Journal of Computer Science and Frontier Technologies
ISSN Print: 3104-4204
ISSN Online: 3104-4212
CD ISSN: 3104-4220
https://doi.org/10.63313/JCSFT.9037

# MATE-YOLO: Multi-scale Attention Task-aligned Enhanced Detection Network for Apple Fruit in Complex Orchard Environments

**Shuai Dong*. Xiyin Liang**
College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China
Email: 727965100@qq.com

**How to cite this paper:** Dong, S., & Liang, X. Y. (2026). MATE-YOLO: Multi-scale Attention Task-aligned Enhanced Detection Network for Apple Fruit in Complex Orchard Environments. *Journal of Computer Science and Frontier Technologies*, 2(2), 8-24. ISSN Print: 3104-4204; ISSN Online: 3104-4212.
https://doi.org/10.63313/JCSFT.9037
**Published: 2026-01-16**



## Abstract

Accurate apple detection in complex orchards remains challenging due to foliage occlusion, illumination variations, and cluttered backgrounds. This study proposes an enhanced YOLOv11n framework integrating three architectural innovations. First, the EMCSP (EMA-enhanced Cross-Stage Partial) module is introduced into the backbone, synergistically incorporating multi-scale attention within cross-stage partial topology to strengthen discriminative feature extraction. Second, the ELA-HSFPN (Efficient Local Attention enhanced Hierarchical Scale Feature Pyramid Network) is devised for the neck, leveraging decoupled spatial attention and bidirectional hierarchical fusion to enhance multi-scale representation. Third, the TADDH (Task-Aligned Dynamic Detection Head) supersedes the conventional head, employing task decomposition, dynamic deformable convolution, and probabilistic feature modulation to achieve optimal classification-localization alignment. Extensive experiments demonstrate substantial improvements over baseline YOLOv11n: Precision+1.4%, Recall+2.3%, mAP@0.5+3.0%, and mAP@0.5:0.95 +1.7%. These results validate the efficacy of our methodology for intelligent fruit harvesting applications.

## Keywords

Apple detection; YOLOv11; attention mechanism; feature pyramid network; task-aligned detection head; deep learning

## 1. Introduction

The apple industry constitutes a critical component of global agricultural production, with worldwide cultivation exceeding 4.7 million hectares and annual production surpassing 95 million metric tons [1]. Apple harvesting, accounting for approximately 30-40% of total production costs, faces increasing pressure from labor shortages and rising costs, intensifying the demand for automated harvesting systems [2]. Accurate fruit detection serves as the fundamental prerequisite for

intelligent harvesting robots, directly determining picking success rates and system efficiency.

Recent advances in deep learning have revolutionized agricultural object detection. The YOLO series has garnered substantial attention owing to its exceptional balance between accuracy and efficiency[3]. YOLOv11n, as the latest nano variant optimized for edge deployment, offers promising potential for real-time fruit detection[4]. Nevertheless, deploying YOLOv11n in complex orchard environments encounters non-trivial challenges that compromise detection performance.

The first challenge stems from dense foliage occlusion, where fruits are frequently obscured by leaves and branches, resulting in incomplete visual representations. The Cross-Stage Partial topology in YOLOv11n lacks explicit attention mechanisms for handling occlusion-induced feature degradation[5]. The second challenge arises from substantial scale variations in natural orchard scenes, where conventional Feature Pyramid Networks fail to adaptively allocate attention across different scales[6]. The third challenge pertains to the inherent conflict between classification and localization tasks, where decoupled head architectures process tasks independently without explicit alignment mechanisms [7].

To address these challenges, this study proposes an enhanced YOLOv11n framework incorporating three synergistic modules. The primary contributions are summarized as follows:

(1) We propose the EMCSP module for the backbone network, integrating Efficient Multi-scale Attention into cross-stage partial topology to enhance discriminative feature extraction under occlusion conditions[8].

(2) We develop the ELA-HSFPN for the neck architecture, introducing decoupled spatial attention with bidirectional hierarchical fusion to achieve superior multi-scale representation while reducing computational complexity from $O(H^2W^2)$ to $O(H+W)$[9].

(3) We design the TADDH to replace the conventional detection head, incorporating task decomposition, dynamic deformable convolution, and probabilistic feature modulation for optimal classification-localization alignment[10].

## 2. Related Work

### 2.1. Deep Learning-based Fruit Detection

Current approaches to fruit detection can be categorized into traditional image processing and deep learning paradigms. Traditional methods rely on handcrafted features including edge detection, color space transformation, and threshold segmentation. Despite stable performance in static scenarios, their robustness against complex environmental factors such as illumination variations, foliage occlusion, and cluttered backgrounds remains insufficient, limiting practical applications.

In contrast, deep learning-based methods exhibit superior feature extraction

capability and adaptability. Single-stage detection algorithms achieve effective balance between speed and accuracy through dense prediction strategies. The YOLO series, pioneered by Redmon et al., reformulated object detection as a unified regression problem, achieving real-time performance breakthroughs [11]. Recent iterations including YOLOv8 and YOLOv9 have demonstrated remarkable performance in agricultural applications [12]. Wang et al. [13] proposed YOLOv10 with consistent dual assignments, further optimizing the accuracy-efficiency trade-off. Despite these advances, deploying general-purpose detectors in complex orchard scenarios without domain-specific adaptations often yields suboptimal performance due to the unique challenges posed by agricultural environments.

### 2.2. Attention Mechanisms in Object Detection

Attention mechanisms have become instrumental in enhancing feature representation capabilities of deep neural networks. Channel attention mechanisms enable adaptive recalibration of channel responses, while spatial attention captures location-dependent information. Coordinate Attention [14] decomposed spatial attention into two 1D encoding processes along horizontal and vertical directions, preserving precise positional information while reducing computational complexity. The Efficient Multi-scale Attention (EMA) [15] further advanced this paradigm by introducing parallel multi-scale feature aggregation combined with cross-spatial learning, enabling comprehensive multi-scale feature capture with reduced computational overhead. However, integrating such attention mechanisms into Cross-Stage Partial topology for fruit detection remains unexplored, presenting opportunities for synergistic enhancements.

### 2.3. Multi-scale Feature Fusion Networks

Multi-scale feature representation constitutes a critical challenge in object detection, particularly for targets exhibiting substantial size variations. Gold-YOLO [16] proposed Gather-and-Distribute mechanism for efficient information exchange across different levels. RT-DETR [17] introduced hybrid encoder architecture enabling cross-scale feature interaction with reduced latency. Despite these advances, existing feature pyramid networks predominantly employ simple concatenation or addition operations for feature fusion, treating all spatial locations uniformly without explicit attention mechanisms. Furthermore, the computational complexity of 2D spatial attention operations $O(H^2W^2)$ poses significant challenges for real-time deployment in resource-constrained agricultural robots. These limitations motivate the development of efficient attention-enhanced feature pyramid architectures that achieve superior multi-scale representation while maintaining computational efficiency.

## 3. Method

### 3.1. MATE-YOLO Architecture

Accurate apple fruit detection in complex orchard environments has emerged as a critical technological prerequisite for intelligent harvesting systems. Although YOLOv11n has achieved remarkable progress in general object detection tasks through its cross-stage partial topology and optimized inference architecture, the unique characteristics of orchard scenarios—severe foliage occlusion, substantial scale variations, and task misalignment—pose fundamental challenges to existing detection paradigms. Accordingly, this study proposes the MATE-YOLO (Multi-scale Attention Task-aligned Enhanced YOLO) architecture, which achieves performance breakthroughs through three synergistic innovations:

(1) EMCSP Module: The EMA-enhanced Cross-Stage Partial module is embedded into the backbone network, synergistically integrating efficient multi-scale attention mechanisms within the cross-stage partial topology. Through parallel multi-scale feature aggregation via dual-branch grouped convolutions and cross-spatial learning with channel dimension preservation, EMCSP enables adaptive feature recalibration that effectively emphasizes discriminative regions while suppressing background interference under occlusion conditions.

(2) ELA-HSFPN Architecture: The Efficient Local Attention enhanced Hierarchical Scale Feature Pyramid Network is developed for the neck architecture, introducing a novel decoupled spatial attention mechanism decomposed along horizontal and vertical axes. This decomposition strategy reduces computational complexity from $O(H^2W^2)$ to $O(H+W)$, enabling efficient deployment on resource-constrained platforms.

(3) TADDH Detection Head: The Task-Aligned Dynamic Detection Head supersedes the conventional detection head, incorporating task decomposition for generating task-specific features guided by global context, dynamic deformable convolution (DyDCNv2) for geometry-aware spatial alignment, and probabilistic feature modulation for achieving optimal alignment between classification confidence and localization precision.

The aforementioned innovations constitute a multi-level synergistic optimization paradigm, fundamentally breaking through the technical bottlenecks of fruit detection under complex orchard conditions. MATE-YOLO redefines the accuracy-efficiency boundary through feature-attention-task joint optimization.

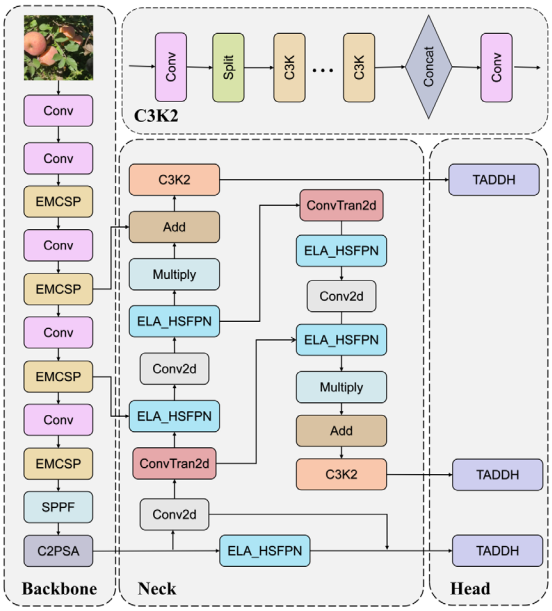



**Figure 1.** MATE-YOLO Architecture Diagram

## 3.2. EMCSP: EMA-enhanced Cross-Stage Partial Module for Robust Feature Extraction

In complex orchard environments, apple fruit detection confronts substantial challenges stemming from dense foliage occlusion, variable illumination conditions, and cluttered backgrounds. The vanilla C3k2 module in YOLOv11n, while computationally efficient through its cross-stage partial architecture, exhibits inherent limitations under such challenging scenarios. Specifically, the standard bottleneck structure processes features through sequential convolutions without explicit mechanisms for adaptive feature recalibration, leading to suboptimal representation learning when target fruits are partially occluded or exhibit appearance variations due to environmental factors.

To address these limitations, we propose the EMCSP (EMA-enhanced Cross-Stage Partial) module, which synergistically integrates an Efficient Multi-scale Attention mechanism into the cross-stage partial framework. The core innovation lies in augmenting each bottleneck unit with a lightweight yet effective attention mechanism that enables the network to dynamically emphasize salient features while suppressing background interference.For the enhanced bottleneck unit, given an input feature map $X \in R^{C\times H\times W}$, the transformation can be formulated as:

$$Y = X + A_{EMA}(\phi_{3\times3}(\phi_{3\times3}(X))) \quad (1)$$

where $\phi_{3\times3}(\cdot)$denotes the convolution operation with batch normalization and SiLU activation, and $A_{EMA}(\cdot)$ represents the EMA attention operator. The EMA mechanism decomposes the channel dimensions into multiple groups and performs parallel multi-scale feature aggregation. Formally, the attention weights are computed through a dual-branch architecture:

$$A_{EMA}(F) = \sigma(g_{1D}(F_g^{1\times1}) + g_{1D}(F_g^{3\times3})) \odot F \quad (2)$$

Where $F_g^{1\times1}$ and $F_g^{3\times3}$ denote features processed by $1\times1$ and $3\times3$ grouped convolutions respectively, $g_{1D}(\cdot)$ represents one-dimensional global average pooling for spatial compression, and $\sigma(\cdot)$ is the sigmoid activation function. The element-wise multiplication $\odot$ recalibrates the original features based on the learned attention coefficients. Subsequently, cross-spatial learning is performed to establish long-range dependencies:

$$F_{out} = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \quad (3)$$

where Q, K and Vare derived from grouped feature representations with dimension dd d. This formulation enables the module to capture both local textural patterns and global contextual information essential for distinguishing apple fruits from complex backgrounds. The complete EMCSP module stacks nn n enhanced bottleneck units within the cross-stage partial topology:

$$Z = Conact(X_{main,}B_{EMCSP}^{(n)} \cdots B_{EMCSP}^{(1)}(X_{split})) \quad (4)$$

Where $B_{EMCSP}^{(i)}$ denotes the $i-th$ enhanced bottleneck operation. By replacing all C3k2 modules in the backbone with our proposed EMCSP, the network acquires enhanced capability for multi-scale feature extraction and adaptive attention allocation, which proves particularly beneficial for detecting fruits exhibiting substantial scale variations and partial occlusions in natural orchard environments.

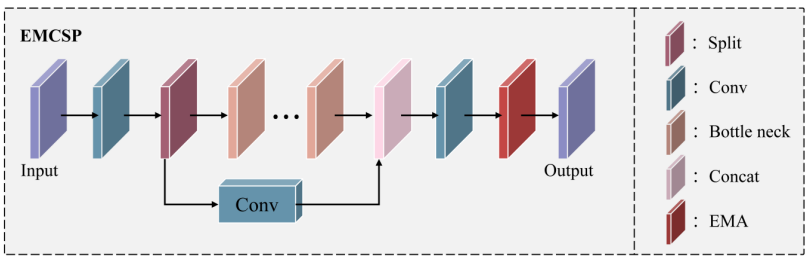



**Figure 2.** EMCSP Structural Diagram

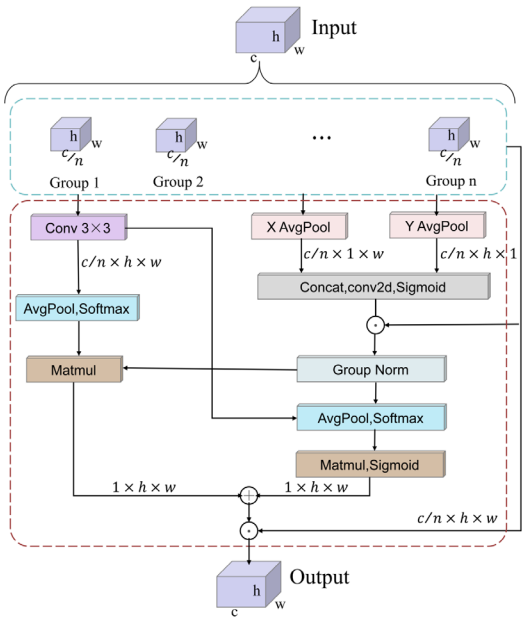



**Figure 3.** EMA Structure Diagram

### 3.3. ELA-HSFPN: Efficient Local Attention Enhanced Hierarchical Scale Feature Pyramid Network

The conventional neck architecture in YOLOv11n employs a straightforward Feature Pyramid Network (FPN) paradigm with sequential upsampling and concatenation operations for multi-scale feature fusion. However, this vanilla design exhibits critical deficiencies in complex orchard environments: (1) the simple concatenation strategy treats all spatial locations and channels equally, failing to adaptively emphasize discriminative features while suppressing background noise; (2) the unidirectional top-down pathway inadequately captures fine-grained spatial details essential for detecting small or partially occluded apple fruits; and (3) the lack of explicit attention mechanisms leads to semantic dilution during cross-scale feature propagation.

To overcome these limitations, we propose the ELA-HSFPN (Efficient Local Attention enhanced Hierarchical Scale Feature Pyramid Network), which introduces a novel decoupled spatial attention mechanism coupled with a bidirectional feature interaction strategy. The core ELA module decomposes the conventional 2D spatial attention into two orthogonal 1D operations along the horizontal and vertical axes, significantly reducing computational complexity while preserving expressive capability. Given an input feature map $F \in R^{C\times H\times W}$the ELA attention mechanism first performs directional pooling to capture global context along each spatial dimension:

$$z^h = \frac{1}{W}\sum_{j=1}^{W} F(:,:,j) \in R^{C\times H\times 1},\ \ z^w = \frac{1}{H}\sum_{j=1}^{W} F(:,i,:) \in R^{C\times 1\times W} \tag{5}$$

Subsequently, a shared 1D convolution with kernel size $k$ is applied to model local spatial dependencies, followed by group normalization and sigmoid activation to generate attention coefficients:

$$a^h = \sigma(GN(Conv1D_k(z^h))), \qquad a^w = \sigma(GN(Conv1D_k(z^w))) \tag{6}$$

where $\sigma(\cdot)$ denotes the sigmoid function and $GN(\cdot)$ represents group normalization. The final attended feature map is obtained through element-wise multiplication:

$$\hat{F} = F \odot a^h \odot a^w \tag{7}$$

This decoupled formulation reduces the computational complexity from $O(H^2W^2)$ of conventional 2D attention to $O(H+W)$, enabling efficient deployment on resource-constrained devices while maintaining strong spatial modeling capability.

Beyond the attention mechanism, ELA-HSFPN introduces a hierarchical bidirectional feature fusion paradigm that facilitates comprehensive information exchange across scales. For adjacent feature levels $P_l$ and $P_{l+1}$, the fusion process is formulated as:

$$P_l^{fused} = P_l^{up} + (A_{ELA}(P_l^{up}) \otimes \phi_{1\times1}(A_{ELA}(P_{l+1}))) \tag{8}$$

Where $P_l^{up}$ represents the upsampled feature from the deeper layer via transposed convolution, $\phi_{1\times1}(\cdot)$ denotes channel alignment through 1×1 convolution, and $\otimes$ indicates element-wise multiplication that enables adaptive gating of cross-scale information flow. This multiplicative interaction allows the network to selectively emphasize semantically consistent features while suppressing conflicting information arising from scale discrepancies. By replacing the original neck structure with our proposed ELA-HSFPN, the network achieves superior multi-scale feature representation with enhanced attention to salient regions, which proves particularly effective for detecting apple fruits exhibiting significant scale variations and complex occlusion patterns in natural orchard environments.

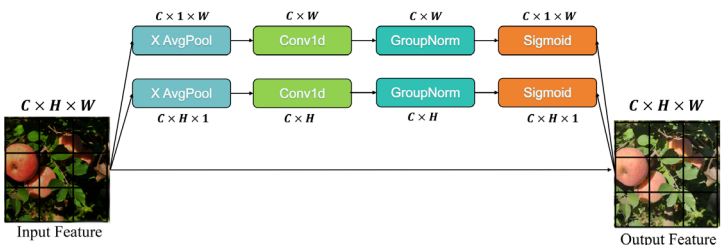



**Figure 4.** ELA Structure Diagram

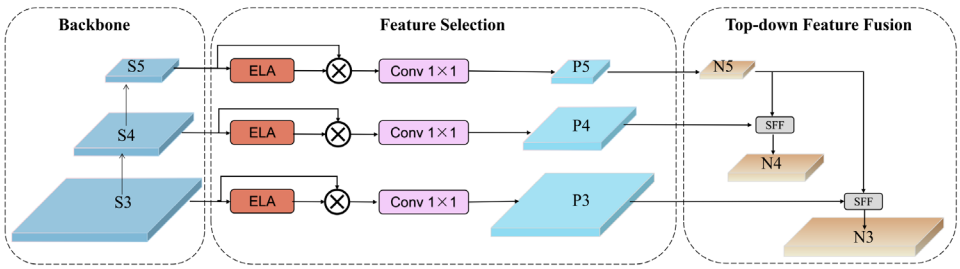



**Figure 5.** Schematic Diagram of the ELA-HSFPN Structure

### 3.4. TADDH: Task-Aligned Dynamic Detection Head for Enhanced Localization and Classification

The conventional detection head in YOLOv11n adopts a decoupled architecture that separately processes classification and regression tasks through independent convolutional branches. While this design alleviates the inherent conflict between semantic classification and precise localization, it suffers from critical limitations in complex orchard scenarios: (1) the static convolutional kernels lack spatial adaptability to handle fruits with irregular shapes and diverse poses; (2) the absence of explicit task alignment mechanisms leads to inconsistency between classification confidence and localization accuracy, where high-confidence predictions may correspond to imprecise bounding boxes; and (3) the uniform treatment of spatial locations fails to emphasize discriminative regions crucial for distinguishing densely clustered or partially occluded fruits.

To address these challenges, we propose the TADDH (Task-Aligned Dynamic Detection Head), which introduces a synergistic framework integrating task decomposition, dynamic spatial alignment, and probabilistic feature modulation. The architecture begins with a shared feature extraction stage that constructs hierarchical representations through cascaded group-normalized convolutions:

$$F_{shared} = Concat(\phi_1(X), \phi_2(\phi_1(X))) \tag{9}$$

where $\phi(\cdot)$ denotes thei − th convolutional layer with group normalization and SiLU activation, and X represents the input feature from the neck network. Subsequently, a task decomposition mechanism adaptively generates task-specific features by leveraging global context information. For the classification and regression branches, the decomposed features are computed as:

$$F_{cls} = T_{cls}(F_{shared}, GAP(F_{shared})), \quad F_{reg} = T_{reg}(F_{shared}, GAP(F_{shared})) \tag{10}$$

where$T_{cls}(\cdot)$and$T_{reg}(\cdot)$ represent the task decomposition operators that modulate spatial features based on channel-wise global statistics obtained through Global Average Pooling (GAP). This decomposition enables the network to dynamically allocate representational capacity according to task-specific requirements.

For precise localization, we introduce a dynamic deformable convolution module that adaptively adjusts sampling locations based on object geometry. The spatial offsets and modulation masks are jointly predicted from the shared features:

$$\Delta\rho, m = Split(\psi_{3\times3}(F_{shared})), \quad \hat{F}_{reg} = DyDCN(F_{reg}, \Delta\rho, \sigma(m)) \tag{11}$$

where $\psi_{3\times3}(\cdot)$ denotes a $3 \times 3$ convolution for offset prediction, $\Delta\rho \in R^{2N}$ represents the learnable offsets for $N = 9$sampling points, $m \in R^{N}$denotes the modulation scalars, and $\sigma(\cdot)$ is the sigmoid activation. The Dynamic Deformable Convolution (DyDCN) enables geometry-aware feature aggregation that adapts to the spatial extent and orientation of apple fruits.

To achieve task alignment between classification and localization, we introduce a probabilistic spatial attention mechanism that generates a classification probability

map highlighting regions with high objectness:

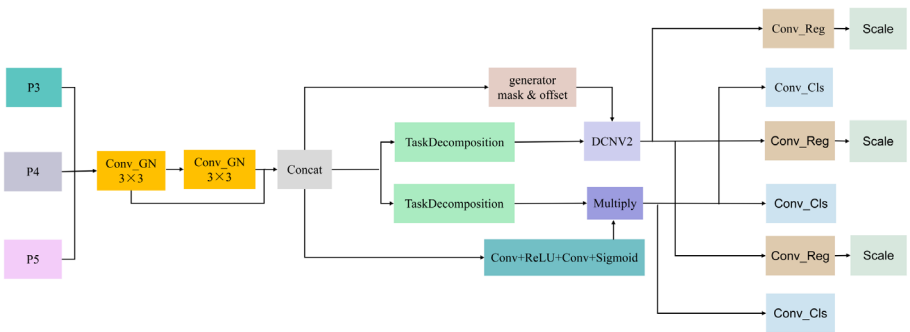
$$P_{cls} = \sigma(\psi_{3\times3}(\mathrm{ReLU}(\psi_{1\times1}(F_{shared})))) \quad (12)$$

The final detection outputs are obtained by modulating the classification features with this probability map and applying scale-adaptive transformations:

$$O_i = Concat(s_i \cdot \phi_{reg}(\hat{F}_{reg}), \phi_{cls}(F_{cls} \odot P_{cls})) \quad (13)$$

where $s_i$ represents a learnable scale factor for the $i-th$ detection level, enabling adaptive calibration across different feature pyramid levels. By replacing the original detection head with TADDH, the network achieves superior alignment between classification confidence and localization precision, which proves particularly effective for detecting apple fruits with substantial appearance variations and complex spatial distributions in natural orchard environments.

**Figure 6.** Schematic Diagram of TADDH Structure




## 4. Experimental Design and Results Analysis

### 4.1. Experimental Dataset and Environment Parameter Settings

To evaluate the proposed methodology, we constructed a comprehensive apple fruit dataset comprising 3,151 images captured from two geographically distinct orchards in Shandong Province, China: a modern standardized orchard in Zibo and a traditional orchard in Tai'an. This dual-source acquisition strategy enriches phenotypic diversity and enhances model generalization across heterogeneous environmental contexts. The dataset systematically encompasses four predominant detection challenges: illumination effects (variable lighting conditions), viewpoint effects (multi-angle observations with scale variations), occlusion effects (partial obstruction by foliage and branches), and dense clustering effects (overlapping fruit boundaries). Following standard protocols, the dataset was partitioned into training, validation, and test subsets with a 8:1:1 ratio, ensuring balanced representation across all challenge scenarios.

The experimental hardware environment for this paper consists of an Ubuntu 22.04 operating system, utilizing an NVIDIA GeForce RTX 4090 GPU with 24GB of VRAM. The software framework is based on PyTorch 2.1.2, CUDA 11.8, Python 3.8, and a CPU configuration of 16vCPU Intel® Xeon® Platinum 8352VCPU@2.10Hz.

Experimental parameters are detailed in Table 4 below:

**Table 1.** Experimental Environment Configuration Parameters Table

| Experimental | Value |
|---|---|
| Operating system<br>Deep learning framework<br>Programing language<br>image size | Ubuntu22.0<br>Pytorch2.1.2/cuda11.8<br>Python 3.8<br>640×640 |
| Epoch | 500 |
| Batch size | 32 |
| Workers | 4 |
| Momentum | 0.937 |
| Weight_decay | 0.0005 |
| lr0 | 0.01 |

## 4.2. Experimental Evaluation Criteria

To provide an accurate and objective evaluation of the proposed algorithm in this paper, multiple evaluation metrics were established to assess the model. These metrics include Precision (P), Recall (R), Mean Average Precision (mAP), model inference speed (FPS), floating-point operations (GFLOPs), and number of parameters. The calculation formulas are as follows:

$$P=\frac{TP}{TP+FP} \tag{31}$$

$$\mathrm{R}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{32}$$

$$AP=\int_0^1 P(R)\,dR \tag{33}$$

$$mAP=\frac{1}{N}\sum_{i=1}^{N} AP(i) \tag{34}$$

In the above equation, TP represents the number of samples that are actually positive and predicted as positive by the model; FP denotes the number of actual negative samples predicted as negative by the model; FN denotes the number of actual positive samples predicted as negative by the model. N represents the number of classes in the dataset. AP is the average precision value on the PR curve, while mAP denotes the average precision calculated by averaging across all classes in the dataset. This metric provides a comprehensive evaluation of the model's detection performance.

## 4.3. Experimental Results and Visualization Analysis

### 4.3.1. Ablation Experiment

To systematically investigate the individual contributions of each proposed module, comprehensive ablation experiments were conducted on the apple fruit dataset. Table 2 presents the quantitative results with various module combinations, where YOLOv11n serves as the baseline model.

When independently integrated into the baseline, each proposed module demonstrates consistent performance improvements. The EMCSP module yields

gains of 0.9%, 1.5%, 1.0%, and 0.8% in Precision, Recall, mAP@0.5, and mAP@0.5:0.95, respectively, validating the efficacy of multi-scale attention mechanisms in enhancing discriminative feature extraction under occlusion conditions. The ELA-HSFPN module achieves improvements of 0.7%, 1.7%, 1.4%, and 1.1% across the four metrics, demonstrating superior multi-scale feature representation capability through decoupled spatial attention and hierarchical fusion. Notably, the TADDH module exhibits the most substantial improvement in mAP@0.5:0.95 (+1.6%), indicating enhanced localization precision attributed to dynamic deformable convolution and task-aligned feature modulation.

The pairwise combinations further substantiate the complementary nature of the proposed modules. The integration of EMCSP and ELA-HSFPN yields cumulative improvements (mAP@0.5: +1.7%), suggesting synergistic enhancement between attention-augmented backbone features and hierarchical neck fusion. The combination of EMCSP and TADDH achieves notable Recall improvement (+2.1%), demonstrating that enhanced feature extraction coupled with task-aligned detection effectively reduces false negatives.

The full integration of all three modules achieves optimal performance with Precision of 81.4%, Recall of 67.5%, mAP@0.5 of 79.2%, and mAP@0.5:0.95 of 45.3%, corresponding to improvements of 1.4%, 2.3%, 3.0%, and 1.7% over the baseline, respectively. The parameter overhead increases moderately from 2.29M to 3.01M (+31.2%), while FLOPs increase from 5.9G to 9.7G, representing an acceptable computational trade-off for the substantial accuracy gains.

A systematic ablation study was conducted to validate the efficacy of the proposed modules and their synergistic effects. As illustrated, the baseline model achieves 76.2%, 43.6%, 80.0%, and 65.2% on mAP@0.5, mAP@0.5:0.95, Precision, and Recall, respectively.Single-module ablation reveals that EMCSP, ELA-HSFPN, and TADDH each contribute positively to model performance, with TADDH demonstrating the most substantial independent gain on mAP@0.5:0.95 (+1.6%), indicating its superiority in multi-scale object detection. Notably, dual-module combinations (EMCSP+ELA-HSFPN, EMCSP+TADDH) further unleash complementary potential, exhibiting consistent performance improvements.

**Table 2.** Results of ablation experiments on the dataset

| Methods | | | | P(%) | mAP@0.5 | mAP@0.5-0.95 | Params(M) | FLOPs(G) |
|---|---|---|---|---|---|---|---|---|
| Yolov11n | EMCSP | ELA-HSFPN | TADDH | | | | | |
| ✓ | | | | 80.0 | 76.2 | 43.6 | 2.292979 | 5.9 |
| ✓ | ✓ | | | 80.9 | 77.2 | 44.4 | 2.584267 | 6.4 |
| ✓ | | ✓ | | 80.7 | 77.6 | 44.7 | 3.430635 | 9.6 |
| ✓ | | | ✓ | 80.8 | 77.8 | 45.2 | 2.875299 | 6.1 |
| ✓ | ✓ | ✓ | | 81.2 | 77.9 | 44.8 | 3.594507 | 6.0 |
| ✓ | ✓ | | ✓ | 81.0 | 78.1 | 45.0 | 3.049968 | 8.6 |
| ✓ | ✓ | ✓ | ✓ | 81.4 | 79.2 | 45.3 | 3.006828 | 9.7 |

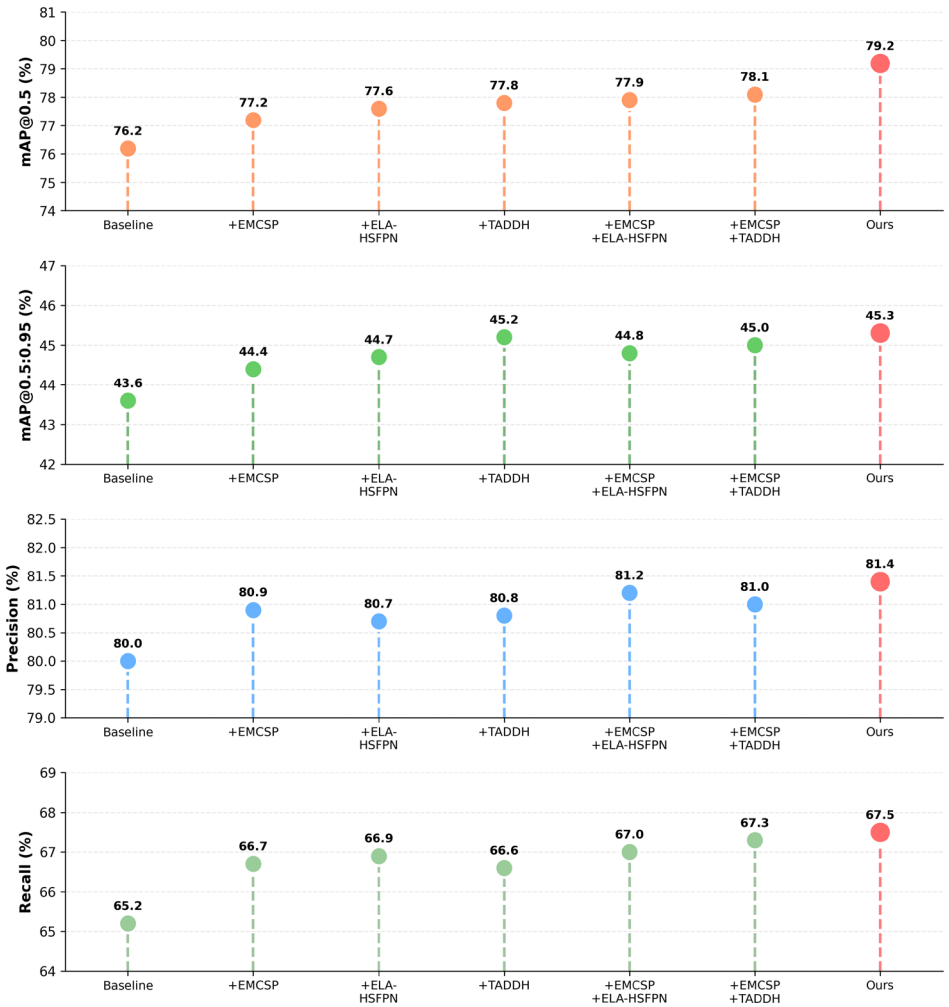



**Figure 8.** Visual comparison of experimental results for mAP@0.5, mAP@0.5:0.95, Precision, and Recall on the dataset

### 4.3.2. Comparative Experiments

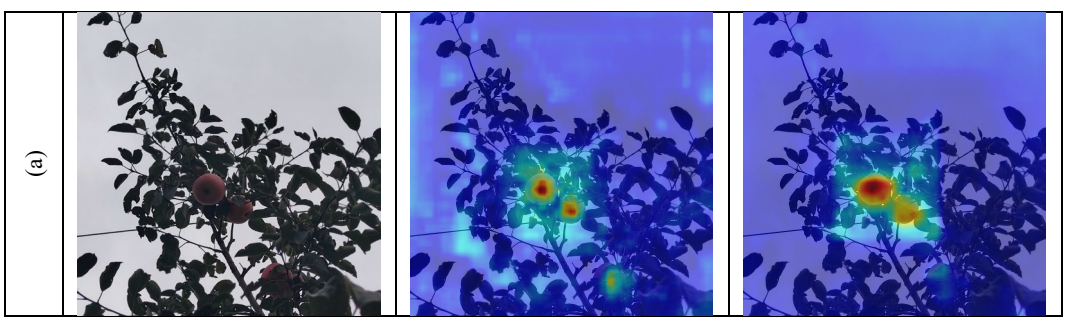
To systematically validate the detection performance of MATE-YOLO in complex orchard scenarios, this study establishes a comprehensive benchmark encompassing classical two-stage detectors (Faster R-CNN), single-stage detectors (SSD), and the latest iterations of the YOLO series . The experimental results are presented in Table three .Quantitative analysis demonstrates that MATE-YOLO achieves significant performance breakthroughs across all four core evaluation metrics, attaining 81.4%, 67.5%, 79.2%, and 45.3% for Precision, Recall, mAP@0.5, and mAP@0.5:0.95, respectively, comprehensively surpassing all comparative methods. Compared to the baseline YOLOv11s, the proposed method yields a

substantial improvement of 2.6 percentage points in mAP@0.5 while maintaining a compact architecture with only 3.01M parameters and 9.7 GFLOPs computational overhead, achieving Pareto optimality between detection accuracy and inference efficiency.

To deeply analyze the feature learning mechanism of MATE-YOLO, this study employs Grad-CAM++ technology to visualize the deep feature responses of the model. Figure nine presents the heatmap comparison results between MATE-YOLO and the YOLOv11s baseline across four representative orchard scenarios.This visualization evidence intuitively validates the synergistic enhancement mechanism of the three proposed core innovations from the feature representation dimension: the EMCSP module achieves adaptive recalibration of discriminative features through efficient multi-scale attention; the ELA-HSFPN architecture strengthens hierarchical fusion of multi-scale semantics via decoupled spatial attention; the TADDH detection head ensures optimal coupling between classification confidence and spatial localization through task-aligned dynamic modulation. The organic synergy of these three components enables MATE-YOLO to precisely capture critical semantic information of targets under extreme conditions such as complex occlusion and drastic scale variation, providing reliable technical support for intelligent orchard detection systems.

**Table 3.** Comparative experiments across different network models on the dataset

| num | Methods | P(%) | R(%) | mAP@0.5 | mAP@0.5-0.95 | Params(M) | GFLOPs |
|---|---|---|---|---|---|---|---|
| 1 | SSD | 50.8 | 57.3 | 39.0 | 15.8 | 103.8 | 162.74 |
| 2 | Faster-RCNN | 58.2 | 64.2 | 46.2 | 19.3 | 137.1 | 170.21 |
| 3 | YOLOv3-tiny | 79.9 | 57.8 | 68.9 | 39.7 | 9.525282 | 14.5 |
| 4 | YOLOv5s | 80.4 | 66.3 | 75.9 | 44.1 | 2.181859 | 5.8 |
| 5 | YOLOv6s | 80.5 | 66.0 | 75.6 | 43.5 | 4.155123 | 11.5 |
| 6 | YOLOv8s | 80.5 | 67.9 | 76.2 | 44.4 | 2.684563 | 6.8 |
| 7 | YOLOv9s | 80.8 | 66.4 | 76.4 | 44.5 | 6.194035 | 22.1 |
| 8 | YOLOv10s | 79.4 | 66.2 | 76.1 | 43.8 | 7.218387 | 21.4 |
| 9 | YOLOv11s | 80.1 | 66.2 | 76.6 | 44.2 | 2.582347 | 6.3 |
| 10 | YOLOv12s | 78.7 | 63.3 | 74.6 | 42.8 | 2.538491 | 6.3 |
| 11 | Ours | 81.4 | 67.5 | 79.2 | 45.3 | 3.006828 | 9.7 |



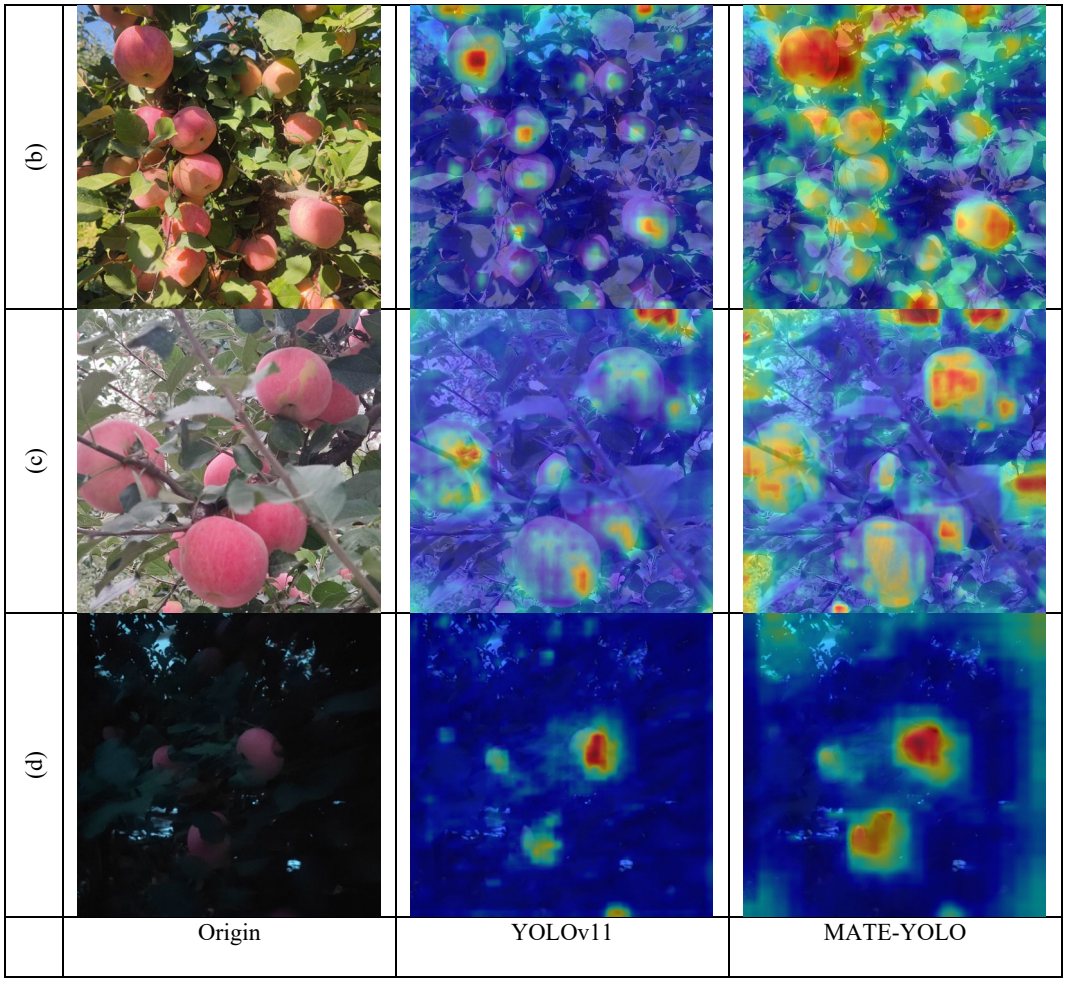



**Figure 9.** Comparative Experimental Diagram

## 5. Conclusion

This study addresses the critical challenges of severe0 foliage occlusion, substantial scale variations, and task misalignment in apple fruit detection under complex orchard environments. An enhanced YOLOv11n framework integrating three synergistic modules is proposed. The EMCSP module is embedded into the backbone network, incorporating efficient multi-scale attention within cross-stage partial topology to strengthen discriminative feature extraction under occlusion conditions. The ELA-HSFPN is designed for the neck architecture, leveraging decoupled spatial attention with bidirectional hierarchical fusion to enhance multi-scale feature representation while reducing computational complexity from $O(H^2W^2)$ to $O(H+W)$. The TADDH replaces the conventional detection head, employing task decomposition, dynamic deformable convolution, and probabilistic feature modulation to achieve optimal classification-localization alignment.

In practical applications, this study provides technical support for the vision systems of apple harvesting robots, though real-time performance and

environmental robustness require further validation. Future research will focus on integrating the detection model with robotic arm control algorithms, testing real-time multi-scale fruit detection in dynamic orchard scenarios, and expanding datasets under diverse illumination and occlusion conditions to further enhance generalization capability.

## References


[1] FAO. FAOSTAT: Crops and livestock products. Food and Agriculture Organization of the United Nations, 2023. https://www.fao.org/faostat/

[2] Zhang, C., Kang, F., & Wang, Y. (2024). Review of apple picking robots: Current status and future perspectives. Computers and Electronics in Agriculture, 217, 108584

[3] Wang, C.Y., Bochkovskiy, A., & Liao, H.Y.M. (2023). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7464-7475.

[4] Jocher, G., Qiu, J., & Chaurasia, A. (2024). Ultralytics YOLO11. https://github.com/ultralytics/ultralytics

[5] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z. (2023). Efficient multi-scale attention module with cross-spatial learning. In ICASSP 2023-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1-5.

[6] Tan, M., Pang, R., & Le, Q.V. (2020). EfficientDet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 10781-10790.

[7] Feng, C., Zhong, Y., Gao, Y., Scott, M.R., & Huang, W. (2021). TOOD: Task-aligned one-stage object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 3490-3499.

[8] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13713-13722.

[9] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018). Path aggregation network for instance segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 8759-8768.

[10] Dai, X., Chen, Y., Xiao, B., Chen, D., Liu, M., Yuan, L., & Zhang, L. (2021). Dynamic head: Unifying object detection heads with attentions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 7373-7382.

[11] Redmon, J., & Farhadi, A. (2018). YOLOv3: An incremental improvement. arXiv preprint arXiv:1804.02767.

[12] Wang, C.Y., Yeh, I.H., & Liao, H.Y.M. (2024). YOLOv9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision (ECCV).

[13] Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., & Ding, G. (2024). YOLOv10: Real-time end-to-end object detection. In Advances in Neural Information Processing Systems (NeurIPS).

[14] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13713-13722.

[15] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z. (2023). Efficient multi-scale attention module with cross-spatial learning. In ICASSP 2023-IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 1-5.

[16] Wang, C., He, W., Nie, Y., Guo, J., Liu, C., Han, K., & Wang, Y. (2024). Gold-YOLO: Efficient object detector via gather-and-distribute mechanism. In Advances in Neural Information Processing Systems (NeurIPS).

[17] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., & Chen, J. (2024). DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).