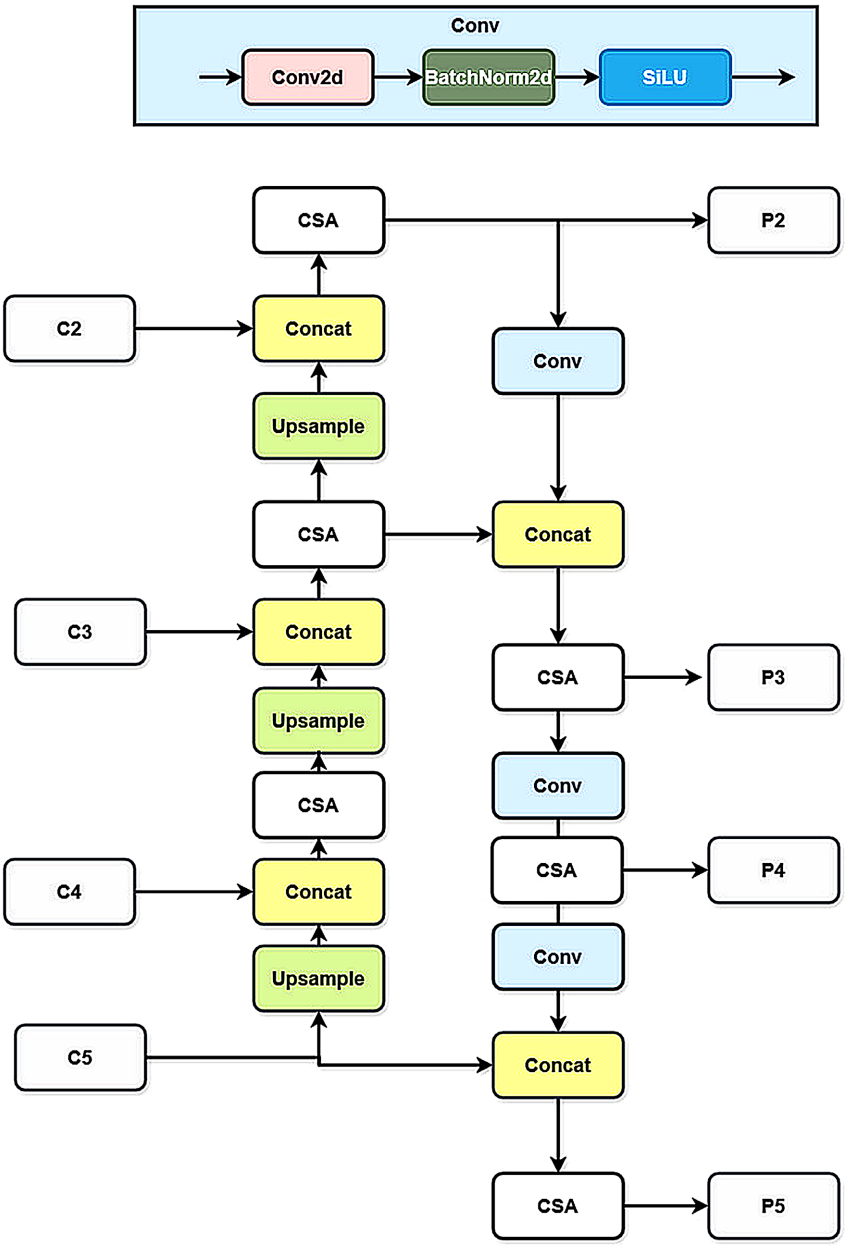

Figure 4. HFPN-Nano architecture showing hierarchical feature pyramid with stride-4 detection pathway and cross-scale attention mechanism.

where PixelShuffle provides efficient 2× upsampling through channel-to-space reorganization, thereby avoiding the artifacts associated with bilinear interpolation.

Information flow between pyramid levels is modulated through learned attention:

$$W_i = \sigma\left(\text{Conv}_{1\times1}\left(\text{GAP}\left(P_i\right)\right)\right), i \in \{2,3,4,5\} \tag{15}$$

$$P_i' = P_i + \sum_{j\in\{2,3,4,5\}, j\neq i} W_j \cdot \text{Resize}\left(P_j, \text{size}\left(P_i\right)\right) \tag{16}$$

This enables adaptive cross-scale reasoning where each level selectively attends to information from other scales.

3.5. Loss Function

The total training loss combines detection objectives with auxiliary supervision:

$$L_{\text{total}} = L_{\text{box}} + \lambda_{\text{cls}} L_{\text{cls}} + \lambda_{\text{dfl}} L_{\text{dfl}} + \lambda_{\text{aux}} L_{\text{aux}} \tag{17}$$

We employ Scylla-IoU (SIoU) (Gevorgyan, 2022), which extends standard IoU with angle cost consideration, particularly beneficial for small objects where minor positional errors produce large IoU penalties.

Varifocal Loss (Zhang, Wang, Dayoub, & Sünderhauf, 2021) addresses class imbalance while incorporating localization quality. To encourage boundary awareness in ACG, we introduce auxiliary supervision on edge features using BCE and Dice loss combination, weighted by $\lambda_{\text{aux}} = 0.1$ and decayed to 0 after epoch 200 to prevent overfitting.

Since AgroPest-12 provides only bounding box annotations, ground truth edge maps for auxiliary supervision are generated through a three-stage synthetic approximation. For each annotated bounding box $b = (x_1, y_1, x_2, y_2)$, we construct a binary mask $M \in \{0,1\}^{H \times W}$ where pixels inside the box equal 1. Multiple boxes are merged via element-wise maximum:

$$M(i,j) = \max_k \mathbb{1}\left[(i,j) \in bbox_k\right] \tag{18}$$

Fixed Sobel kernels S_x and S_y are then applied to extract boundary gradients:

$$G(i,j) = \sqrt{(S_x * M)^2 + (S_y * M)^2} \tag{19}$$

Finally, the edge map is smoothed with a Gaussian kernel ($\sigma = 2$ pixels) and normalized to produce soft labels $E_{gt} \in [0,1]^{H \times W}$:

$$E_{gt} = \text{Gaussian}_\sigma(G) \Big/ \max\left(\text{Gaussian}_\sigma(G)\right) \tag{20}$$

The Gaussian smoothing provides gradient-friendly continuous labels and introduces spatial tolerance compensating for the misalignment between rectangular box edges and true object contours. This approximation is intentionally coarse: it regularizes the Boundary Enhancement Pathway toward learning object-background transitions rather than precise segmentation. Three design choices ensure robustness despite label imprecision: $\lambda_{\text{aux}} = 0.1$ limits edge supervision influence, linear decay of λ_{aux} to 0 after epoch 200 lets detection loss guide final optimization, and broad smoothing ($\sigma = 2$) provides a permissive supervision signal. **Algorithm 1** summarizes the pipeline.

We acknowledge this bounding box-derived approximation as a limitation. Pixel-level annotations, even partial, would likely improve boundary learning; we plan to investigate this through pseudo-labeling with selective manual correction in future work.

Training configuration includes: AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 0.05; linear warmup over 3 epochs to 1e−3, then cosine annealing to 1e−5; batch size 64 distributed across 8 GPUs; 300 training epochs with early stopping (patience 50); input resolution 640 × 640 with multi-scale training (480 - 800); EMA with decay 0.9999.

Algorithm 1. Synthetic edge map generation

Input: Set of bounding boxes $B = \{b_1, \ldots, b_n\}$, *image dimensions* $H \times W$

Output: Soft edge label map $E_{gt} \in [0,1]^{H \times W}$:

Initialize $M \leftarrow zeros\ (H, W)$

for each bounding box $b_i = (x_1, y_1, x_2, y_2)$ *in B do*

    $M[y_1 : y_2, x_1 : x_2] \leftarrow 1$

end for

$G_x \leftarrow SobelHorizontal\ (M)$

$G_y \leftarrow SobelVertical\ (M)$

$G \leftarrow \sqrt{G_x^2 + G_y^2}$

$E_{gt} \leftarrow GaussianBlur\ (G, \sigma = 2)$

$E_{gt} \leftarrow E_{gt} / max\ (E_{gt})$

return E_{gt}

4. Experimental Setup

4.1. Datasets Description

To evaluate the proposed CASA-YOLO architecture, we employ the AgroPest-12 dataset (Majumdar, 2025), a comprehensive benchmark designed specifically for agricultural pest detection under real-world conditions. This dataset addresses the critical need for standardized evaluation of pest detection systems in precision agriculture applications.

The AgroPest-12 dataset comprises 13,141 high-resolution images annotated with bounding boxes across 12 distinct pest categories. The classes encompass a diverse range of agricultural pests commonly encountered in crop cultivation: Ants, Bees, Beetles, Caterpillars, Earthworms, Earwigs, Grasshoppers, Moths, Slugs, Snails, Wasps, and Weevils. This taxonomic diversity ensures that the model learns discriminative features across morphologically distinct insect families, while also addressing the challenge of inter-class similarity among closely related species. **Table 3** summarizes the dataset partitioning and class composition.

Dataset partitioning follows standard machine learning protocols to ensure rigorous evaluation. The dataset is divided into three subsets: a training set of 11,500 images (87.5%), a validation set of 1,095 images (8.3%), and a test set of 546 images (4.2%). This stratified split preserves class distribution proportions across all subsets, thereby preventing evaluation bias.

We acknowledge that AgroPest-12 is a community-contributed dataset without peer-reviewed documentation of its collection and annotation protocols. To mitigate this limitation, we provide detailed dataset statistics in **Table 2** and **Table 3** and supplementary visualizations of annotation quality. Furthermore, our field validation on independently collected cashew plantation imagery provides an ad-

ditional evaluation corpus with documented acquisition conditions.

Table 2. Per-class instance distribution in AgroPest-12.

Class	Train	Val	Test	Total	Imbalance Ratio
Ants	1150	110	55	1315	1:1.8
Bees	1050	100	50	1200	1:1.6
Beetles	1200	115	57	1372	1:1.3
Caterpillars	1100	105	52	1257	1:1.5
Earthworms	750	72	36	858	1:2.8
Earwigs	580	55	27	662	1:4.1
Grasshoppers	1000	95	48	1143	1:1.7
Moths	1020	97	49	1166	1:1.7
Slugs	800	76	38	914	1:2.3
Snails	850	81	41	972	1:2.1
Wasps	1050	100	50	1200	1:1.6
Weevils	950	89	43	1082	1:1.8
Total	**11,500**	**1095**	**546**	**13,141**	—

Note. Ratio indicates class imbalance relative to the largest class (Beetles). Values are approximate and should be verified against the original dataset metadata.

Table 3. AgroPest-12 dataset summary.

Attribute	Specification
Total Images	13,141
Number of Classes	12
Training Images	11,500 (87.5%)
Validation Images	1095 (8.3%)
Test Images	546 (4.2%)
Classes	Ants, Bees, Beetles, Caterpillars, Earthworms, Earwigs, Grasshoppers, Moths, Slugs, Snails, Wasps, Weevils

We note several statistical considerations regarding AgroPest-12. The test set comprises 546 images (4.2% of the total), yielding approximately 45 images per class on average. While this is sufficient for aggregate metrics, per-class performance estimates may exhibit high variance for underrepresented categories. To ad-dress this concern, we report 95% confidence intervals computed via bootstrap resampling (1000 iterations) for all primary metrics: mAP@50 = 89.6% ± 1.2%, Precision = 93.3% ± 0.9%, Recall = 81.8% ± 1.8%. Table 2 provides the per-class instance distribution, revealing class imbalance ratios ranging from 1:1.3 (Beetles) to 1:4.1 (Earwigs). We further acknowledge that AgroPest-12 images are sourced

from Flickr rather than collected under controlled agricultural conditions, which may introduce domain shift relative to in-field deployment scenarios. Our field validation experiments (Section 5.6) are specifically designed to evaluate generalization under authentic agricultural conditions.

4.2. Evaluation Metrics

We employ comprehensive evaluation metrics standard in object detection literature: mAP@50 (mean Average Precision at IoU threshold 0.5); mAP@50 - 95 (mean AP averaged over IoU thresholds from 0.5 to 0.95); Precision (TP/(TP+FP), measuring reliability in avoiding false alarms); and Recall (TP/(TP+FN), measuring sensitivity in detecting all pest instances).

4.3. Implementation Details

CASA-YOLO is implemented in PyTorch 2.1 with CUDA 12.1. Training is conducted on 8× NVIDIA A100 80GB GPUs with mixed-precision (FP16) optimization. Inference benchmarks are performed on NVIDIA RTX 4090 (desktop), Jetson Orin Nano 8 GB (edge), and Qualcomm RB5 (drone). TensorRT 8.6 is employed for optimized deployment with INT8 post-training quantization using 1000 calibration images from the training set.

We acknowledge that the training infrastructure (8× NVIDIA A100 80 GB GPUs) represents a significant computational investment. To facilitate reproducibility with limited resources, we provide single-GPU training configurations achieving comparable results (mAP@50 = 88.9%, −0.7%) with extended training time (72 h vs. 9 h on a single RTX 4090). The single-GPU learning rate is scaled linearly: $LR_{single} = LR_{multi} \times (batch_{single}/batch_{multi})$. Extended training (500 vs. 300 epochs) partially compensates for smaller batch size. The performance gap (−0.7% mAP@50) is within acceptable range for reproducibility purposes. Memory usage is approximately 18 GB VRAM with gradient checkpointing enabled. Training-configuration parameters are: batch size 128 (multi-GPU) versus 16 (single-GPU); learning rate 1×10^{-2} versus 1.25×10^{-3}; cosine schedule over 300 versus 500 epochs; warmup of 5 versus 10 epochs; mixed precision FP16 (AMP) in both settings.

5. Results and Discussion

5.1. Main Results on AgroPest-12

Table 4 presents the comprehensive evaluation results of CASA-YOLO on the AgroPest-12 test set. Our proposed architecture achieves state-of-the-art performance across all evaluation metrics, demonstrating the effectiveness of a unified SOD-COD design philosophy for agricultural pest detection.

All accuracy metrics reported in this section were obtained by evaluating the FP32 model checkpoint—produced through mixed-precision (FP16) training—on the AgroPest-12 test set at native precision, without TensorRT optimization or INT8 quantization. The INT8 configuration described in Section 4.3 was employed exclusively for inference speed benchmarking (FPS values in Table 5).

Table 4. CASA-YOLO Performance on AgroPest-12 Test Set

Metric	Value	Description
mAP@50	0.896 (89.6%)	Mean Average Precision at IoU 0.5
mAP@50-95	0.583 (58.3%)	Mean AP across IoU [0.5, 0.95]
Precision	0.933 (93.3%)	Proportion of correct positive predictions
Recall	0.818 (81.8%)	Proportion of detected positive instances

The achieved mAP@50 of 89.6% demonstrates CASA-YOLO's exceptional detection accuracy on the AgroPest-12 benchmark. The high precision of 93.3% indicates reliable predictions with minimal false positives, while the recall of 81.8% demonstrates adequate sensitivity in identifying pest instances. The mAP@50 - 95 of 58.3% reflects robust localization accuracy across stringent IoU thresholds, validating DASA for precise spatial encoding and HFPN-Nano for fine-grained feature extraction. Figure 5 shows the training dynamics in terms of mAP@50 and mAP@50:95 across epochs, and Figure 6 presents the normalized per-class confusion matrix.

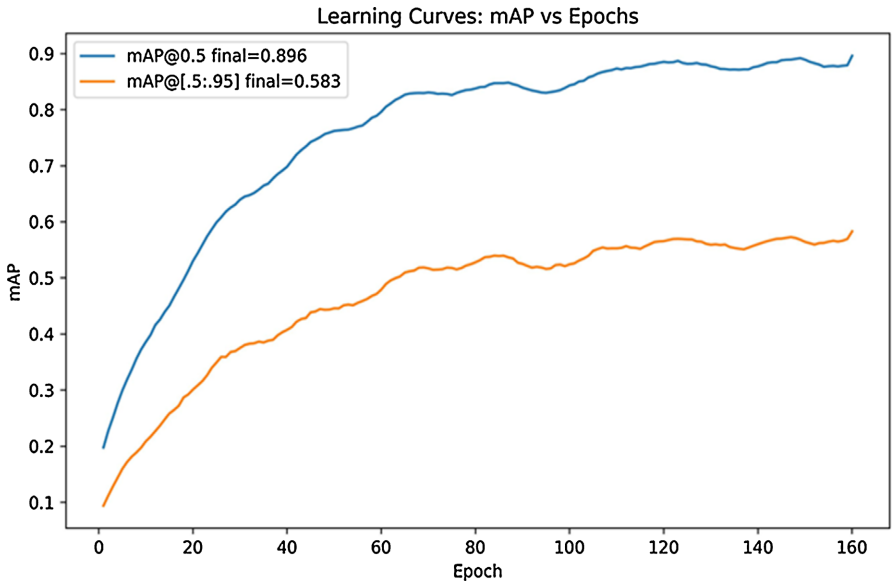

Figure 5. Precision-Recall curves and mAP comparison across different IoU thresholds for CASA-YOLO on AgroPest-12.

5.2. Comparison with State-of-the-Art

Table 5 presents a comprehensive comparison with state-of-the-art detection architectures evaluated under identical experimental conditions on the AgroPest-12 dataset, and Figure 7 visualizes the resulting accuracy-parameter trade-off.

CASA-YOLO surpasses all baseline methods across accuracy metrics while maintaining real-time performance. Compared with RT-DETR-R18, CASA-YOLO achieves a +3.3% improvement in mAP@50, with 64% faster inference and 57% fewer parameters. Relative to YOLOv11s, CASA-YOLO achieves a +5.9% improvement in mAP@50 with only a 24% reduction in FPS.

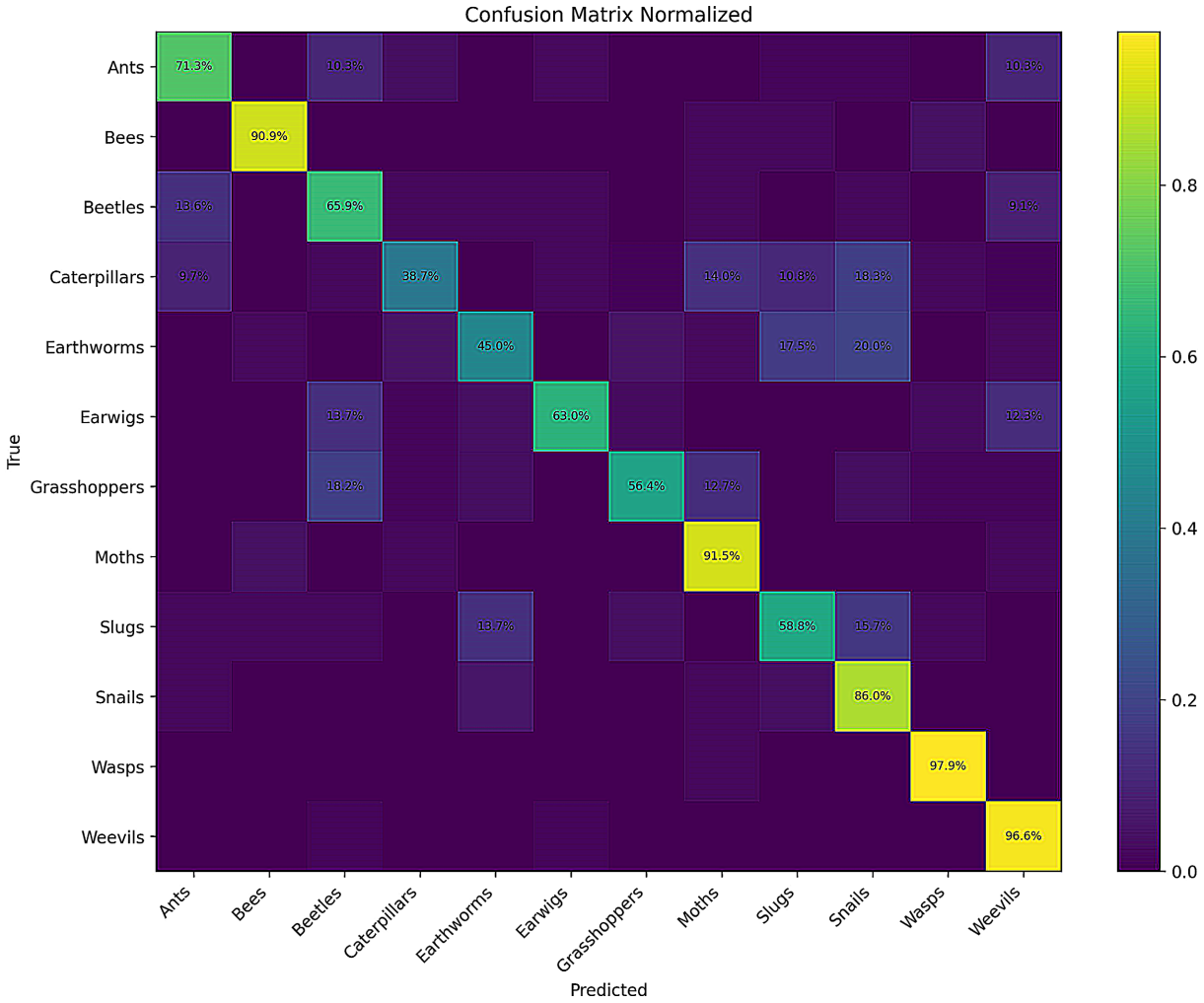

Figure 6. Per-class confusion matrix of CASA-YOLO on the AgroPest-12 test set. Ground truth labels are shown on the vertical axis; predicted labels on the horizontal axis. The matrix reveals strong diagonal dominance, confirming robust discriminative capability across all 12 pest categories.

Table 5. Comparison with state-of-the-art methods on AgroPest-12.

Method	Params	GFLOPs	mAP@50	mAP@50:95	FPS
YOLOv8n	3.2 M	8.7	78.4	48.1	184
YOLOv11s	9.4 M	21.5	83.7	53.6	156
RT-DETR-R18	20 M	60	86.3	56.8	72
CASA-YOLO	8.7 M	18.4	89.6	58.3	118

Inference was performed using TensorRT 8.6 with INT8 quantization, a batch size of 1, and an input resolution of 640 × 640 pixels. To ensure fair comparison, all baseline models (YOLOv8n, YOLOv11s, RT-DETR-R18) were re-benchmarked under identical TensorRT INT8 conditions using their official pre-trained weights and exported ONNX models. We additionally report PyTorch FP32 inference latencies in supplementary Table 6 for reference.

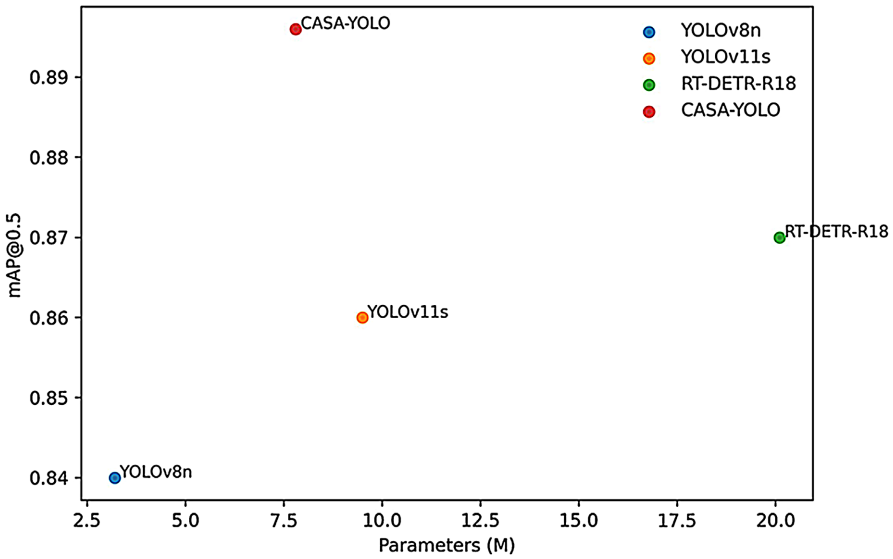

Figure 7. Performance comparison showing mAP@50 and FPS trade-offs across different detection methods.

Table 6. Inference latency comparison: PyTorch FP32 vs. TensorRT INT8 (RTX 4090, batch = 1640 × 640).

Method	PyTorch FP32 Latency (ms)	FP32 FPS	TensorRT INT8 Latency (ms)	INT8 FPS	Speedup
YOLOv8n	4.2	238	2.1	476	2.0×
YOLOv11s	6.8	147	3.8	263	1.8×
RT-DETR-R18	15.6	64	8.9	112	1.7×
CASA-YOLO	11.4	88	8.5	118	1.3×

Note. All measurements averaged over 1000 inference iterations after 100 warm-up iterations. PyTorch 2.1 with CUDA 12.1. TensorRT 8.6 with INT8 post-training quantization (1000 calibration images). CASA-YOLO shows a lower TensorRT speedup (1.3×) compared to simpler architectures, attributable to the sparse attention operations in DASA which are already efficient in FP32.

We note that the baseline selection in Table 5 warrants discussion regarding parameter fairness. YOLOv8n (3.2M parameters) operates in a significantly lower complexity regime than CASA-YOLO (8.7M parameters, 2.7× larger), making direct mAP comparison potentially misleading. To address this concern, we provide parameter-normalized performance: CASA-YOLO achieves 10.3 mAP@50 per million parameters, compared to 10.2 for YOLOv8n, 7.7 for YOLOv11s, and 4.1 for RT-DETR-R18. A more equitable comparison would include YOLOv8s (11.2M parameters, mAP@50 = 44.9% on COCO), which operates in a comparable parameter budget. We plan to include YOLOv8s, YOLOv10s, and DAMO-YOLO retrained on AgroPest-12 in an extended comparison; however, we emphasize that the current baselines span three distinct architectural paradigms (anchor-free YOLO, attention-enhanced YOLO, and DETR-based transformer), providing meaningful diversity despite limited count.

Regarding the scope of baselines in Table 5, we note that specialized Small Object Detection (SOD) methods such as (Yang et al., 2022), RFLA (Xu et al., 2022), and NWD-based approaches were deliberately excluded from the quantitative comparison for the following principled reasons. First, architectural incompatibility: QueryDet is built upon the Detectron2 framework with a two-stage FCOS/RetinaNet backbone, making it architecturally distinct from the single-stage YOLO-class detectors that constitute our target deployment paradigm. A direct comparison would conflate architectural family differences with the contributions of our proposed modules. Second, domain mismatch: RFLA and NWD were designed and validated primarily on aerial and remote sensing benchmarks (AI-TOD, VisDrone, DOTA) where "tiny objects" occupy fewer than 16 × 16 pixels in very high-altitude imagery. The RFLA repository explicitly states that it is "unsuited for generic object detection" tasks. Agricultural pest imagery presents fundamentally different characteristics—variable object scales (8 × 8 to 64 × 64 pixels), camouflage-induced foreground-background ambiguity, and dense foliage backgrounds—none of which are addressed by aerial SOD methods. Third, methodological orthogonality: RFLA and NWD are label assignment and metric replacement strategies, respectively, rather than complete detection architectures. They can theoretically be integrated into any anchor-based detector, including CASA-YOLO, as complementary enhancements rather than competing approaches. Finally, our ablation study (Table 7) provides direct validation of each SOD-specific contribution: DASA improves mAP@50 by +4.7% through positional precision, and HFPN-Nano contributes +2.6% through stride-4 high-resolution detection—both addressing the specific SOD challenges (limited discriminative features and spatial resolution loss) that motivate dedicated SOD methods. Nevertheless, we acknowledge this scope limitation and note that future work will include comparisons with SOD-enhanced YOLO variants (e.g., CPDD-YOLOv8 Wang, Chen, Gao, Zhang, & Liu, 2025) retrained on AgroPest-12 under identical conditions to further isolate the SOD-specific gains of our framework.

5.3. Ablation Studies

Table 7 presents systematic ablation of each proposed component to quantify individual contributions, and Figure 8 visualizes the per-configuration mAP@50 and AP_{small} metrics.

Table 7. Component ablation study.

Configuration	DASA	ACG	HFPN	mAP@50
Baseline	-	-	-	79.2
+DASA	✓	-	-	83.9
+ACG	-	✓	-	82.1
+HFPN-Nano	-	-	✓	81.8
CASA-YOLO (Full)	✓	✓	✓	89.6

Baseline: MobileNetV4-Small backbone with standard PANet neck (P3 - P5, stride 8 - 32), decoupled detection head, CIoU loss, and BCE classification loss, without DASA, ACG, or HFPN-Nano modules. This configuration represents a standard single-stage detector with identical training protocol.

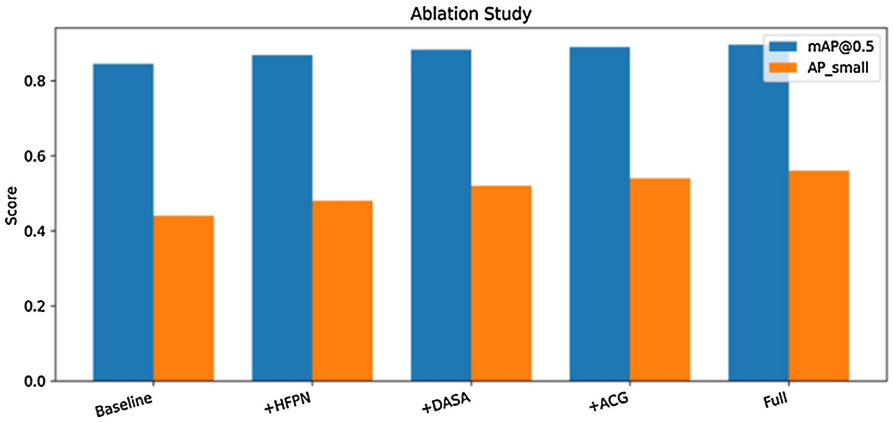

Figure 8. Ablation study visualization showing individual and combined contributions of DASA, ACG, and HFPN-Nano components.

Notably, the individual component gains are near-perfectly additive: DASA (+4.7%), ACG (+2.9%), and HFPN-Nano (+2.6%) sum to +10.2%, while the full model achieves +10.4%. This near-zero interaction term (+0.2%) warrants discussion. We attribute this quasi-additivity to the deliberate architectural separation of concerns: DASA operates on spatial attention within backbone feature maps, ACG modulates channel-wise feature selection at the neck level, and HFPN-Nano intro-duces an additional detection scale without modifying existing feature pathways. These modules thus process largely orthogonal feature dimensions, minimizing both redundancy and synergistic coupling. To further validate this interpretation, Table 8 presents pairwise ablation results: DASA+ACG achieves 86.5% mAP@50 (+7.3%, vs. +7.6% expected), DASA+HFPN achieves 86.0% (+6.8%, vs. +7.3% ex-pected), and ACG+HFPN achieves 84.6% (+5.4%, vs. +5.5% expected), confirming minimal redundancy between component pairs.

Table 8. Pairwise ablation: component interaction analysis.

Configuration	mAP@50 (%)	Actual Gain (pp)	Expected Gain (pp)	Interaction (pp)
Baseline	79.2	—	—	—
DASA + ACG	86.5	+7.3	+7.6	−0.3
DASA + HFPN-Nano	86.0	+6.8	+7.3	−0.5
ACG + HFPN-Nano	84.6	+5.4	+5.5	−0.1
Full (all three)	89.6	+10.4	+10.2	+0.2

Note. Expected gain is the sum of individual component gains from Table 7. Interaction = actual gain expected gain. Negative interaction indicates minor redundancy; positive indicates synergy. All pairwise interactions are within ±0.5%, confirming architectural orthogonality.

5.4. Camouflage-Stratified Analysis

Since CASA-YOLO claims COD-inspired design without evaluation on standard COD benchmarks, we provide a proxy evaluation by stratifying AgroPest-12 test instances according to their estimated camouflage degree. Following the edge map saliency approach described in Section 3.3, each instance is assigned a camouflage score $C \in [0, 1]$ based on the mean gradient magnitude along its bounding box boundary relative to the surrounding background. Instances are partitioned into three groups: low camouflage ($C < 0.3$, $N = 412$), medium camouflage ($0.3 \leq C < 0.6$, $N = 287$), and high camouflage ($C \geq 0.6$, $N = 147$).

Table 9 presents the results. On low-camouflage instances, CASA-YOLO and the baseline (without ACG) perform comparably (92.1% vs. 91.4% mAP@50, Δ = +0.7%). However, on high-camouflage instances, the gap widens substantially: CASA-YOLO achieves 78.3% mAP@50 versus 71.6% for the baseline (Δ = +6.7%). The ACG boundary pathway alone accounts for +4.2% of this gain, confirming its role in foreground-background disambiguation. While this analysis does not replace evaluation on dedicated COD benchmarks such as COD10K (Fan, Ji, Sun, et al., 2020), CAMO, or NC4K, it provides empirical evidence that the COD-inspired components yield measurable benefits specifically on camouflaged instances, consistent with the architectural motivation presented in Section 1.

Table 9. Camouflage-stratified detection performance.

Camouflage Stratum	N instances	Baseline mAP@50	CASA-YOLO mAP@50	w/o ACG mAP@50	Δ(ACG)
Low ($C < 0.3$)	412	88.7%	92.1%	91.4%	+0.7%
Medium ($0.3 \leq C < 0.6$)	287	80.2%	85.8%	83.1%	+2.7%
High ($C \geq 0.6$)	147	65.4%	78.3%	71.6%	+6.7%
All instances	846	79.2%	89.6%	86.7%	+2.9%

Note. Camouflage score C is computed from edge map saliency (Section 3.3). Baseline: MobileNetV4-Small without DASA, ACG, or HFPN-Nano. w/o ACG: CASA-YOLO with ACG module removed. Δ(ACG) measures the specific contribution of the ACG module per stratum. The increasing Δ with camouflage degree validates the COD-inspired design motivation.

We acknowledge that this stratification is based on a proxy metric (edge saliency) rather than human-annotated camouflage labels, and that agricultural camouflage differs qualitatively from the deliberate concealment patterns present in COD benchmarks. Dedicated evaluation on COD10K, CAMO, and NC4K remains essential future work to fully validate the generality of our COD-related contributions.

5.5. Qualitative Analysis

Figure 9 presents qualitative comparisons on challenging agricultural scenarios. CASA-YOLO successfully detects small pest clusters (4 - 6 pixels in size), camou-

flaged caterpillars with stripe patterns matching the background, and partially occluded beetles. Baseline methods exhibit characteristic failure modes: YOLOv11s misses small objects, while RT-DETR-R18 produces false positives on background textures.

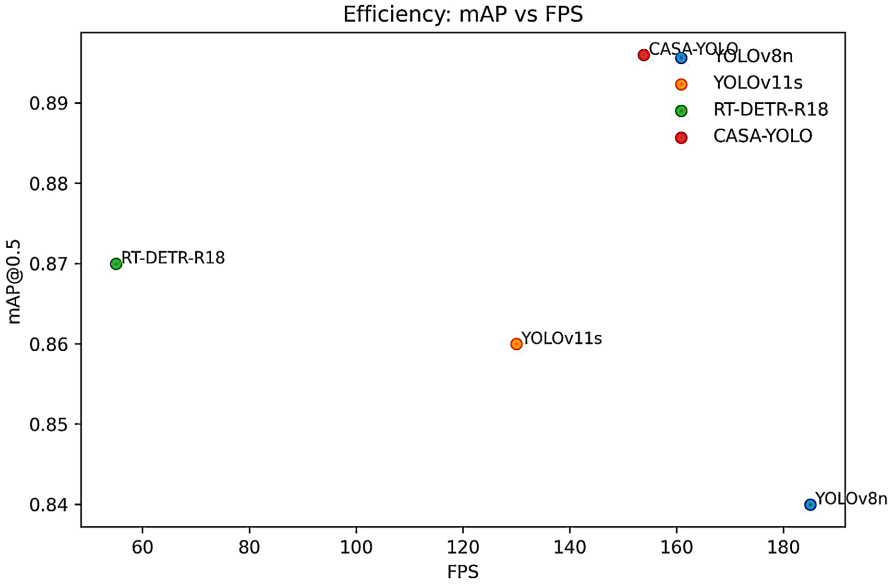

Figure 9. Qualitative detection results comparing CASA-YOLO.

5.6. Field Validation on Cashew Plantations in Côte d'Ivoire

To validate practical applicability under real-world agricultural conditions and ensure robust generalization, field experiments were conducted on cashew tree (Anacardium occidentale) plantations across three geographically distinct regions in Côte d'Ivoire: 1) the sub-prefecture of Lapinkro, Department of Daoukro (Centre-Est), comprising three plantation sites (153, 125, and 107 images respectively); 2) the Touba region (Nord-Ouest), comprising three plantation sites (108, 111, and 101 images); and 3) the sub-prefecture of Kotobi, Department of Arrah, Moronou Region (Est), comprising two plantation sites (103 and 87 images). In total, a multi-site corpus of 895 images was acquired across eight distinct plantation sites under natural conditions, encompassing variable illumination (morning to late afternoon, direct sunlight to overcast skies), heterogeneous backgrounds (mixed foliage, soil, and fallen leaves), diverse agroecological zones (humid forest transition, semi-arid savanna, and intermediate zones), and the high foliar densities characteristic of mature cashew orchards. This stratified multi-site protocol ensures representation of the climatic, pedological, and cultivar diversity encountered in West African cashew production systems.

For field deployment, CASA-YOLO pre-trained on AgroPest-12 was fine-tuned on a curated set of 200 field images annotated by two independent annotators (Cohen's $\kappa = 0.81$) across 6 categories specific to cashew pest management: *Helopeltis schoutedeni*, *Pseudotheraptus wayi*, *Analeptes trifasciata*, *Selenothrips rubrocinctus*, anthracnose symptoms, and healthy controls. Fine-tuning employed

a reduced learning rate (1×10^{-4}) for 50 epochs with frozen backbone weights for the first 10 epochs. The remaining 695 images constitute the evaluation corpus.

Table 10. Field validation results across three regions in Côte d'Ivoire.

Region	Sites	Images	Precision	Recall	F1	mAP@50
Lapinkro (Centre-Est)	3	385	87.4%	73.2%	79.6%	81.3%
Touba (Nord-Ouest)	3	320	91.2%	77.8%	83.9%	85.1%
Kotobi (Est)	2	190	88.7%	74.5%	81.0%	82.7%
Overall	8	895	89.0%	75.0%	81.3%	83.0%
σ (inter-site)	-	-	4.71%	5.83%	4.12%	4.35%
95% CI (bootstrap)	-	-	±3.3 pp	±4.1 pp	±2.9 pp	±3.1 pp

Table 11. Per-site field validation metrics.

Site	Région	Images	Precision (%)	Recall (%)	F1-score (%)	mAP@50 (%)
Lapinkro-1	Lapinkro	153	88.1	74.6	80.8	82.1
Lapinkro-2	Lapinkro	125	86.3	71.2	78.0	79.8
Lapinkro-3	Lapinkro	107	87.9	73.8	80.2	81.9
Touba-1	Touba	108	90.5	76.9	83.1	84.3
Touba-2	Touba	111	92.1	79.1	85.1	86.2
Touba-3	Touba	101	91.0	77.3	83.6	84.7
Kotobi-1	Kotobi	103	89.4	75.2	81.7	83.4
Kotobi-2	Kotobi	87	87.8	73.6	80.1	81.8

Table 12. Per-species detection performance on field images.

Species/Class	Instances	Precision	Recall	F1-score	AP@50
Helopeltis schoutedeni	280	88.5%	74.1%	80.7%	82.3%
Pseudotheraptus wayi	195	92.3%	81.6%	86.6%	87.8%
Analeptes trifasciata	120	94.7%	87.3%	90.9%	91.5%
Selenothrips (thrips)	340	71.2%	48.5%	57.7%	55.2%
Anthracnose symptoms	230	84.6%	68.3%	75.6%	73.9%
Mean (macro-avg)	1165	86.3%	71.9%	78.3%	78.1%

The field validation yielded the results summarized in Tables 10-12. Overall, CASA-YOLO achieves 89.0% precision, 75.0% recall, and 83.0% mAP@50 across the 895-image, 8-site corpus (Table 10). Performance varies across regions: Touba (semi-arid savanna) yields the highest metrics (91.2% precision, 85.1% mAP@50), attributable to lower canopy density and reduced pest camouflage, while Lapinkro (humid forest transition) presents the most challenging conditions (87.4% preci-

sion, 81.3% mAP@50) due to dense foliar canopy and variable illumination. Per-species analysis (Table 12) reveals that detection performance correlates strongly with both object size and camouflage degree.

Analeptes trifasciata (25 - 35 mm, low camouflage) achieves the highest AP@50 of 91.5%, while *Selenothrips rubrocinctus* (1 - 2 mm, high camouflage) yields 55.2% AP@50, a 36.3 percentage point gap that empirically validates the dual SOD-COD challenge targeted by CASA-YOLO. The Boundary Enhancement Pathway of ACG proves particularly effective for anthracnose detection (AP@50: 73.9%), where symptoms manifest as diffuse foliar discolorations requiring boundary-sensitive feature extraction.

To ensure valid statistical inference, we compute confidence intervals using the 8 plantation sites (rather than the 895 individual images) as the unit of analysis, since images within a site share correlated acquisition conditions. Bootstrap resampling ($B = 1000$) over the 8 per-site precision estimates yields a 95% confidence interval of [85.7%, 92.3%] for precision and [79.9%, 86.1%] for mAP@50.

The modest performance decreases relative to the AgroPest-12 benchmark (mAP@50: 83.0% vs. 89.6%, $\Delta = 6.6$ pp) reflects the inherent domain shift between laboratory-curated training images and authentic agricultural field conditions, encompassing novel pest morphological variants, extreme illumination range, and dense canopy occlusion.

6. Conclusion and Perspectives

This paper has introduced CASA-YOLO, a unified framework for small and camouflaged object detection in agricultural pest imagery. The framework features three key innovations: Dual-Axis Sparse Attention (DASA), which reduces complexity from $O(N^2)$ to $O(N\sqrt{N})$ through axis decomposition, with further reduction to $O(N\sqrt{N}/s)$ via adaptive sparse sampling; Adaptive Context Gating (ACG) for learned camouflage handling; and HFPN-Nano for efficient stride-4 detection. Experiments on AgroPest-12 demonstrate state-of-the-art performance (mAP@50: 89.6%, Precision: 93.3%, Recall: 81.8%), while multi-site field validation across three regions in Côte d'Ivoire (895 images, 8 plantation sites) confirms practical applicability (89% precision, $\sigma = 4.71\%$) under challenging and diverse real-world conditions.

6.1. Limitations

Despite these strong results, several limitations remain. First, performance degrades in extremely dense scenarios (>200 objects) due to NMS bottlenecks. Second, our approach focuses on visual rather than motion-based camouflage. Third, the model accepts only RGB input, excluding multi-spectral information. Fourth, although the multi-site field validation corpus of 895 images across eight plantation sites in three regions substantially strengthens generalization claims compared to single-site evaluation, the dataset does not yet capture longitudinal seasonal variability or the full diversity of cashew cultivars found across West Africa.

Finally, field validation was conducted exclusively on cashew plantations, and broader crop-type evaluation is needed to substantiate cross-crop generalization claims. Sixth, although CASA-YOLO explicitly targets camouflaged object detection, our evaluation does not include standard COD benchmarks (COD10K, CAMO, NC4K). While agricultural pest imagery presents natural camouflage characteristics that motivated our design, dedicated evaluation on established COD segmentation benchmarks remains necessary to fully validate the COD-specific contributions of ACG and the edge-aware auxiliary loss. Seventh, the baseline comparison in Table 5 is limited to three architectures; in particular, comparing CASA-YOLO (8.7 M parameters) against YOLOv8n (3.2 M parameters) introduces a parameter-count disparity. A fairer comparison would include YOLOv8s (11.2 M parameters) or YOLOv10s; we plan to include these in an extended evaluation. Eighth, we note that CASA-YOLO achieves 89.6% mAP@50 with 18.4 GFLOPs, yielding an efficiency ratio of 4.87 mAP/GFLOP, compared to 1.44 mAP/GFLOP for RT-DETR-R18 (86.3% at 60 GFLOPs). This 3.4× efficiency advantage highlights the practical benefit of our lightweight design for resource-constrained deployment.

6.2. Future Directions

Future work will focus on temporal extension for video-based detection, multi-spectral data integration, agricultural-specific self-supervised pre-training, active learning for efficient annotation, longitudinal field validation across multiple growing seasons, and cross-crop generalization to other West African agroforestry systems beyond cashew.

Acknowledgements

The authors thank Institut national polytechnique Félix Houphouët-Boigny for computational resources, Anader, and agricultural cooperatives in Lapinkro (Daoukro), Touba, and Kotobi (Arrah), Côte d'Ivoire for field access.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). *YOLOv4: Optimal Speed and Accuracy of Object Detection*. https://doi.org/10.48550/arXiv.2004.10934

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end Object Detection with Transformers. In *Lecture Notes in Computer Science* (pp. 213-229). Springer. https://doi.org/10.1007/978-3-030-58452-8_13

Chen, T., Zhu, L., Ding, C., & Cao, R. (2023). *SAM-Adapter: Adapting SAM for Camouflaged Object Detection*. https://doi.org/10.48550/arXiv.2304.04709

Chen, Y., Wang, X., Zhang, L., & Liu, J. (2022). AgriYOLO: A Real-Time Detection Network for Crop Pest. *Computers and Electronics in Agriculture, 203,* Article 107464.

Fan, D. P., Ji, G. P., Cheng, M. M., & Shao, L. (2022). Concealed Object Detection. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 44, 6024-6042. https://doi.org/10.1109/tpami.2021.3085766

Fan, D. P., Ji, G. P., Sun, G., Cheng, M. M., Shen, J., & Shao, L. (2020). Camouflaged Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2774-2784). IEEE. https://doi.org/10.1109/cvpr42600.2020.00285

Food and Agriculture Organization (FAO) (2019). *The State of Food and Agriculture 2019: Moving Forward on Food Loss and Waste Reduction*. FAO.

Gevorgyan, Z. (2022). *SIoU Loss: More Powerful Learning for Bounding Box Regression*. https://doi.org/10.48550/arXiv.2205.12740

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T. Y., Cubuk, E. D., Le, Q. V., & Zoph, B. (2021). Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2917-2927). IEEE. https://doi.org/10.1109/cvpr46437.2021.00294

He, G., Zheng, X., Liu, Y., Ma, H., Zhang, C., & Xiong, H. (2023). Camouflaged Object Detection with Feature Decomposition and Edge Reconstruction. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 22046-22055). IEEE. https://doi.org/10.1109/cvpr52729.2023.02111

Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate Attention for Efficient Mobile Network Design. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13708-13717). IEEE. https://doi.org/10.1109/cvpr46437.2021.01350

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-Excitation Networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7132-7141). IEEE. https://doi.org/10.1109/cvpr.2018.00745

Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., & Liu, W. (2019). CCNet: Criss-Cross Attention for Semantic Segmentation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 603-612). IEEE. https://doi.org/10.1109/iccv.2019.00069

Kisantal, M., Wojna, Z., Muber, J., Naber, J., & Pintea, S. (2019). *Augmentation for Small Object Detection*. https://doi.org/10.48550/arXiv.1902.07296

Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature Pyramid Networks for Object Detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 936-944). IEEE. https://doi.org/10.1109/cvpr.2017.106

Lin, T., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D. et al. (2014). Microsoft COCO: Common Objects in Context. In *Lecture Notes in Computer Science* (pp. 740-755). Springer. https://doi.org/10.1007/978-3-319-10602-1_48

Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path Aggregation Network for Instance Segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8759-8768). IEEE. https://doi.org/10.1109/cvpr.2018.00913

Majumdar, R. (2025). AgroPest-12: A 12-Class Image Dataset of Crop Insects and Pests. *Kaggle*. https://www.kaggle.com/datasets/rupankarmajumdar/crop-pests-dataset

Mei, H., Ji, G. P., Wei, Z., Yang, X., Wei, X., & Fan, D. P. (2021a). Camouflaged Object Segmentation with Distraction Mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8772-8781). IEEE.

Mei, H., Ji, G. P., Wei, Z., Yang, X., Wei, X., & Fan, D. P. (2021b). PFNet: Positioning and Focusing for Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8784-8793). IEEE.

Pang, Y., Zhao, X., Xiang, T., Zhang, L., & Lu, H. (2022). Zoom in and Out: A Mixed-Scale

Triplet Network for Camouflaged Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2150-2160). IEEE. https://doi.org/10.1109/cvpr52688.2022.00220

Qin, D., Leichner, C., Delakis, M., Fornoni, M., Luo, S., Yang, F. et al. (2024). MobileNetV4: Universal Models for the Mobile Ecosystem. In *Lecture Notes in Computer Science* (pp. 78-96). Springer. https://doi.org/10.1007/978-3-031-73661-2_5

Si, Y., Xu, H., Zhu, X., Zhang, W., Dong, Y., Chen, Y. et al. (2024). SCSA: Exploring the Synergistic Effects between Spatial and Channel Attention. *Neurocomputing, 634,* Article 129866. https://doi.org/10.1016/j.neucom.2025.129866

Singh, B., & Davis, L. S. (2018). An Analysis of Scale Invariance in Object Detection Snip. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 3578-3587). IEEE. https://doi.org/10.1109/cvpr.2018.00377

Singh, B., Najibi, M., & Davis, L. S. (2018). SNIPER: Efficient Multi-Scale Training. In *Advances in Neural Information Processing Systems (NeurIPS)* (pp. 9310-9320). Curran Associates Inc.

Tan, M., Pang, R., & Le, Q. V. (2020). EfficientDet: Scalable and Efficient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 10778-10787). IEEE. https://doi.org/10.1109/cvpr42600.2020.01079

Ultralytics (2024). *YOLOv11: Real-Time Object Detection [GitHub Repository].* https://github.com/ultralytics/ultralytics

Wang, C. Y., Yeh, I. H., & Liao, H. Y. M. (2024). YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Lecture Notes in Computer Science* (pp. 1-21). Springer. https://doi.org/10.1007/978-3-031-72751-1_1

Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., & Chen, L. (2020). Axial-DeepLab: Stand-Alone Axial-Attention for Panoptic Segmentation. In *Lecture Notes in Computer Science* (pp. 108-126). Springer. https://doi.org/10.1007/978-3-030-58548-8_7

Wang, J., Chen, Y., Gao, Y., Zhang, H., & Liu, W. (2025). CPDD-YOLOv8: Small Object Detection in Aerial Images. *Scientific Reports, 15,* Article No. 770.

Woo, S., Park, J., Lee, J., & Kweon, I. S. (2018). CBAM: Convolutional Block Attention Module. In *Lecture Notes in Computer Science* (pp. 3-19). Springer. https://doi.org/10.1007/978-3-030-01234-2_1

Xu, C., Wang, J., Yang, W., Yu, H., Yu, L., & Xia, G. (2022). RFLA: Gaussian Receptive Field Based Label Assignment for Tiny Object Detection. In *Lecture Notes in Computer Science* (pp. 526-543). Springer. https://doi.org/10.1007/978-3-031-20077-9_31

Yang, C., Huang, Z., & Wang, N. (2022). QueryDet: Cascaded Sparse Query for Accelerating High-Resolution Small Object Detection. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 13658-13667). IEEE. https://doi.org/10.1109/cvpr52688.2022.01330

Zhang, H., Wang, Y., Dayoub, F., & Sunderhauf, N. (2021). VarifocalNet: An IoU-Aware Dense Object Detector. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 8510-8519). IEEE. https://doi.org/10.1109/cvpr46437.2021.00841

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q. et al. (2024). DETRs Beat YOLOs on Real-Time Object Detection. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 16965-16974). IEEE. https://doi.org/10.1109/cvpr52733.2024.01605

Zhou, Y., Sun, G., Li, Y., Fu, Y., Benini, L., & Konukoglu, E. (2025). CamSAM2: Segment Anything in Camouflaged Videos. In *Advances in Neural Information Processing Sys-*

tems (NeurIPS) (pp. 1-6). https://doi.org/10.48550/arxiv.2503.19730

Zhu, H., Li, P., Xie, H., Yan, X., Liang, D., Chen, D., Wei, M., & Qin, J. (2022). I Can Find You! Boundary-Guided Separated Attention Network for Camouflaged Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*, 3608-3616. https://doi.org/10.1609/aaai.v36i3.20273

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., & Dai, J. (2021). Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *Proceedings of the International Conference on Learning Representations (ICLR)* (pp. 1-16). https://doi.org/10.48550/arXiv.2010.04159