Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Hierarchical Fusion of RGB and Estimated Depth for Apple Instance Segmentation in Complex Orchards

Tao Zhang

ztncwu@126.com

North China University of Water Resources and Electric Power

Jiaxuan Xia

North China University of Water Resources and Electric Power

Jinxing Niu

North China University of Water Resources and Electric Power

Chang Liu

North China University of Water Resources and Electric Power

Yuhang Wang

North China University of Water Resources and Electric Power

Xingyue Xu

North China University of Water Resources and Electric Power

Dingyi Ma

North China University of Water Resources and Electric Power

Yingchun Deng

Henan Provincial Institute of Food and Salt Inspection Technology

Research Article

Keywords: apple instance segmentation, edge-aware downsampling, feature fusion, monocular depth estimation, RGB-D perception

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11462623/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Hierarchical Fusion of RGB and Estimated Depth for Apple Instance Segmentation in Complex Orchards

Tao Zhang[1]*, Jiaxuan Xia[1], Jinxing Niu[1],Chang Liu[1], Yuhang Wang[1], Xingyue Xu[1], Dingyi Ma[1], Yingchun Deng[2]

Correspondence:
Tao Zhang
ztncwu@126.com
[1]North China University of Water Resources and Electric Power, Zhengzou,Henan 450045,China

[2]Henan Provincial Institute of Food and Salt Inspection Technology, Zhengzhou,Henan 450003,China

Abstract: Illumination variation, foliage occlusion, and fruit overlap pose substantial challenges to apple instance segmentation in complex orchard environments. This study proposes HEF-YOLO (Hierarchical Heterogeneous Fusion and Edge-aware Downsampling You Only Look Once), an RGB-D multimodal segmentation network based on YOLO11. Red-green-blue (RGB) images are combined with depth maps estimated from the same images using Depth Anything V2, where the estimated depth serves as an auxiliary geometric representation rather than an independent metric measurement. An asymmetric dual-branch backbone separately extracts appearance and geometric features. Lightweight fusion is performed at the P3 feature level, while frequency-domain feature enhancement fusion strengthens cross-modal interaction at P4. An omnidirectional information integration attention module further refines spatial feature responses, and EdgeLAWDS (Edge-aware Lightweight Adaptive-Weight Downsampling) reduces boundary information loss during feature downsampling. Experiments were conducted on 1,264 samples from the AmodalAppleSize_RGB-D dataset using a 7:2:1 training-validation-test split. HEF-YOLO achieved mask mean average precision (mAP) values of 92.5% at an intersection-over-union threshold of 0.50 (mAP50) and 72.0% averaged over thresholds from 0.50 to 0.95 (mAP50-95), outperforming YOLO11n-seg by 1.7 and 2.1 percentage points, respectively. The results demonstrate the effectiveness of estimated-depth-guided multimodal feature learning for apple instance segmentation.

Keywords: apple instance segmentation; edge-aware downsampling; feature fusion; monocular depth estimation; RGB-D perception

1 INTRODUCTION

Apple harvesting is labour intensive and concentrated within a short seasonal window. Harvesting robots require reliable estimates of fruit location, visible contour and spatial relationships with branches, leaves and neighbouring fruit to support target selection and end-effector approach.

Traditional methods based on colour thresholds, texture, edges and circularity are sensitive to illumination, shadows, occlusion and fruit maturity. Deep learning detectors, including Faster R-CNN and YOLO, have improved orchard object recognition[1,2]. Instance segmentation methods such as Mask R-CNN additionally produce pixel-level masks that distinguish neighbouring instances and describe their visible contours [3].

Instance segmentation supports robotic perception beyond bounding-box detection. Fruit masks facilitate estimation of centres, visible areas and occlusion, while reducing contamination from neighbouring fruit during depth extraction and grasp-point selection. Kang and Chen introduced DaSNet and DaSNet-v2 for apple detection and fruit and branch segmentation, with RGB-D data supporting three-dimensional visualisation [4,5]. Wang and He improved apple detection and segmentation using Mask Scoring R-CNN [6], and Tang et al. used an improved SOLOv2 model for overlapping apple segmentation and picking-point localisation [7]. These studies connect visual recognition with subsequent robotic operations.

Most existing methods rely primarily on RGB appearance. Strong illumination, shadows and similar colours between green apples and leaves can weaken visual contrast. Independently acquired depth can complement RGB observations for fruit separation and localisation, as illustrated by multimodal Fuji apple detection [8] and the review of RGB-D sensing in orchards [9]. Several systems use depth after detection or segmentation to recover coordinates[5,7,10]. The present study addresses a different setting: an auxiliary depth representation is inferred from the RGB image itself. It therefore investigates the usefulness of a learned geometric prior for segmentation, rather than the benefit of an additional depth sensor.

Combining RGB features with predicted depth does not automatically improve segmentation. The two representations have different numerical distributions, while depth estimation can introduce errors at occlusions and boundaries. Dual encoders and multilevel interaction have been investigated in RGB-D scene segmentation[11], and UMIS-YOLO incorporates frequency-domain feature enhancement and amplitude-phase fusion [12]. These designs motivate investigating how fusion complexity should be allocated across feature scales. In the present setting, any benefit must be evaluated against RGB-only and simpler configurations because the auxiliary depth is derived from the same input image.

Downsampling is another source of boundary degradation. Strided convolution and pooling reduce resolution while increasing receptive fields, but may suppress small targets, short contour arcs and occlusion boundaries. Such information is important for separating touching apples. Local importance pooling, anti-aliasing, adaptive pooling and frequency decomposition offer alternative downsampling mechanisms [13-16], motivating explicit treatment of task-relevant edge information.

This study proposes HEF-YOLO, an RGB-D apple instance segmentation

network guided by monocular depth estimation. The proposed architecture employs an asymmetric dual-branch backbone to separately extract RGB appearance features and depth-guided geometric representations. A hierarchical fusion strategy is then introduced, using lightweight spatial fusion at the P3 level to preserve fine-grained local information and frequency-domain feature fusion at P4 to enhance cross-modal interaction in deeper semantic features. The frequency-domain fusion component is adapted from UMIS-YOLO [12] and integrated into the proposed scale-specific fusion framework. In addition, EdgeLAWDS incorporates local content weighting, Sobel edge priors, directional gradients, and edge residual connections into the P3-to-P4 downsampling path to reduce the loss of boundary and structural information. The effectiveness of the overall architecture and its individual components is evaluated through comparative and ablation experiments.

2 Related Work

2.1 Apple Detection and Instance Segmentation

Early orchard vision systems typically identified fruit using handcrafted features such as colour thresholds, region growing, morphological operations, edge information, and circular fitting. Although these approaches are computationally inexpensive, their performance is strongly affected by illumination variation, shadows, fruit maturity, and occlusion. With the development of deep learning, convolutional neural networks have substantially improved fruit recognition in complex orchard environments. Bargoti and Underwood applied deep object detection to orchard fruit [1], while the computational efficiency of the YOLO framework [2] has made single-stage detectors attractive for real-time robotic applications. However, bounding-box detection provides only coarse spatial information and cannot accurately describe fruit boundaries or separate adjacent instances.

Instance segmentation addresses this limitation by combining object recognition with pixel-level mask prediction. Kang and Chen proposed DaSNet for joint fruit detection and semantic segmentation [4], and subsequently developed DaSNet-v2 to perform lightweight fruit and branch instance segmentation with three-dimensional orchard visualisation [5]. Wang and He introduced the MS-ADS model based on Mask Scoring R-CNN to improve apple instance segmentation in natural environments [6]. Tang et al. combined an EfficientNetV2 backbone with lightweight spatial attention and SOLOv2 to improve the segmentation and localisation of overlapping apples [7]. These studies demonstrate the importance of accurate instance separation for downstream tasks such as depth extraction, target localisation, and robotic picking. Nevertheless, most existing approaches rely primarily on RGB appearance information, which remains vulnerable to weak colour contrast, shadows, and severe occlusion.

2.2 RGB and Estimated Depth Feature Fusion

Depth information provides complementary geometric cues that can

alleviate the limitations of RGB-only perception. RGB-D sensing has therefore been widely investigated for fruit detection, localisation, and robotic harvesting, although its performance depends on depth quality, sensing range, and camera calibration [9]. Gené-Mola et al. combined RGB and depth information for multimodal Fuji apple detection [8]. Kang and Chen used RGB-D data for orchard scene reconstruction and three-dimensional visualisation [5], while Hu et al. incorporated RGB-D information into an improved YOLOX framework for apple detection and localisation [10]. Tang et al. further used depth information to support picking-point localisation after instance segmentation [7].

Unlike sensor-acquired depth, monocular depth estimation derives geometric structure directly from an RGB image. Recent models such as Depth Anything V2 [17] can generate dense depth representations with strong scene-level structural consistency, providing an alternative source of geometric cues when direct depth sensing is unavailable or unreliable. However, because the predicted depth originates from the same RGB observation, it should be regarded as an auxiliary learned representation rather than an independent geometric measurement.

Effective use of such auxiliary information requires appropriate feature interaction. ESANet employs separate RGB and depth encoders with multilevel fusion for RGB-D semantic segmentation [11]. UMIS-YOLO introduces multimodal feature interaction in the frequency domain by jointly modelling amplitude and phase information [12]. These studies show that feature-level fusion can improve complementary representation learning across modalities. However, the effectiveness of transferring such fusion mechanisms to RGB and monocular-depth representations remains insufficiently explored, particularly for orchard instance segmentation where fine boundaries and occlusion relationships are critical.

2.3 Multiscale Fusion and Downsampling

Multiscale feature aggregation plays a central role in modern detection and segmentation networks. Feature Pyramid Networks (FPN) propagate high-level semantic information through top-down pathways and lateral connections [18]. PANet further introduces a bottom-up aggregation path to strengthen information flow across feature levels [19], while BiFPN employs learnable bidirectional fusion to improve multiscale feature integration [20]. These architectures enhance communication across spatial resolutions, but multimodal segmentation additionally requires effective interaction between heterogeneous representations within the same feature level.

In parallel, downsampling determines how much spatial detail is retained as feature resolution decreases. Local importance-based pooling learns adaptive spatial weights to preserve informative regions [13], anti-aliased downsampling improves shift consistency by suppressing high-frequency aliasing [14], and AdaPool dynamically adjusts pooling according to local feature content [15]. Haar wavelet downsampling decomposes features into low- and high-frequency components, allowing directional details to be retained

during resolution reduction[16]. These approaches demonstrate that conventional strided convolution or pooling can discard task-relevant fine-grained information.

For apple instance segmentation, such information loss is particularly important because fruit contours, contact boundaries, and partially visible regions often occupy only a small number of pixels. Preserving these structures during bottom-up feature aggregation is therefore essential for separating overlapping or occluded fruit. This motivates the incorporation of content-adaptive weighting and explicit edge information into the downsampling process.

3 Materials and Methods

3.1 Dataset

The experiments used the Fuji subset of the publicly available AmodalAppleSize_RGB-D dataset[21,22], which contains 3,925 RGB-D samples from six Fuji apple trees, the dataset contains apples captured at different maturity stages and under varying orchard conditions. RGB images have a resolution of 1300 × 1300 pixels.**Fig 1** illustrates representative orchard scenes.The dataset is publicly available from the CORA Research Data Repository under a Creative Commons Attribution 4.0 International (CC BY 4.0) license at https://doi.org/10.34810/DATA916.

.

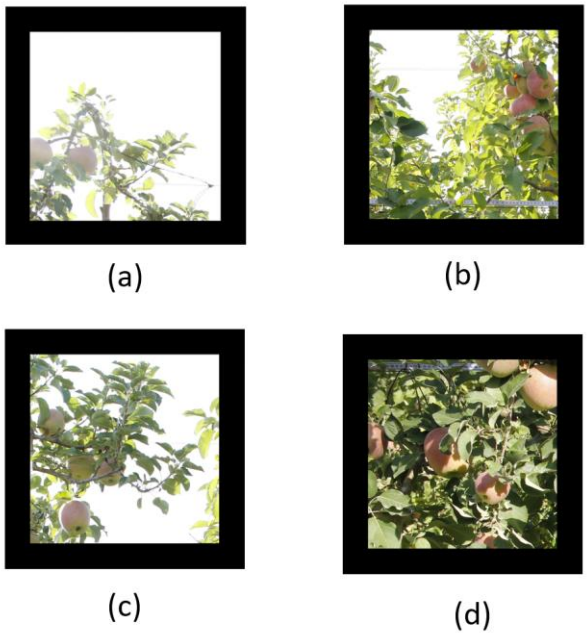

Figure 1 Dataset Samples and Challenging Scenarios from [22]:(a) Backlit scenes (b) Overlapping fruits (c) Occlusion by branches and leaves (d) Complex backgrounds Reproduced under the Creative Commons Attribution 4.0 International License (CC BY 4.0)

Depth Anything V2[17] was used to generate dense depth estimates from the RGB images. These predictions replaced the original dataset depth maps

as the auxiliary network input. Their role is to provide a learned representation of relative scene geometry for segmentation. Because both branches originate from the same RGB image, their fusion combines two representations of a common observation.

To limit the computational cost of the comparative and ablation experiments, 1,264 samples were randomly selected from the full dataset. This subset includes red and green apples, overlapping fruit, foliage occlusion and varying illumination. It was divided into training, validation and independent test sets in a 7:2:1 ratio, corresponding to approximately 70%, 20% and 10% of the selected samples. RGB images, predicted depth maps and instance annotations remained paired. The same partition was used across experiments. **Tab 2** and **3** report performance on the independent test set.

3.2 Network Architecture

HEF-YOLO comprises an asymmetric dual-branch backbone, hierarchical fusion, EdgeLAWDS downsampling and a segmentation head. First, Depth Anything V2 maps each RGB image to an estimated-depth image. The RGB image and its depth estimate are then encoded separately. Features interact through scale-specific fusion at P3 and P4, followed by spatial attention and C3k2 integration. EdgeLAWDS replaces the strided convolution in the bottom-up P3-to-P4 path; the P4-to-P5 stage retains conventional strided convolution. The segmentation head produces the instance masks. **Fig 2** shows the segmentation network after depth generation.

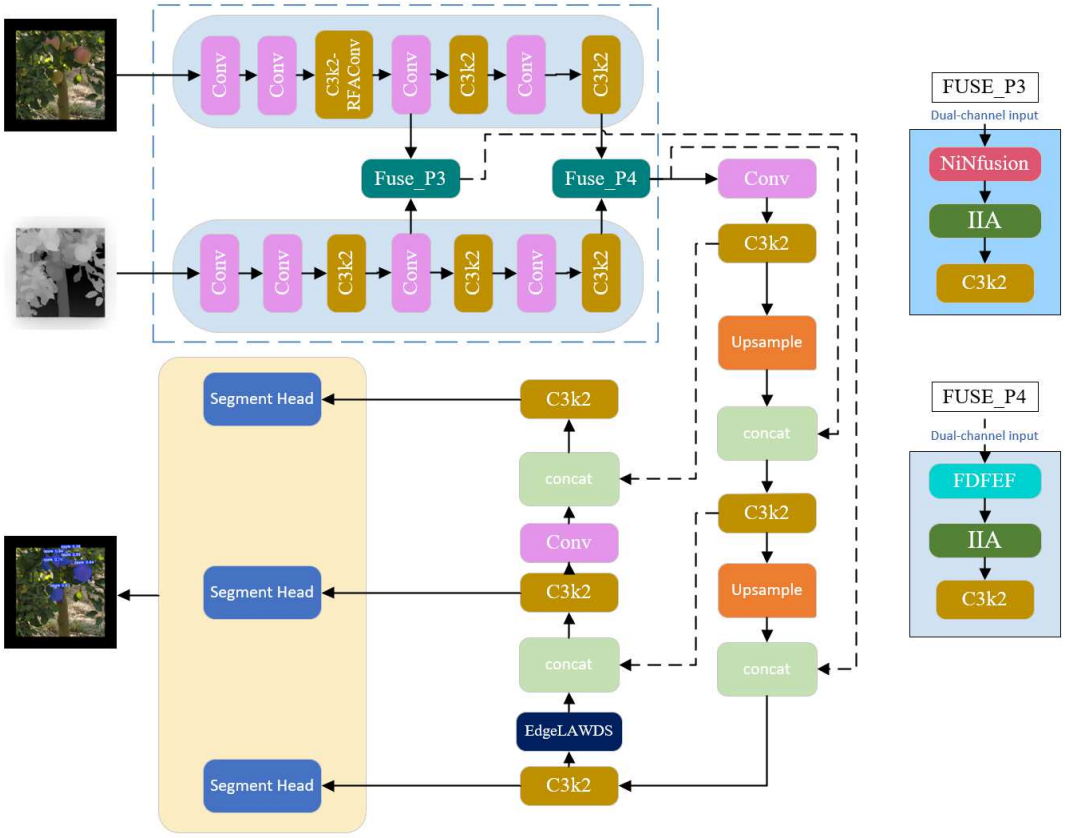

Figure 2 The overall network architecture of the HEF-YOLO model

3.3 Asymmetric Dual-Branch Feature Extraction

The backbone contains separate RGB and estimated-depth branches derived from YOLO11, as shown in **Fig 3**. RGB images retain colour and texture, while the estimated-depth input provides a learned geometric representation. The shallow RGB branch incorporates C3k2 with receptive-field attention convolution, denoted C3k2_RFAConv. The estimated-depth branch retains the conventional convolutional structure. This asymmetric design is intended to accommodate the different representations before fusion.

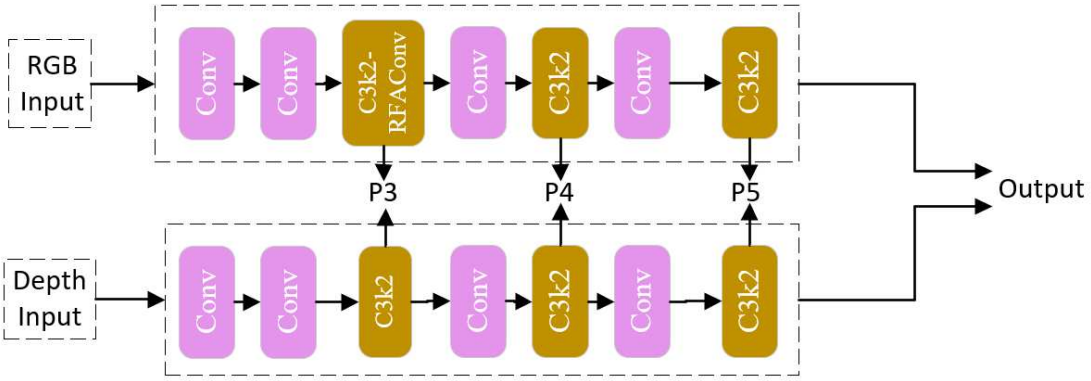

Figure 3 Asymmetric dual-branch feature extraction backbone network structure

3.4 Hierarchical Fusion of RGB and Estimated Depth

P3 retains relatively high spatial resolution, whereas P4 has a lower-resolution representation. Network-in-Network Fusion (NiNfusion) is used at P3, and frequency-domain feature enhancement fusion (FDFEF), adapted from [12], is used at P4. Both fusion outputs undergo omnidirectional information integration attention (IIA) and C3k2 processing. This configuration is evaluated as a combined fusion design; the individual contributions of its operators are not isolated by the present ablation table.

3.4.1Lightweight Fusion at P3

NiNfusion concatenates RGB and depth features along the channel dimension, applies a bias-free convolutional projection and uses SiLU activation for nonlinear mapping (**Fig 4**). Let F_R^3 and F_D^3 denote the P3 features. Channel concatenation is defined in Eq.(1), and the projected feature is given in Eq.(2).

$$F_C^3 = Concat(F_R^3, F_D^3) \tag{1}$$

$$F_N^3 = SiLU(W_N * F_C^3) \tag{2}$$

Here, Concat denotes channel-wise concatenation, W_N denotes the projection parameters and the asterisk denotes convolution. The supplied implementation uses a 1 × 1 kernel, unit stride and one convolution group by default, with no bias or batch-normalisation layer. The convolution input-channel count equals the sum of the two branch channel counts. Under these default settings, the operation preserves the spatial dimensions of P3 and limits the cost of intermediate-resolution fusion.

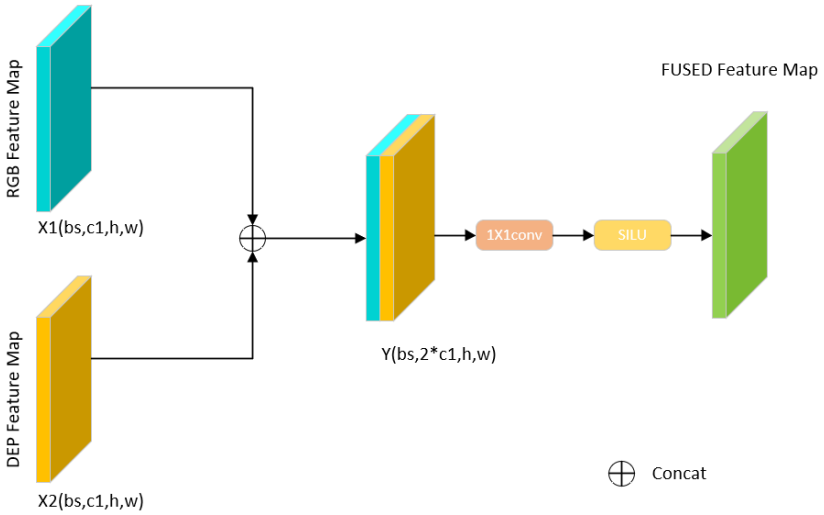

Figure 4 Architecture of the Network-in-Network Fusion

3.4.2Frequency-Domain Fusion at P4

FDFEF consists of modality-specific frequency enhancement followed by amplitude–phase fusion, following the approach in [12].**Fig 5** outlines this process. For input feature X_i, where i identifies the RGB or estimated-depth representation, an orthonormal real two-dimensional fast Fourier transform is applied as in Eq.(3).

$$\mathrm{F}_i = \mathrm{FFT2}(X_i), \mathrm{F}_i \in \pounds^{C\times H\times\left(\left\lfloor \frac{W}{2}\right\rfloor +1\right)} \tag{3}$$

The transform uses orthonormal scaling. Learnable real and imaginary parameters form the complex spectral weight $W^{(i)}_{\text{complex}}$ Element-wise modulation, inverse transformation, an input residual and the DyT layer produce the enhanced feature in Eq.(4).

$$\hat{X}_i = \mathrm{DyT}\left(\mathrm{irFFT2}\left(\mathrm{F}_i \, e \, W^{(i)}_{\text{complex}}\right) + X_i\right) \tag{4}$$

The enhanced feature is transformed again, giving $\hat{\mathrm{F}}_i$ .Its amplitude and phase are explicitly separated in Eq.(5). Amplitude describes the distribution of spectral energy, while phase contributes spatial and structural information.

$$A_i = \left|\hat{\mathrm{F}}_i\right|, \quad \varphi_i = \mathrm{angle}\left(\hat{\mathrm{F}}_i\right), \quad i \in \{\mathrm{R}, \mathrm{D}\} \tag{5}$$

Learnable amplitude and phase coefficients are normalised across modalities by Softmax. The fused amplitude is calculated using Eq.(6), where s denotes the Softmax operation and α denotes its learnable amplitude parameters.

$$A_{\text{fused}} = s(\alpha)_1 A_R + s(\alpha)_2 A_D \tag{6}$$

Phase is periodic. Direct averaging of angles near the negative and positive pi boundary can introduce discontinuities. The two phases are therefore mapped to unit complex vectors, combined using normalised weights and converted back to an angle, as in Eq.(7). Here, β contains the learnable phase parameters and j is the imaginary unit.

$$\varphi_{\text{fused}} = \text{angle}\left(s(\beta)_1 e^{j\varphi_R} + s(\beta)_2 e^{j\varphi_D}\right) \tag{7}$$

The fused spectrum is reconstructed from its amplitude and phase, transformed back to the spatial domain and processed by an output DyT layer, giving Eq.(8). Applying frequency-domain fusion at the lower-resolution P4 level limits its computational burden while allowing interaction between deeper RGB and depth features.

$$Y_{P4} = \text{DyT}\left(\text{irFFT2}\left(A_{\text{fused}} \text{ e } e^{j\varphi\text{fused}}\right)\right) \tag{8}$$

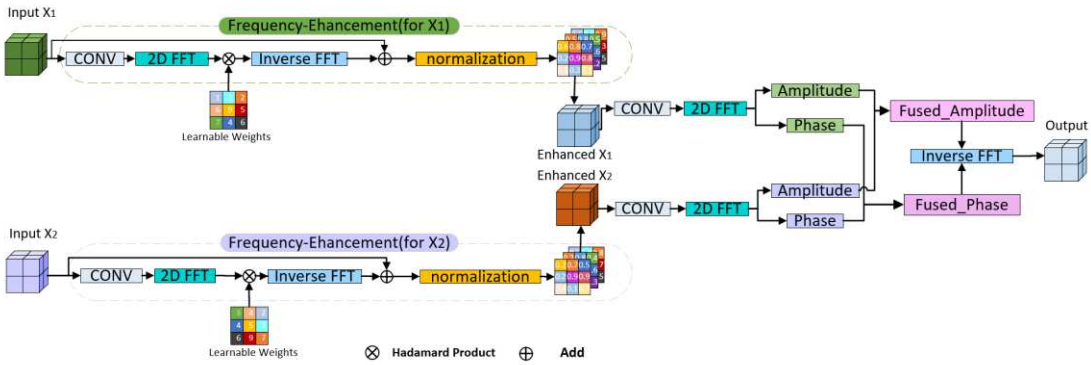

Figure 5 Architecture of the frequency-domain feature enhancement fusion (FDFEF) module

3.4.3 Omnidirectional Spatial Attention

The IIA module applies spatial attention to the fused features through horizontal, vertical, local two-dimensional and omnidirectional two-dimensional branches (**Fig 6**). The different spatial supports are intended to represent boundaries with different orientations. Their benefit for specific occlusion or contour conditions is not separately quantified in the present experiments.

Channel-wise average and maximum pooling of the input feature F are concatenated in Eq.(9) to form a two-channel spatial descriptor P.

$$P = \text{Concat}\left(\text{AvgPool}_c(F), \text{MaxPool}_c(F)\right) \tag{9}$$

The descriptor is processed by four convolutional branches as specified in Eq.(10),(11). The kernel sizes are k for the larger directional and omnidirectional operations and k_l for the local operation. These kernel sizes determine the spatial support of the corresponding branches.

$$A_h = \text{Conv}_{1\times k}(P), \quad A_v = \text{Conv}_{k\times 1}(P) \tag{10}$$

$$A_l = \text{Conv}_{k_l\times k_l}(P), \quad A_o = \text{Conv}_{k\times k}(P) \tag{11}$$

The four responses are concatenated and passed through a 1 × 1 convolution, batch normalisation and a sigmoid function, producing spatial attention A in Eq.(12). In this equation, σ denotes the sigmoid function rather than Softmax.

$$A = \sigma\left(\mathrm{BN}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Concat}(A_h, A_v, A_l, A_o)\right)\right)\right) \tag{12}$$

The output in Eq.(13) uses residual attention with a learnable scalar γ, initialised to zero. The symbol e denotes element-wise multiplication.

$$\hat{F} = F + \gamma(F \mathrm{e}\ A) \tag{13}$$

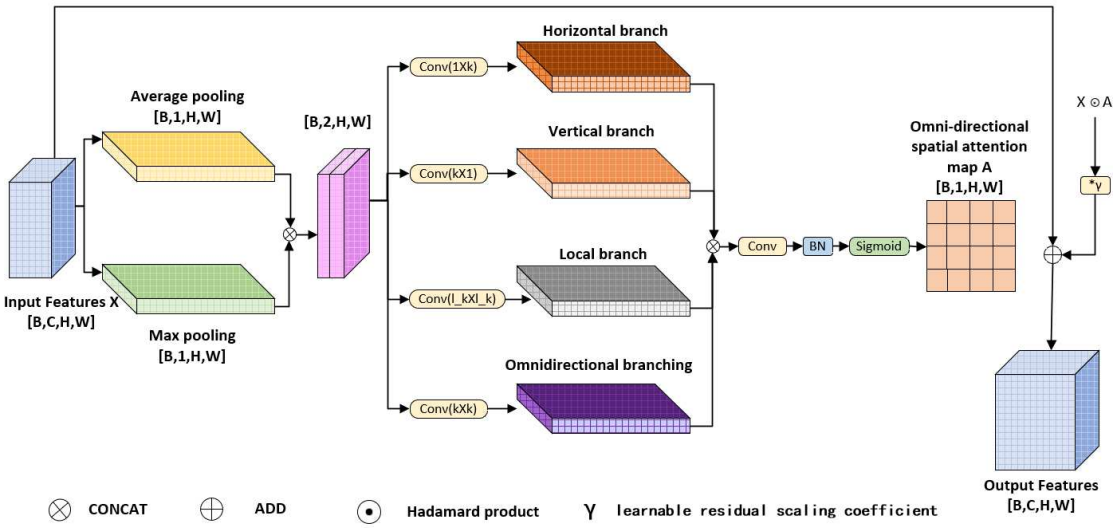

Figure 6 Omnidirectional information integration attention. Four spatial branches form one attention map, followed by residual feature weighting

3.5 Edge-Guided Locally Adaptive Weighted Downsampling

Downsampling reduces spatial resolution and computational cost, but can weaken the short contour segments needed to separate occluded or touching fruit. EdgeLAWDS is introduced only in the bottom-up P3-to-P4 path. The P4-to-P5 path remains unchanged to avoid repeating the additional computation throughout the network.**Fig 7** shows the local content branch, edge-prior branch, adaptive candidate aggregation and edge residual.

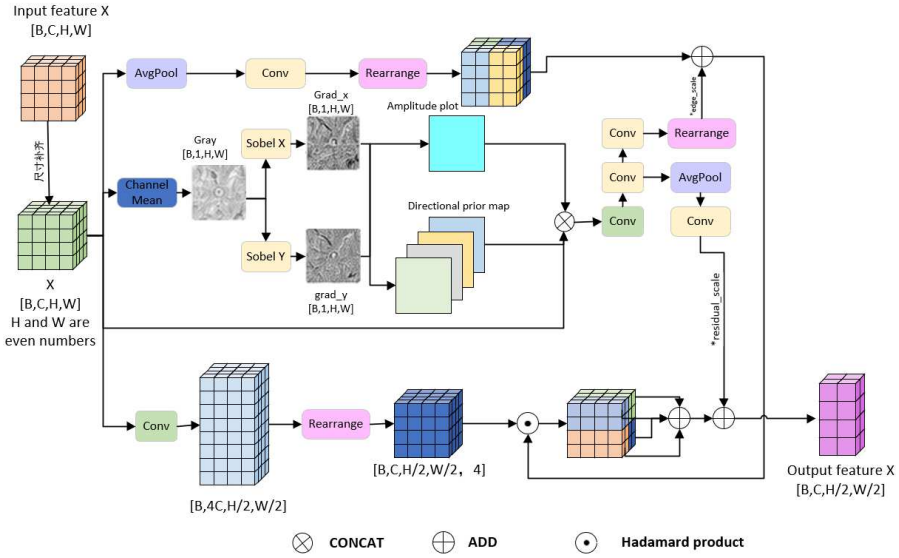

Figure 7 EdgeLAWDS processing flow. Content and edge responses predict four-

position weights; an edge residual supplements the weighted downsampling output.

3.5.1 Edge Priors

For input X ,channel averaging produces the spatial descriptor G Horizontal and vertical Sobel convolution gives the gradient responses in Eq.(14), with the kernels specified in Eq.(15).

$$G_x = G * K_x, \quad G_y = G * K_y \tag{14}$$

$$K_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix}, \quad K_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{15}$$

A small positive smoothing term ò stabilises the gradient magnitude in Eq.(16). Four directional descriptors are formed from the absolute horizontal and vertical gradients and their sum and difference, as shown in Eq.(17).

$$M = \sqrt{G_x^2 + G_y^2 + \text{ò}} \tag{16}$$

$$O = \text{Concat}\left(|G_x|, |G_y|, |G_x - G_y|, |G_x + G_y|\right) \tag{17}$$

The magnitude channel and four directional channels are concatenated with the input feature and encoded as an edge prior. The directional descriptors provide gradient responses beyond the horizontal and vertical axes. This construction is intended to support boundary-sensitive downsampling; the resulting boundary accuracy requires evaluation in addition to overall mask mAP.

3.5.2 Content and Edge-Guided Weights

The local content branch applies 3 × 3 average pooling and convolution, then rearranges spatial responses into four candidate weights for each output location. The edge encoder processes the original feature, gradient magnitude and directional descriptors to generate a corresponding edge bias. The two responses determine the sampling weights in Eq.(18), where A_L and A_E are the local content and edge responses and λ_e is a learnable scale. Softmax acts over the four candidate positions.

$$A = \text{Softmax}\left(A_L + \lambda_e A_E\right) \tag{18}$$

3.5.3 Adaptive Aggregation and Edge Residual

A grouped strided convolution generates four candidate feature groups, rearranged into Z_i for each output position. Eq.(19) aggregates these candidates using the adaptive weights A_i . Their contributions therefore depend on local content and the edge response.

$$Y_L = \sum_{i=1}^{4} A_i \text{e} \ Z_i \tag{19}$$

The encoded edge feature is also downsampled and added as a residual, as in Eq.(20). Here, R_E is the edge residual, λ_r is a learnable coefficient and Proj maps the output channels. The residual coefficient is initialised to a small value so that adaptive aggregation dominates at the start of training. Together,

weighting and residual compensation seek to retain boundary information for subsequent multiscale processing.

$$Y = \mathrm{Proj}\left(Y_L + \lambda_r R_E\right) \tag{20}$$

4 RESULTS AND DISCUSSION

4.1 Evaluation Metrics

Mask mAP at an intersection-over-union (IoU) threshold of 0.5 and mask mAP averaged over thresholds from 0.5 to 0.95 are the principal segmentation metrics. IoU is calculated between a predicted instance mask M_{p} and its ground-truth mask using Eq.(21), where vertical bars denote pixel cardinality.

$$\mathrm{IoU} = \frac{\left|M_p \cap M_g\right|}{\left|M_p \cup M_g\right|} \tag{21}$$

Average precision (AP) summarises the precision–recall relationship over confidence thresholds, and mean average precision (mAP) averages AP across classes. The first metric, denoted mask mAP50 in the tables, uses IoU = 0.5. The second, mask mAP50–95, averages results over ten IoU thresholds at intervals of 0.05. It imposes stricter requirements on mask overlap and therefore complements the single-threshold measure. All tabulated mAP values are reported as percentages.

4.2 Experimental Settings

Experiments were conducted with an NVIDIA RTX 3090 GPU with 24 GB of memory and an Intel Xeon Silver 4210R processor. The software environment comprised Ubuntu 20.04, Python 3.10, PyTorch 2.2.2 and CUDA 12.1. All models used the same 7:2:1 training-validation-test partition and the same mask evaluation protocol. Ablation configurations used the same training settings. **Tab 1** lists the reported hyperparameters. The accuracy values in **Tab 2** and **3** are single-run results on the independent test set; repeated-run uncertainty estimates are not available.

Table 1 Reported training hyperparameters

Hyperparameter	Value
Training epochs	300
Input size / pixel	640 × 640
Batch size	16
Optimizer	AdamW
Initial learning rate	0.01
Weight decay	0.0005
Mask ratio	4
Activation function	SiLU

4.3 Ablation Study

Six configurations, A0-A5, were evaluated, including the YOLO11n-seg baseline. In **Tab 2**, Depth denotes the estimated-depth input together with its

feature-processing branch; RFAConv denotes C3k2_RFAConv in the RGB branch; Fusion denotes the hierarchical fusion configuration, including IIA; and EdgeLAWDS denotes its use in the P3-to-P4 path.

Table 2 Ablation results for YOLO11n-seg and the proposed components

Model	Depth	RFAconv	Fusion	EdgeLAWDs	Mask mAP50/%	Mask mAP50-95/%
A0	-	-	-	-	90.8	69.9
A1	yes	-	-	-	91.6	71.6
A2	yes	yes	-	-	91.8	71.8
A3	yes	yes	yes	-	92.1	71.5
A4	yes	yes	-	yes	91.6	71.7
A5	yes	yes	yes	yes	92.5	72.0

4.3.1 Depth Input and RGB Feature Extraction

Introducing the predicted-depth branch in A1 increases mask mAP50 from 90.8% to 91.6% and mask mAP50-95 from 69.9% to 71.6%, corresponding to gains of 0.8 and 1.7 percentage points over A0, respectively. The larger improvement in mAP50-95 shows that depth information mainly contributes to more accurate mask delineation under stricter overlap criteria. By supplementing RGB appearance cues with geometric information, the additional branch improves feature discrimination in scenes involving occlusion, overlap, and appearance ambiguity.

A2 further incorporates RFAConv and achieves mask mAP50 and mask mAP50-95 of 91.8% and 71.8%, respectively. Although the gains over A1 are limited to 0.2 percentage points for both metrics, the consistent improvement reflects enhanced local spatial representation. RFAConv therefore mainly acts as a feature refinement component rather than a dominant source of accuracy improvement.

4.3.2 Hierarchical Fusion

Adding hierarchical fusion to A2 produces A3, for which mask mAP50 increases from 91.8% to 92.1%, while mask mAP50-95 decreases slightly from 71.8% to 71.5%. The improvement at IoU = 0.5 indicates that hierarchical fusion strengthens the interaction between RGB appearance features and depth-derived geometric features, thereby improving instance discrimination at the target level.

The slight reduction in mAP50-95, however, shows that cross-modal fusion alone is insufficient to improve fine-grained mask agreement across stricter IoU thresholds. This result suggests that hierarchical fusion primarily enhances semantic and cross-modal representation, whereas boundary-sensitive information requires additional preservation during feature downsampling.

4.3.3 EdgeLAWDS and the Combined Architecture

A4 evaluates EdgeLAWDS independently on the basis of A2. The resulting mask mAP50 and mask mAP50-95 are 91.6% and 71.7%, respectively, slightly below those of A2. Therefore, EdgeLAWDS does not provide an isolated

improvement in the overall segmentation metrics under this configuration. Its contribution is instead related to preserving local edge responses and spatial details during downsampling, which may not be fully reflected when the module is used without enhanced cross-modal feature interaction.

When EdgeLAWDS is combined with hierarchical fusion in A5, the complete HEF-YOLO model achieves the best performance among all configurations, with mask mAP50 and mask mAP50-95 reaching 92.5% and 72.0%, respectively. Compared with A3, the improvements are 0.4 and 0.5 percentage points, while relative to A2 the combined configuration gains 0.7 and 0.2 percentage points.

The performance of A5 shows that hierarchical fusion and EdgeLAWDS provide complementary functions. Hierarchical fusion enhances the interaction between appearance and geometric information, whereas EdgeLAWDS reduces the loss of boundary and local structural cues during resolution reduction. Their combination enables the fused multimodal features to retain both semantic discrimination and spatial detail.

Overall, A5 outperforms the baseline A0 by 1.7 percentage points in mask mAP50 and 2.1 percentage points in mask mAP50-95. Among the evaluated modifications, the predicted-depth branch contributes the largest individual gain, confirming the importance of geometric information for apple instance segmentation. RFAConv, hierarchical fusion, and EdgeLAWDS subsequently refine local representation, cross-modal interaction, and boundary preservation, respectively, resulting in the highest overall segmentation accuracy.

4.4 Comparison with Other Models

Tab 3 compares HEF-YOLO with several YOLO-based segmentation models and SBA-YOLO [23] on the same apple test set. SBA-YOLO, although originally developed for clothing instance segmentation, was retrained and evaluated using the same apple dataset partition to ensure a consistent comparison.

HEF-YOLO achieves the highest mask mAP50 and mask mAP50-95 among the evaluated models, reaching 92.5% and 72.0%, respectively. Compared with YOLOv8n-seg, YOLOv12n-seg, and YOLOv13n-seg, the corresponding gains in mask mAP50 are 1.6, 1.6, and 2.3 percentage points, while the gains in mask mAP50-95 are 2.1, 3.0, and 5.0 percentage points. These results confirm the effectiveness of incorporating depth-guided geometric information and hierarchical heterogeneous fusion into the segmentation framework.

HEF-YOLO contains 3.45 million parameters and requires 14.6 GFLOPs, with an inference speed of 52.5 FPS. Relative to YOLO11n-seg, the parameter count and computational cost increase by approximately 19.8% and 41.7%, respectively, while the inference speed decreases from 112.4 to 52.5 FPS. The additional computation is mainly introduced by the dual-branch architecture and multimodal fusion operations.

Despite this increase in complexity, the model retains real-time inference

capability while achieving higher segmentation accuracy. Compared with SBA-YOLO, HEF-YOLO provides higher mask accuracy with fewer parameters and lower FLOPs, although its reported throughput is lower. Overall, the proposed architecture achieves a favorable trade-off between accuracy and computational cost for RGB-D apple instance segmentation in complex orchard environments.

Table 3 Comparison of model complexity,throughput and mask accuracy

Model	Parameters/mollion	GFLOPs	FPS	Mask mAP50/%	Mask mAP50-95/%
YOLOv5n-seg	1.88	6.7	156.1	89.4	63.5
YOLOv8n-seg	3.26	11.9	141.6	90.9	69.9
SBA-YOLO	7.89	22.6	87.6	90.6	69.7
YOLOv12n-seg	2.77	9.7	63.2	90.9	69.0
YOLOv13n-seg	2.71	10.1	43.8	90.2	67.0
YOLO11n-seg	2.88	10.3	112.4	90.8	69.9
HEF-YOLO	3.45	14.6	52.5	92.5	72.0

5 CONCLUSION

This study proposed HEF-YOLO, an RGB-D multimodal instance segmentation framework for apple perception in complex orchard environments. The model employs an asymmetric dual-branch backbone to extract complementary appearance and depth-guided geometric features. RFAConv enhances local spatial representation, hierarchical fusion promotes cross-modal feature interaction at different feature scales, and EdgeLAWDS preserves boundary and structural information during downsampling.

Experiments were conducted on a 1,264-sample subset of the AmodalAppleSize_RGB-D dataset using a 7:2:1 training-validation-test split. HEF-YOLO achieved a mask mAP50 of 92.5% and a mask mAP50-95 of 72.0%, outperforming YOLO11n-seg by 1.7 and 2.1 percentage points, respectively. The ablation results show that the predicted-depth branch provides the largest individual improvement, demonstrating the value of complementary geometric information for apple segmentation. RFAConv further refines local feature representation, while hierarchical fusion and EdgeLAWDS provide complementary benefits in cross-modal interaction and boundary preservation. Their combined use yields the best overall performance among the evaluated configurations.

HEF-YOLO contains 3.45 million parameters and requires 14.6 GFLOPs, while maintaining an inference speed of 52.5 FPS. Although the multimodal architecture introduces additional computational cost compared with YOLO11n-seg, the model retains real-time processing capability and provides improved segmentation accuracy, resulting in a favorable accuracy-complexity trade-off for robotic apple perception.

The present evaluation is limited to a single dataset subset and one train-validation-test partition, and the reported results are based on single-run

training. In addition, the auxiliary depth information is estimated from RGB images and has not been validated as metric depth. Future work will therefore include repeated experiments and cross-orchard validation, direct comparison between predicted and sensor-acquired depth, more detailed evaluation of occluded and boundary regions, and end-to-end efficiency analysis including depth generation. The proposed perception framework will also be integrated with apple localization and robotic harvesting experiments to further evaluate its applicability under real orchard conditions.

Author contributions

Tao Zhang:funding,discovery,Jiaxuan Xia: manuscript，implementation，technique,validation, Jinxing Niu: academic paper insights, Chang Liu Yuhang Wang Xingyue Xu Dingyi Ma Yingchun Deng :advancement，implementation.

Funding

This study was supported by the Key R&D Special Project of Henan Province (Project No. 231111112700) and the Key Science and Technology Project of Henan Province (Project No. 262102110383).

Data availability

The dataset analysed during the current study is the Fuji subset of the publicly available AmodalAppleSize_RGB-D dataset, available from the CORA Research Data Repository at https://doi.org/10.34810/DATA916.The dataset is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license. The processed data generated during the current study are available from the corresponding author upon reasonable request.

Materials availability

Not applicable.

Code availability

The code developed and used during the current study is available from the corresponding author upon reasonable request.

Declarations

Ethical approval and accordance statement

This study did not involve human participants, human data, human tissue, or animals. Therefore, ethical approval was notrequired. All methods were carried out in accordance with relevant institutional guidelines and applicable regulations.

Consent to participate

Not applicable. This study did not involve human participants.

Consent to publish

Not applicable. This manuscript does not contain any identifiable personal data requiring consent for publication.

Competing interests

The authors declare no competing interests.

REFERENCES

[1] Bargoti, S., & Underwood, J. (2017, May). Deep fruit detection in orchards. In 2017 IEEE international conference on robotics and automation (ICRA) (pp. 3626-3633). IEEE.

[2] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).

[3] He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969).

[4] Kang, H., & Chen, C. (2019). Fruit detection and segmentation for apple harvesting using visual sensor in orchards. Sensors, 19(20), 4599.

[5] Kang, H., & Chen, C. (2020). Fruit detection, segmentation and 3D visualisation of environments in apple orchards. Computers and Electronics in Agriculture, 171, 105302.

[6] Wang, D., & He, D. (2022). Apple detection and instance segmentation in natural environments using an improved Mask Scoring R-CNN Model. Frontiers in plant science, 13, 1016470.

[7] Tang, S., Xia, Z., Gu, J., Wang, W., Huang, Z., & Zhang, W. (2024). High-precision apple recognition and localization method based on RGB-D and improved SOLOv2 instance segmentation. Frontiers in Sustainable Food Systems, 8, 1403872.

[8] Gené-Mola, J., Vilaplana, V., Rosell-Polo, J. R., Morros, J. R., Ruiz-Hidalgo, J., & Gregorio, E. (2019). Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. Computers and Electronics in Agriculture, 162, 689-698.

[9] Fu, L., Gao, F., Wu, J., Li, R., Karkee, M., & Zhang, Q. (2020). Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. Computers and Electronics in Agriculture, 177, 105687.

[10] Hu, T., Wang, W., Gu, J., Xia, Z., Zhang, J., & Wang, B. (2023). Research on apple object detection and localization method based on improved YOLOX and RGB-D images. Agronomy, 13(7), 1816.

[11] Seichter, D., Köhler, M., Lewandowski, B., Wengefeld, T., & Gross, H. M. (2021, May). Efficient rgb-d semantic segmentation for indoor scene analysis. In 2021 IEEE international conference on robotics and automation (ICRA) (pp. 13525-13531). IEEE.

[12] Yang, Y., Feng, X., Li, M., Hu, X., Qin, J., Gruen, A., ... & Gong, J. (2025). UMIS-YOLO: Underwater multimodal images instance segmentation with YOLO. IEEE Transactions on Geoscience and Remote Sensing.

[13] Gao, Z., Wang, L., & Wu, G. (2019, October). Lip: Local importance-based pooling. In 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (pp. 3354-3363). IEEE.

[14] Zhang, R. (2019, May). Making convolutional networks shift-invariant again. In International conference on machine learning (pp. 7324-7334).

PMLR.
[15] Stergiou, A., & Poppe, R. (2022). Adapool: Exponential adaptive pooling for information-retaining downsampling. IEEE Transactions on Image Processing, 32, 251-266.
[16] Xu, G., Liao, W., Zhang, X., Li, C., He, X., & Wu, X. (2023). Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. Pattern recognition, 143, 109819.
[17] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., & Zhao, H. (2024). Depth anything v2. Advances in neural information processing systems, 37, 21875-21911.
[18] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017, July). Feature pyramid networks for object detection. In 2017 IEEE conference on computer vision and pattern recognition (CVPR) (pp. 936-944). Ieee.
[19] Liu, S., Qi, L., Qin, H., Shi, J., & Jia, J. (2018, June). Path aggregation network for instance segmentation. In 2018 IEEE/CVF conference on computer vision and pattern recognition (pp. 8759-8768). Ieee.
[20] Tan, M., Pang, R., & Le, Q. V. (2020, June). Efficientdet: Scalable and efficient object detection. In 2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR) (pp. 10778-10787). IEEE.
[21] Gené-Mola, J., Ferrer-Ferrer, M., Gregorio, E., Blok, P. M., Hemming, J., Morros, J. R., ... & Ruiz-Hidalgo, J. (2023). Looking behind occlusions: A study on amodal segmentation for robust on-tree apple fruit size estimation. Computers and Electronics in Agriculture, 209, 107854.
[22] Gené-Mola J, Ferrer-Ferrer M, Hemming J, van Dalfsen P, de Hoog D, Sanz-Cortiella R, Rosell-Polo JR, Morros JR, Vilaplana V, Ruiz-Hidalgo J, Gregorio E. AmodalAppleSize_RGB-D dataset: RGB-D images of apple trees annotated with modal and amodal segmentation masks for fruit detection, visibility and size estimation. *Data in Brief*. 2024;52:110000. https://doi.org/10.1016/j.dib.2023.110000
[23] Gu, M., Yao, G., Dong, X., & Pan, Y. (2026). Multi-scale clothing image instance segmentation method based on improved YOLOv8-seg. Pattern Analysis and Applications, 29(1), 14.