Journal of Engineering, Project, and Production Management
2026, 16(5), 2026-0007

DAMF-YOLOv8: A Multi-Class Apple Detection Algorithm Based on Deformable Attention and Multi-Scale Fusion

Ziyan Meng
Undergraduate Student, School of Automation Science and Electrical Engineering, Beihang University, 37 Xueyuan Road, Haidian District, Beijing, 100191, P.R. China, E-mail: 23376046@buaa.edu.cn

Project Management
Received April 14, 2026; revised May 14, 2026; accepted May 14, 2026
Available online June 4, 2026

Abstract: Automated apple detection in orchard environments presents significant challenges due to the prevalence of small targets, frequent occlusions, and the need for multi-category classification under complex conditions. To navigate toward solutions, this study introduces Deformable Attention and Multi-scale Fusion YOLOv8 (DAMF-YOLOv8), an enhanced detection model based on YOLOv8 that incorporates three existing modules. The first is the C2f-DLKA module, which combines deformable convolution with large-kernel attention to improve the extraction of small-target features and contextual modeling. The second is the Multi-Scale Local Channel Attention (MLCA) mechanism, employing multi-scale local-global channel attention to achieve adaptive fusion of spatial and semantic features. The third is the Detect-AFPN-P2345 network, facilitating efficient multi-scale feature fusion across P2 to P5 levels. Through systematic parameter optimization and ablation studies, the integration of these modules is empirically validated, achieving an optimal balance between detection accuracy and model efficiency. Experimental results demonstrate that DAMF-YOLOv8 achieves 91.8% mAP50 and 72.4% mAP50-95, representing improvements of 1.8 and 2.0 percentage points, respectively, over the baseline while maintaining minimal parameter growth. For challenging samples such as green apples, the model improves mAP50 by 4.6% and mAP50-95 by 3.9%. By integrating these modules, DAMF-YOLOv8 offers an effective, lightweight solution for automated orchard monitoring.

Keywords: Apple detection; lightweight model; attention mechanism; feature pyramid network; precision agriculture.

DOI 10.32738/JEPPM-2026-0007

1. Introduction

As an important agricultural product, the apple tree requires proper management, which depends on thorough observation of the individual level of fruit, growth stages and well-being (Huang et al., 2025). Therefore, accurate identification of apples and monitoring of their growth have become crucial in improving the efficiency of orchards. Traditionally, this has been done by manual checking, which is limited by high labor intensity, a great amount of time spent, and subject to the bias introduced by human factors. These constraints have inspired the design of automated detection mechanisms, making this area a research imperative in both academia and industry (Zhang et al., 2024).

The object detection research based on deep learning has become a common technology for orchard target detection (Azizi et al., 2024). Existing methodologies can be divided into two-stage and one-stage detection algorithms. Compared with two-stage algorithms, single-stage methods show significant advantages in detection efficiency. More recent studies have focused on YOLO-based models to detect apples.

Huang et al. (2025) proposed a multi-electrode sensor array system for in-situ precision sensing in orchards. While this work demonstrates the value of detailed fruit-level monitoring, it relies on sensor-based data rather than visual detection, which limits its applicability for direct visual fruit recognition and classification tasks. Wu et al. (2024) presented DNE-YOLO for apple detection in various natural environments and achieved improved accuracy. However, their method primarily targets single-class apple detection and does not address multi-category classification (e.g., intact, green, and decayed apples) within a unified lightweight framework. Yue et al. (2024) developed an improved YOLOv8n that operates in cluttered conditions to achieve a trade-off between accuracy and model lightness. Despite its balanced design, this work did not incorporate dedicated mechanisms for multi-scale feature fusion, nor did it report performance on particularly challenging categories such as green apples with low visual contrast against foliage. Guo et al. (2024) proposed a picking

decision algorithm based on YOLOv8 to manage environmental complexity and reduce decision-making time. Their focus, however, was on the picking decision pipeline rather than on enhancing the detector's ability to extract fine-grained features of small, occluded, and irregularly shaped targets.

Zhou et al. (2026) proposed an effective YOLOv8-based instance segmentation framework for multi-colored apple fruit detection and 3D localization. Their work introduced specialized model variants tailored to red, green, and yellow apples, and successfully integrated them into a unified MCA-YOLO model that achieved strong segmentation performance. The focus of that study was color-based fruit segmentation and spatial localization, which differs from the quality-based multi-category detection task (intact, green, and decayed apples) addressed in the present work.

From the above analysis, a clear research gap emerges in existing work within apple detection and monitoring, either does not employ visual detection (Huang et al., 2025) or, among vision-based YOLO approaches, lacks a systematic integration of deformable attention for small-target feature extraction, adaptive multi-scale channel attention for semantic fusion, and cross-level feature pyramid enhancement, all within a single lightweight model. Table 1 summarizes these comparisons and highlights the gap.

Table 1. Summary of related studies on apple detection and monitoring

Reference	Year	Method	Visual detection	Multi-class	Small-target enhancement	Lightweight design
Huang et al.	2025	Multi-electrode sensor array	×	×	—	—
Wu et al.	2024	DNE-YOLO	√	×	×	×
Yue et al.	2024	Improved YOLOv8n	√	×	×	√
Guo et al.	2024	YOLOv8 (picking)	√	×	×	√
Zhou et al.	2026	MCA-YOLO (segmentation + 3D)	√	color-based	×	√
DAMF-YOLOv8 (Ours)	—	YOLOv8 + C2f-DLKA + MLCA + AFPN	√	√	√	√

In complex orchard environments, existing apple detection methods face difficulties in accurately detecting small, occluded, and multi-category apples while maintaining a lightweight model suitable for deployment. To address this gap, this study proposes Deformable Attention and Multi-scale Fusion YOLOv8 (DAMF-YOLOv8), an enhanced detection model based on YOLOv8 that incorporates three existing modules: the C2f-DLKA module for enhanced small-target feature extraction, the Multi-Scale Local Channel Attention (MLCA) module for dual-branch processing and adaptive fusion, and the Detect-AFPN-P2345 module for cross-scale interaction and weighted fusion. This integration aims to improve detection accuracy on challenging samples such as green apples while remaining lightweight. The effectiveness of the proposed model is validated through systematic parameter optimization, ablation studies, and comparative experiments on a multi-class apple dataset.

2. Materials and Methods

2.1. DAMF-YOLOv8

Based on YOLOv8 (Redmon et al., 2016), we propose DAMF-YOLOv8, which incorporates three existing modules. The structure of the algorithm is illustrated in Fig. 1.

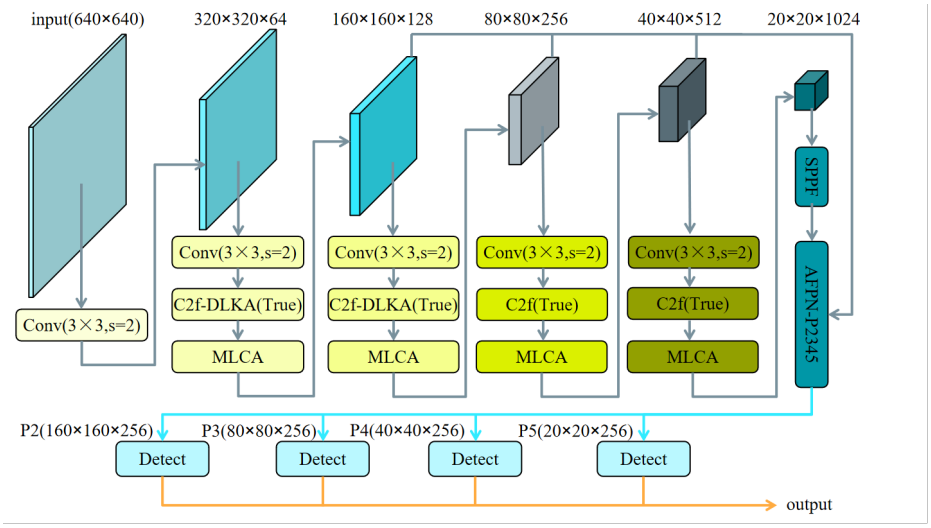

Fig. 1. Schematic diagram of the DAMF-YOLOv8 algorithm architecture

The three improvements are integrated into distinct locations within the YOLOv8 architecture. First, the C2f-DLKA module is applied only to the first two C2f blocks of the backbone to balance parameter overhead with accuracy gains. Second, the MLCA attention mechanism, which introduces negligible parameters, is embedded after every C2f module in the backbone to enhance feature representation via channel-wise recalibration. Finally, the Detect-AFPN-P2345 structure is placed in the neck and detection head to adaptively fuse multi-scale features (P2 to P5) and enable cross-level semantic interaction, improving detection across varied scales, especially for small targets.

2.2. C2f-DLKA

2.2.1. DeformConv

Conventional convolution operations utilize fixed-size and geometrically inflexible kernels. Conversely, deformable convolution adds a convolutional layer to acquire offsets on spatial positions of every sampling point, allowing the kernel to rearrange its receptive field allocation internally when extracting features (Dai et al., 2017). This adaptation mechanism guarantees accuracy in the localization of important features even in small targets with small characteristic information.

The deformable convolution module consists of two parallel convolutional sequences: a spatial offset prediction network, which is used to produce offset field Δ, and then the dynamic feature convolution layer generates improved feature representations. The simplified forms are shown in Eqs. (1) and (2).

$$\Delta \in \mathbb{R}^{B\times 2k^2\times H\times W} = \mathrm{Conv2d}\left(\mathbf{X}, \mathbf{W}_{offset}\right) \tag{1}$$

$$\mathbf{Y} \in \mathbb{R}^{B\times C_{\mathrm{out}}\times H\times W} = \mathrm{DeformConv2d}(\mathbf{X}, \mathbf{W}, \mathbf{\Lambda}) \tag{2}$$

By replacing a standard convolution with a conventional convolution followed by a dynamic convolution, and subsequently encapsulating the result, a readily deployable deformable convolution module is constructed. As an example, the structure of a 3 × 3 deformable convolution is illustrated in the figure below.

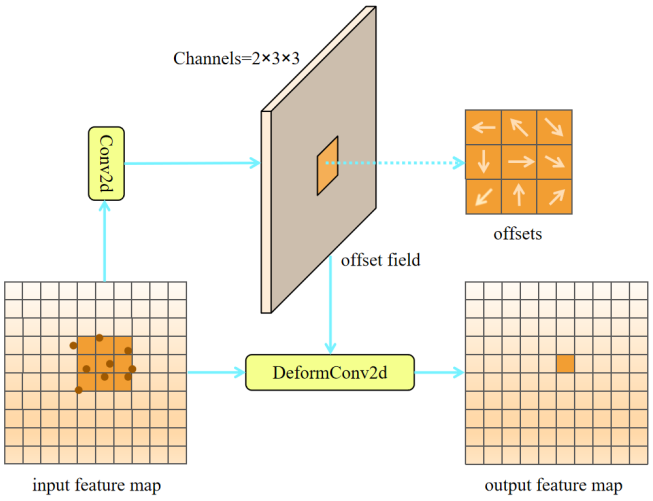

Fig. 2. Schematic diagram of 3×3 deformable convolution

2.2.2. Large kernel attention (LKA)

Traditional methods to model spatial information normally only use the large-kernel convolutions that increase the size of the kernel to adequately encompass the spatial relations across pixels. Thus, in order to mitigate this drawback, the LKA method (Azad et al., 2024) is proposed. LKA is the basic design principle of breaking down a large-kernel convolution into the following consecutive operations. At the first level, the depthwise separable convolution obtains local features. The receptive field is then expanded to a global one by using depthwise dilated convolution to amplify feature localization. It is then immediately succeeded by a 1 × 1 convolution to incorporate channel-wise data, producing attention weights. Last, the attention weights are multiplied with the original features to give the refined feature map.

2.2.3. C2f-DLKA

The DLKA module (Shu et al., 2024) effectively combines deformable convolution with LKA, where the former enables the accurate localization of fine-grained features and the latter then provides efficient global spatial modeling. The DLKA module has three important key layers, which are obtained by substituting the first and the second convolutional layers in LKA with deformable convolution.

However, DLKA introduces more parameters than a standard convolution. To overcome the problem, DLKA is integrated into the bottleneck layer of the C2f module. The original C2f bottleneck uses a 3 × 3 convolution for dimensionality reduction, followed by another 3 × 3 convolution for restoration. By replacing the restoration convolution with DLKA, the computational cost is substantially reduced due to the dimensionality reduction step. The resulting C2f-DLKA module retains the original C2f structure, with the bottleneck dimension set to c and three consecutive Bottleneck_DLKA layers,

as illustrated in the figure below.

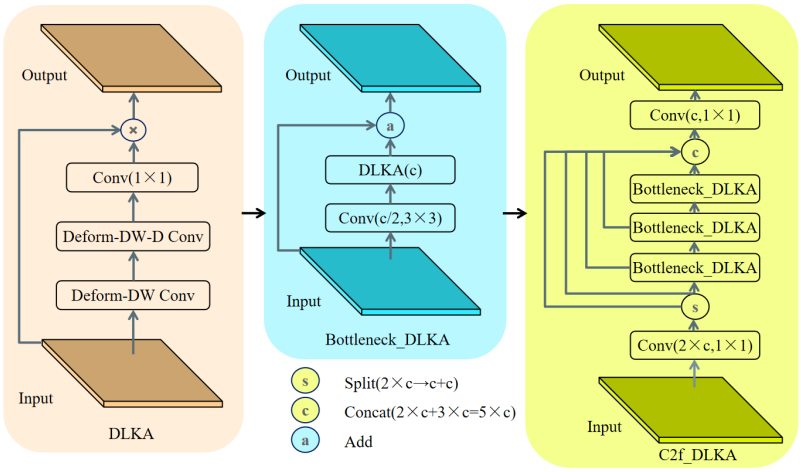

Fig. 3. Schematic diagram of the C2f-DLKA architecture

2.3. MLCA

2.3.1. Efficient channel attention (ECA)

ECA is a lightweight channel attention. It is based on the fundamental idea of substituting fully connected layers with 1D convolution (Wang et al., 2020), which makes computational cost less expensive and inter-channel dependencies modeled.

The input feature map is first spatially compressed using Global Average Pooling (GAP), which converts the input characteristics of form (B, C, H, W) to a feature map of the form (B, C, 1, 1). This step reduces the spatial contents of each channel to one scalar, which neglects the spatial content and only concentrates on channel contents. The (B, C, 1, 1) feature map is then reshaped to (B, 1, C). Afterward, 1D convolution is used to extract local channel dependencies, and a Sigmoid function is used to squeeze the values in the range of [0, 1] to produce attention weights. These weights are reconfigured to (B, C, 1, 1) and reshaped spatially to (B, C, H, W). Last but not least, a multiplication of the weights and original input features is performed to produce feature improvement.

A key innovation of Efficient Channel Attention (ECA) is its ability to dynamically adjust the size of the 1D convolution kernel based on the number of input channels, as shown in Eq. (3).

$$t = \left\lfloor \frac{|\log_2(C)+b|}{\gamma} \right\rfloor \tag{3}$$

In Eq. (3), where C is the number of channels in the input feature map, b is the bias term typically set to 1, and γ is the scaling factor (gamma) usually set to 2. If t is odd, then $k = t$; if t is even, then $k = t + 1$. Through t is a dynamic adaptive adjustment of the convolution kernel size. ECA ensures that the effective modeling range for channel dependencies precisely aligns with the number of channels.

2.3.2. MLCA

MLCA (Lu et al., 2025) has the structure that is shown in Fig. 4. The model used to model the local channel dependence is essentially the same as modeling the global channel dependence. Local adaptive average pooling (LAP) is first used to reduce the number of input features into a feature map of size (B, C, L, L) to maintain local spatial features. It is then reshaped to (B, 1, C×L×L), and 1D convolution is achieved to make local dependencies between channels. The result is reshaped back to (B, C, L, L) and then filtered through a Sigmoid to derive local attention weights. At this point, channel weights of positions L × L are obtained.

Simultaneously, the compressed feature map (B, C, L, L) is processed using the ECA method to derive global channel weights of shape (B, C, 1, 1). These weights are then further processed by a 1D convolution, after which they are spatially expanded to (B, C, L, L). The global channel weights and local channel weights are fused through weighted summation (Wang et al., 2025). The fusion weights are dynamically optimized during training to seek the optimal solution.

The fusion of local and global channel attention is governed by two learnable scalar parameters, α and β, initialized to 0.5. During training, these parameters are updated via standard backpropagation to dynamically balance the contribution of local spatial details and global semantic context. This adaptive mechanism differs from non-adaptive fusion methods, such as fixed-weight summation or channel-wise concatenation, which cannot adjust to varying feature importance across inputs. At inference, α and β are fixed, ensuring deterministic behavior.

C2f-DLKA and MLCA are complementary and synergistic: the former is able to capture the small target details accurately and at a high spatial level, while the latter supports the semantics of the prominent features in terms of local and global channel modeling, which links them to the context of the global semantics. C2f-DLKA exposes important properties of occlusion targets, including arc-shaped shapes and slender stem channels, and MLCA reinforces these minor properties by localized channel modeling and later connects them with large features such as color channels through weighted fusion.

This allows semantic integration in which these finer-grained features are associated with common target characteristics.

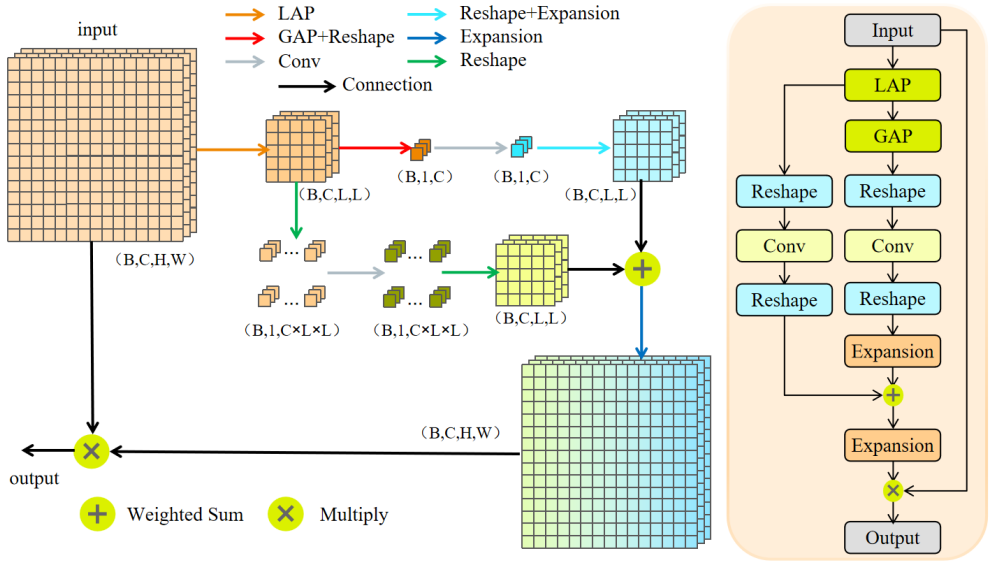

Fig. 4. Architecture of multi-scale local channel attention

2.4. Detect-AFPN-P2345

The neck network of YOLOv8 uses channel-wise feature concatenation and combines channel feature maps at varying levels of a hierarchy with a constant ratio between 1 and 1. However, the Adaptive Spatial Feature Fusion (ASFF) approach has learnable weighting parameters, which are dynamically changed by the use of adaptive spatial weights of individual feature maps throughout the training procedure, thus attaining adaptive feature incorporation (Liu et al., 2019), and this method also supports the combination of n feature maps.

The specific flow of Adaptive Spatial Feature Fusion (ASFF) is as follows: The 1 × 1 convolutions would first be used to compress the dimensions of the feature maps, as shown in the figure, before 8 channels per map are fused together. They are followed by the channel-wise concatenation of the n feature maps to make an 8×n-channel feature map, and subsequently, an n-channel feature map is created through another 1 × 1 convolution. Each channel in this output corresponds to the weight information of its respective input feature map. The n-channel feature map is then normalized through the Softmax function, so that the weights of all the n channels are added to 1. Finally, weighted fusion is performed to generate the combined feature map, which is further refined through a 3 × 3 convolution to produce the final output feature map.

Next, the feature fusion process is introduced. Based on the number of input scales, the feature map is upsampled or downsampled to all other scales. At each target scale, ASFF adaptively fuses the original feature map with all transformed ones, producing fused feature maps for all scales. The 3-scale fusion process is illustrated in the schematic below. The 2-scale and 4-scale fusion processes can be done in the same manner as the 3-scale fusion.

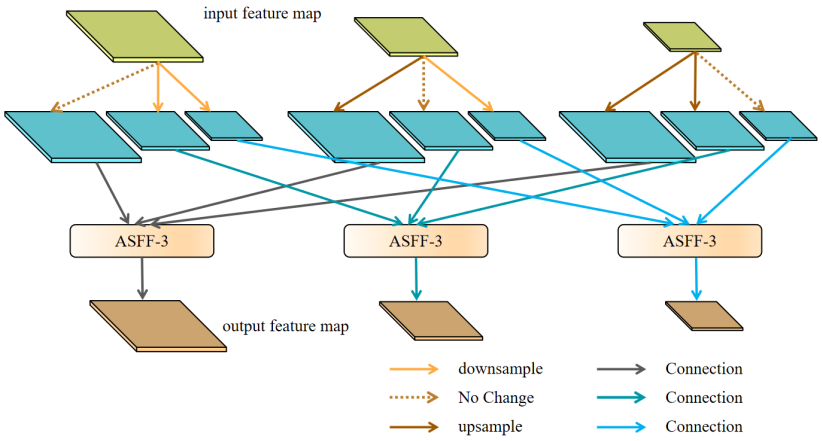

Fig. 5. Schematic illustration of the triple-scale feature fusion process

The entire process of Detect-AFPN-P2345 is then presented (Li et al., 2023). The dimensionality reduction is initially done by 1 × 1 convolutions, which is set to downsample the number of channels in the channel map per feature map by a factor of four. This is completed by 2-scale, 3-scale, and 4-scale fusion operations. Each fusion output undergoes feature

enhancement before serving as input to the subsequent fusion stage (where each fusion process takes the feature-enhanced output from the previous fusion as input). The resulting four fused feature maps are then processed by a 1 × 1 convolution to restore the channel dimension to a predefined size, 256 channels in this study. This single unified scale of 256 channels is ultimately detected by the detection head to produce the final output. The following diagram shows the workflow (Mou et al., 2023).

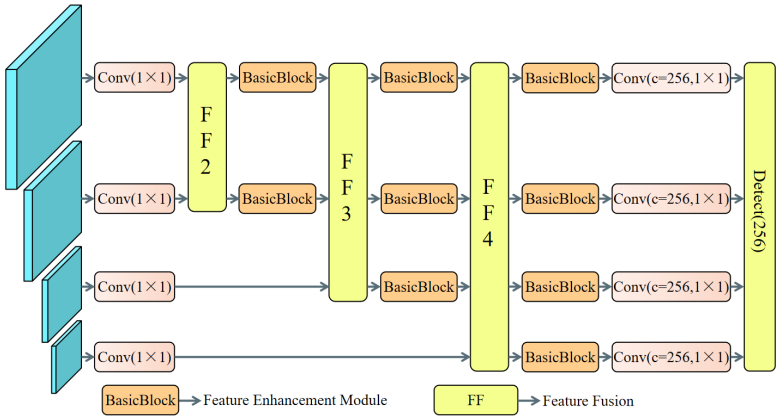

Fig. 6. Architectural diagram of Detect-AFPN-P2345

3. Results and Analysis

3.1. Experimental Preparation

The original images were collected from publicly available online orchard image repositories, totaling approximately 2,000. To enhance data diversity and model generalization, offline data augmentation was applied using the following stochastic operations: random horizontal flipping, random rotation (±15°), random scaling (0.8–1.2), HSV color space jitter (hue ±0.02, saturation ±0.3, value ±0.3), brightness and contrast adjustment (±0.2), and Gaussian noise addition (σ=0.01). These augmentations expanded the dataset to the final 3,053 images. The detailed category distribution (instance counts) with the 3:1 training-validation split is presented in Table 2.

Table 2. Category distribution of the apple dataset

Category	Training	Validation	Total
Intact apple	1,570	677	2,247
Green apple	1,905	629	2,534
Decayed apple	3,087	606	3,693

Manual annotation was performed using LabelImg (Wang et al., 2024), with the following labeling criteria: intact apples, smooth, undamaged skin, unripe green apples, predominantly green appearance and decayed apples with visible rot or surface damage. As the images were sourced from diverse online repositories, specific capture conditions (camera models, lighting, occlusion levels) are not documented; however, the dataset inherently encompasses a broad range of natural orchard environments.

All experiments are conducted on Windows 11 using Python 3.8.20 and PyTorch 1.10.1. The model typically converges around 150 epochs, with a total of 200 epochs set for all models to ensure full convergence. The AdamW optimizer was selected for this study because its decoupled weight decay regularization is particularly beneficial for lightweight architectures, as it helps constrain model complexity without interfering with the gradient updates of the loss function, thereby improving generalization and preventing overfitting. The base learning rate was set to 0.001429, which was determined through a learning rate range test to achieve the optimal balance between convergence speed and training stability. A momentum of 0.9 and a weight decay of 0.0005 were applied. The batch size was set to 16.

It should be noted that YOLOv8 was selected as the baseline architecture for this study based on three considerations. First, YOLOv8 offers a mature, well-documented ecosystem with robust support for pluggable modules, enabling stable and reproducible integration of the proposed components. Second, YOLOv8 remains the most widely adopted benchmark in recent precision agriculture and orchard detection literature, which maximizes the comparability of our results with the existing body of research. Third, the three modules proposed in this work include C2f-DLKA, MLCA, and Detect-AFPN-P2345, which are architecturally decoupled from any specific YOLO implementation and can be readily transferred to newer versions such as YOLOv11 or YOLO26. Cross-version migration is identified as a promising direction for future investigation.

Detection accuracy is evaluated using mean Average Precision (mAP) (Maxwell et al., 2021), while model complexity and storage are assessed by parameter count and model size. mAP_{50} represents the average precision at an IoU threshold

of 0.5. mAP_{50-95} averages precision across IoU thresholds from 0.5 to 0.95.

3.2. Model Performance

Compared with the baseline YOLOv8n, the proposed improved model achieves a 1.8 percentage point increase in mAP_{50} (rising from 90.0% to 91.8%) and a 2.0 percentage point increase in mAP_{50-95} (rising from 70.4% to 72.4%).

Table 3. Performance of YOLOv8 and DAMF-YOLOv8

Algorithm	Category	mAP50 (%)	mAP50-95 (%)	Parameters
YOLOv8	All	90.0	70.4	3006233
	Intact apple	91.5	70.4	
	Green apple	83.7	58.2	
	Decayed apple	94.8	82.5	
DAMF-YOLOv8	All	91.8	72.4	3091101
	Intact apple	91.4	70.6	
	Green apple	88.3	62.1	
	Decayed apple	95.7	84.4	

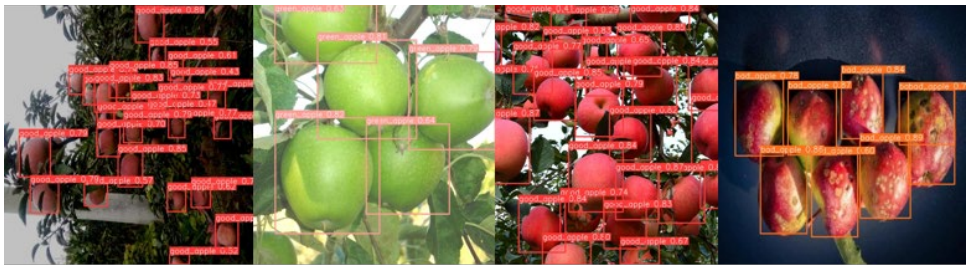

Fig. 7. Representative detection results of DAMF-YOLOv8 across all three categories.

Note: From left to right: intact apples (sub-images 1 and 3), green apples (sub-image 2), and decayed apples (sub-image 4). Multiple instances are successfully detected in each sub-image under natural orchard conditions with partial occlusion.

According to the fine-grained categorical detection performance perspective, the enhanced model has distinct benefits. The mAP_{50} and mAP_{50-95} measures also come in at 4.6 percent and 3.9 percent improvement in the task of green apple detection. The results of this finding have two meanings. Firstly, it confirms the better ability of the model to capture features in smaller-sized targets, and secondly, and more importantly, it demonstrates that the augmented modules are efficient in dealing with the issue of revealing challenging samples that arise because of the visual resemblance of targets and the background. Green apples form generic representations of the difficult samples - small size and low contrast with the environment. As demonstrated by the example figures, the refined model (right picture) is inherently more accurate in localization accuracy when detecting green apples, but also more adequate in semantically isolating targets and the backdrop.

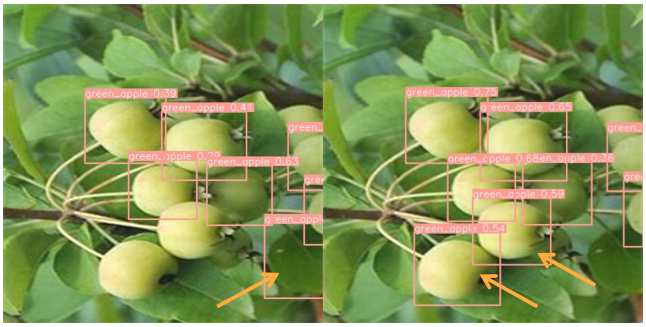

Fig. 8. Comparison of green apple detection between YOLOv8 (left) and DAMF-YOLOv8 (right).

Note: DAMF-YOLOv8 achieves more accurate localization and better semantic separation of the targets from the background.

3.3. Ablation Study

Based on a systematic analysis of ablation experiments, this study conducts a rigorous evaluation of the effectiveness and synergistic effects of each proposed improvement module in the DAMF-YOLOv8 algorithm, as detailed in Table 4.

Table 4. Ablation study

YOLOv8	MLCA	C2f-DLKA	Detect-AFPN-P2345	mAP50 (%)	mAP50-95 (%)	Parameters
√				90.0	70.4	3006233
√	√			90.8	71.3	3006265
√		√		91.3	71.2	3478325
√			√	90.5	70.4	2618977
√	√	√		91.8	71.5	3478357
√	√		√	91.3	71.3	2619009
√		√	√	91.2	70.9	3091069
√	√	√	√	91.8	72.4	3091101

The baseline achieves 90.0% mAP_{50} and 70.4% mAP_{50-95}. Integrating MLCA increases mAP_{50} by 0.8 percentage points with a slight parameter increase. C2f-DLKA boosts mAP_{50} by 1.3 percentage points while significantly enhancing small-target feature extraction. Detect-AFPN-P2345 reduces parameters by 12.9% while maintaining baseline accuracy. Regarding module combinations, MLCA with Detect-AFPN-P2345 achieves a favorable trade-off between lightweight design and accuracy, yielding a 12.9% parameter reduction and a 1.3 percentage point mAP_{50} gain, making it suitable for resource-constrained scenarios. The combination of MLCA and C2f-DLKA further improves mAP_{50} by 0.5 percentage points over C2f-DLKA alone, demonstrating effective complementarity.

Finally, the combination of all three modules achieves optimal performance, delivering 91.8% mAP_{50} and 72.4% mAP_{50-95} with a similar parameter count.

3.4. Comparative Experiments

Table 5. Comparative experiments

Model	mAP50 (%)	mAP50-95 (%)	Size of Model (MB)
DAMF-YOLOv8	91.8	72.4	6.61
SSD	79.9	55.0	96.08
Faster RCNN	77.8	54.1	113.51
YOLOv3n	87.4	63.3	4.79
YOLOv5n	89.0	67.2	4.74
YOLOv11n	90.1	69.2	6.24

The results of the comparative experiments are shown in Table 5 (Kuznetsova et al., 2020). DAMF-YOLOv8 has far exceeded both traditional detectors and modern lightweight models in terms of its state-of-the-art detection accuracy, achieving 91.8 percent mAP_{50} and 72.4 percent mAP_{50-95}. Additionally, the model size is much smaller than that of SSD (Sun et al., 2021) and Faster RCNN (Fu et al., 2020). Of special significance is its higher performance in mAP_{50} by 1.7 percent compared to YOLOv11n, without a significant change in its model size. The architecture has managed to balance computational efficiency and detection precision, achieving an optimal trade-off that is especially applicable to practical orchard applications where accuracy and deployment efficiency are important.

4. Discussion and Prospects

4.1. Interpretation of Module Contributions

This paper proposes DAMF-YOLOv8, an enhanced model integrating three key modules for automated apple detection in orchard environments.

The C2f-DLKA module enhances the detection of diverse apples. The deformable convolution unit focuses on the entire contours of intact apples, the small-scale features of green apples, and the irregular deteriorated regions of decayed apples, and the large kernel attention provides better separation results among the green apples and leaf backgrounds and the contextual information of the deteriorated region of decayed apples. The MLCA module applies to exact feature enrichment on a semantic level: it reinforces the responses of the red-yellow channel on intact apples with strong color features, concentrates on the arc-shaped contour features on the edge areas on green apples and at the same time, it averts the global color features of intact areas and local texture differences on the decayed areas on decayed apples. The introduction of the P2 feature layer and the multi-scale adaptive fusion mechanism in the Detect-AFPN-P2345 module is effective for retaining the fine edges of green apples and the small-scale decay structure of decayed apples, and for providing full flexibility in integrating the overall shape semantics and local texture of intact apples.

While MLCA and Detect-AFPN-P2345 are themselves existing modules built upon ECA and ASFF, respectively. The contribution of DAMF-YOLOv8 lies in how these components are extended and, more importantly, how they work together as an integrated system. MLCA extends ECA by introducing a parallel local attention branch with learnable fusion weights, achieving adaptive multi-scale channel attention that ECA alone cannot provide. Detect-AFPN-P2345 extends ASFF by incorporating the P2 feature layer and cascaded multi-scale fusion across P2 to P5 levels. Most critically, the three modules are complementary and synergistic: C2f-DLKA captures fine-grained spatial details of small and occluded targets, MLCA enriches channel-wise semantic responses, and Detect-AFPN-P2345 adaptively fuses these features across resolution levels. As demonstrated in the ablation study, the complete DAMF-YOLOv8 achieves 91.8% mAP50, surpassing all partial combinations. This confirms that the integrated model delivers improvements beyond the sum of its individual components.

4.2. Threats to Validity

Several factors should be considered when interpreting the findings of this study. Internal validity: Training outcomes can be sensitive to hyperparameter configurations. Although we employed a learning rate range test to determine the optimal learning rate and used AdamW with standard momentum and weight decay settings, small variations in batch size or augmentation intensity could influence the reported performance figures.

External validity: The dataset was sourced from publicly available online images with limited documented provenance regarding cultivars, geography, and weather conditions. Consequently, the model's generalizability to other apple varieties or substantially different orchard environments has not yet been fully established.

Construct validity: While mAP50 and mAP50-95 are standard metrics for object detection, practical orchard deployment may require additional measures such as inference latency and per-class false-negative rates. These aspects are partially addressed in this study through parameter count and model size comparisons, but a comprehensive real-time evaluation remains to be conducted.

Conclusion validity: All reported experimental results were averaged over three independent training runs to mitigate random variation. However, formal statistical significance testing (e.g., paired t-tests with confidence intervals) was not performed and is planned for future work.

4.3. Limitations and Future Work

Regarding robustness to domain shifts, the diverse data augmentation strategies employed in this study, particularly HSV color jitter and geometric transforms, provide a degree of invariance to illumination and viewpoint variations within the current dataset. However, systematic evaluation on external datasets collected under substantially different conditions (e.g., different apple cultivars, heavy rain, nighttime, and varied phenological stages) has not yet been conducted. Consequently, future research will proceed along three principal directions. First, building more varied datasets encompassing a wider range of deviations in illumination, weather, and phenological levels will be essential to further validate and strengthen the model's environmental robustness. Second, extending the detection framework to handle more than three categories, including more detailed growth stages and health conditions, will enhance its practical utility for comprehensive orchard monitoring. Third, investigating more sophisticated compression methods such as knowledge distillation and structured pruning will help achieve an optimal balance between detection quality and inference speed, facilitating real-time deployment on resource-constrained edge devices.

5. Conclusion

Overall, this paper introduces an enhanced model, DAMF-YOLOv8, which is specifically designed to generate accurate results for apples in the complex orchard environment. The model integrates complementary modules to address key challenges in apple detection. The proposed approach includes three well-established existing modules: C2f-DLKA, MLCA, and Detect-AFPN-P2345. Experimental results show that, despite having virtually the same model size, DAMF-YOLOv8 achieves superior overall detection performance of 91.8% and 72.4% at mAP50 and mAP50-95, respectively, which are 1.8 and 2.0 percentage points higher than the baseline YOLOv8 model. This study provides a feasible technical solution for automated orchard monitoring and, accordingly, offers useful references for developing systems to advance agricultural visions.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The author used Grammarly solely for language editing and readability improvement. The author reviewed and verified all content and takes full responsibility for the accuracy and integrity of the manuscript.

References

Azad, R., Niggemeier, L., Hüttemann, M., Kazerouni, A., Aghdam, E. K., Velichko, Y., Bagci, U., and Merhof, D. (2024). Beyond self-attention: Deformable large kernel attention for medical image segmentation. In Proceedings of the *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1287-1297.

Azizi, A., Zhang, Z., Hua, W., Li, M., Igathinathane, C., Yang, L., Ampatzidis, Y., Ghasemi-Varnamkhasti, M., Radi, R., Zhang, M. and Li, H. (2024). Image processing and artificial intelligence for apple detection and localization: A comprehensive review. *Computer Science Review*, 54, 100690.

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., and Wei, Y. (2017). Deformable convolutional networks. In Proceedings of the *IEEE International Conference on Computer Vision (ICCV)*,764-773.

Fu, L., Majeed, Y., Zhang, X., Karkee, M., and Zhang, Q. (2020). Faster R‐CNN‐based apple detection in dense-foliage fruiting-wall trees using RGB and depth features for robotic harvesting. *Biosystems Engineering*, 197, 245-256.

Guo, Z., Liu, Z., Dong, Q., Wang, K., Yan, X. A., and Su, Y. (2024). Research on picking decision algorithm of apple picking robot based on improved YOLOv8. In 2024 *12th International Conference on Information Systems and Computing Technology (ISCTech)*, 1-7.

Huang, W., Yang, H., Wang, Y., Ding, P., Nawi, N. M., and Zhang, X. (2025). In-situ precision sensing for smart agriculture using multi-electrode sensor array systems in orchards. *Sensors and Actuators A: Physical*, 382, 116134.

Kuznetsova, A., Maleva, T., and Soloviev, V. (2020). Using YOLOv3 algorithm with pre- and post-processing for apple detection in fruit-harvesting robot. *Agronomy*, 10(7), 1016.

Li, Y., Wang, L., Lu, H., Li, K., and Wang, S. (2023). AFPN: Adaptive feature pyramid network for object detection.

Liu, S., Huang, D., and Wang, Y. (2019). ASFF: Learning spatial fusion for single-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),* 2996-3005.

Lu, F., Yao, S., Sun, G., Zhou, T., and Huang, Y. (2025). FMS-YOLO: A lightweight safety belt detection algorithm for high-altitude workers based on attention mechanism and efficient architecture. *Journal of Real-Time Image Processing*, 22, 90.

Maxwell, A. E., Warner, T. A., and Guillén, L. A. (2021). Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—Part 1: Literature review. *Remote Sensing*, 13(13), 2450.

Mou, C., Wang, K., and Wang, L. (2023). AFPN: Asymptotic feature pyramid network for object detection. In Proceedings of the *IEEE International Conference on Multimedia and Expo (ICME)*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* 779-788.

Shu, R., Chen, L., Su, L., Li, T., and Yin, F. (2024). DLCH-YOLO: An object detection algorithm for monitoring the operation status of circuit breakers in power scenarios. *Electronics*, 13(19), 3949.

Sun, H., Xu, H., Liu, B., He, D., He, J., and Zhang, H. (2021). Mean-SSD: A novel real-time detector for apple leaf diseases using improved light-weight convolutional neural networks. *Computers and Electronics in Agriculture*, 189, 106379.

Wang, G., Li, H., Li, P., Lang, X., Feng, Y., Ding, Z., and Xie, S. (2024). M4SFWD: A multi-faceted synthetic dataset for remote sensing forest wildfires detection. *Expert Systems with Applications*, 248, 123489.

Wang, M., Luo, J., Lin, K., Chen, Y., Huang, X., Liu, J., Wang, A., and Xiao, D. (2025). Colony-YOLO: A lightweight micro-colony detection network based on improved YOLOv8n. *Microorganisms*, 13(7), 1617.

Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W., and Hu, Q. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR,)* 11534-11542.

Wu, H., Mo, X., Wen, S., Wu, K., Ye, Y., and Wang, Y. (2024). DNE-YOLO: A method for apple fruit detection in diverse natural environments. *Journal of King Saud University - Computer and Information Sciences*, 36(9), 102220.

Yue, Y. R., Cui, S. H., and Shan, W. (2024). Apple detection in complex environment based on improved YOLOv8n. *Engineering Research Express*, 6(4), 045259.

Zhang, G., Tian, Y., Yin, W., and Zheng, C. (2024). An apple detection and localization method for automated harvesting under adverse light conditions. *Agriculture*, 14(3), 485.

Zhou, J., Chen, M., Zhang, M., Zhang, Z., Zhang, Y., and Wang, M. (2026). Improved YOLOv8 for multi-colored apple fruit instance segmentation and 3D localization. *Artificial Intelligence in Agriculture*, 16(1), 381-396.

Ziyan Meng is currently an undergraduate student in Automation at Beihang University, expected to receive his Bachelor of Engineering degree in 2027. He has received the first prize in the Beijing College Student Mathematics Competition, the First Prize in the Beijing College Student Physics Competition and the Honorable Mention in the Mathematical Contest in Modeling (MCM/ICM). His research interests include computer vision, object detection, and intelligent robotics.