BiGCFormer: A Unified Multimodal Framework for Robust Raspberry Detection in Complex Agricultural Scenes

Qian Wu[1,2,3,4], Heming Huang[1,3,4*], Dongke Song[1,3,4], Jie Zhou[1,3,4]

[1]The School of Computer, Qinghai Normal University, Xining 810008, China
[2]Shandong Facility Horticulture Bioengineering Research Center, Weifang University of Science and Technology, Weifang 262700, China
[3]The State Key Laboratory of Tibetan Intelligence, Xining 810008, China
[4]The Ministry of Education Key Laboratory of Tibetan Information Processing, Xining 810008, China

Keywords: BiGCFormer; YOLO; Raspberry detection; Bi-Gate; Bi-CA

Citation: Wu Q., Huang H.M., Song D.K., et al.: BiGCFormer: A Unified Multimodal Framework for Robust Raspberry Detection in Complex Agricultural Scenes. Data Intelligence, 2026, Vol. 8, Art. No.: 20250304. DOI: https://doi.org/10.3724/2096-7004.di.2025.0304

ABSTRACT

High-precision fruit recognition and localization in complex agricultural settings are a key element for intelligent orchard management and precise harvesting. However, traditional unimodal detection is constrained in meeting robustness and accuracy needs due to natural interferences like uneven illumination and foliage occlusion. Therefore, *BiGCFormer,* a *Bi-stream Graph Cross-attention Fusion Detection Network,* is proposed. It is a fusion-driven multimodal raspberry detection system integrating multi-source visual perception and deep learning, thereby, enhancing detection stability in complex scenes. Building upon the YOLOv11 architecture, it incorporates dual-stream input backbone networks for the respective processing of RGB and depth modal features, and there are three key technological innovations: (1) spatial-semantic structural alignment capabilities between RGB and depth features are strengthened by a modal-aware positional encoding module; (2) complementary enhancement pathways between RGB-D (depth) modalities are established by a bidirectional cross-attention mechanism to extract multimodal information interaction features; and (3) dynamic adjustment of fusion intensity is implemented by a dual-gated residual fusion mechanism through learnable gating functions, with robustness against modal heterogeneity and noise interference enhanced. Experimental evaluations on raspberry fruit detection datasets demonstrate that the proposed *BiGCFormer* model achieves a

*Corresponding author: Heming Huang (E-mail: huanghm@qhnu.edu.cn; ORCID: 0000-0002-6339-7135).

© 2026 Chinese Academy of Sciences. Published under a Creative Commons Attribution 4.0 International (CC BY 4.0) license.

mean Average Precision at IoU threshold 0.5 (mAP@.5) of 70.9%, representing a 2.1 percentage point improvement over the Fusion-Mamba (FMB) methodology, with comprehensive enhancements across key metrics such as precision and recall. Compared with existing multimodal detection models, the proposed *BiGCFormer* exhibits superior detection stability and generalization capabilities under varying fruit sizes, poses, and occlusion conditions.

1. INTRODUCTION

Intelligent fruit detection is recognized as a core task in smart agriculture applications which encompasses domains such as automated harvesting, yield assessment, and precision management, all of which exert a direct influence on crop collection efficiency and agricultural product quality. In practical orchard environments, fruits typically exist against unstructured backgrounds and are confronted with multiple challenges, including severe occlusion, uneven illumination, color similarity between fruits and backgrounds, and significant scale variations. Traditional unimodal object detection methods based on RGB images are rendered inadequately by these conditions for achieving stable and high-precision fruit recognition. Although certain effectiveness has been demonstrated by deep learning models (such as the YOLO series, Faster R-CNN, and SSD) in agricultural vision tasks in recent years [1-5], these models still exhibit insufficient generalization capabilities under complex environments. Notably, significant accuracy degradation is observed in scenarios involving multi-fruit occlusion and small-target detection.

To enhance perceptual robustness and recognition accuracy, attempts have been made by a growing number of researchers to introduce multimodal fusion mechanisms, with complementary characteristics between RGB images and depth maps leveraged to enhance feature representation capabilities [6-10]. Color and texture details are excellently captured by RGB images while spatial geometric and structural information are provided by depth maps, this makes depth maps particularly suitable for addressing issues of inter-fruit occlusion and background interference. However, limitations continue to be exhibited by current mainstream RGB-D fusion methods, including simplistic fusion strategies, the lack of modal-aware mechanisms, and unidirectional static information interaction channels [11-14]. Feature concatenation or linear weighting is commonly employed by existing approaches, which fail to model deep semantic complementary relationships between modalities, exhibit poor robustness when confronting modal degradation, and struggle to adapt to variations in modal dominance across different scenarios.

In recent years, methods including IGNet, FMB, and CDDFuse have been developed, with attention mechanisms or feature selection modules introduced to attempt to enhance the quality of multimodal fusion [15-20]. However, relatively coarse fusion granularity, unidirectional interaction patterns, and limited capabilities of suppressing information redundancy are maintained by these approaches. Notably, there remains room for improvement in scenarios involving dense fruit occlusion and small target overlapping.

To address these challenges, a novel dual-backbone multimodal Raspberry detection model, ***BiGCFormer***, is proposed. The core innovation of this model lies in the establishment of a unified theoretical framework for multimodal geometric interaction, achieved through three key technical

innovations: (1) the introduction of a modal-aware positional encoding module which enables explicit modeling of spatial perception deviations between RGB and Depth modalities; (2) the design of bidirectional cross-attention mechanisms which construct bidirectional dynamic information flows across modalities to enhance the accuracy of semantic alignment and the efficacy of contextual modeling; and (3) the proposal of dual-gated residual fusion modules which integrate learnable gating with residual mechanisms to explicitly regulate modal contributions during feature fusion. The detection robustness of ***BiGCFormer*** is thereby effectively improved in scenarios involving modal degradation or noise interference.

Compared with existing methods, three distinct advantages are demonstrated by ***BiGCFormer***: (1) structural support for modal decoupled modeling and semantic-level interaction while avoiding the accumulation of information redundancy; (2) possession of fusion strategies with dynamic response capabilities, enabling adaptive importance weighting across modalities; and (3) construction of efficient detection backbones based on YOLOv11, balancing detection speed and accuracy to maintain practical feasibility in embedded deployment scenarios.

Systematic evaluations are conducted on constructed multimodal fruit detection datasets, with results indicating that an mAP@.5 of 70.9% is achieved by ***BiGCFormer***, representing a 2.1 percentage point improvement over FMB methods and a 3.1 percentage point improvement over YOLOv11 + Two-Stream. Under various conditions of modal quality degradation and occlusion, smaller reductions in detection accuracy are observed compared with existing methods [21-25], validating the robustness and adaptability of the proposed fusion mechanism.

The ***BiGCFormer*** framework is centered on structural optimization, supports semantic interaction, and targets robust fusion. It provides multimodal solutions with engineering applicability and research promotion value for fruit detection in complex natural environments. This framework can be extended to multimodal object recognition tasks, including those in forestry monitoring, intelligent breeding, and security inspection applications [26-30].

The main contributions are as follows: (1) a modal-aware positional encoding module is employed to enhance spatial-semantic structural alignment between RGB and Depth features; (2) a bidirectional cross-attention mechanism is introduced to establish complementary enhancement pathways between RGB-D modalities, facilitating the extraction of multimodal interaction features; (3) a dual-gated residual fusion mechanism is used to dynamically adjust fusion intensity via learnable gating functions, thereby enhancing robustness against modal heterogeneity and noise interference.

2. RELATED WORKS

2.1 RGB-D Fusion Object Detection

RGB-D fusion has been recognized as a significant approach for enhancing the accuracy and robustness of object detection, particularly in complex natural scenarios. In such scenarios, depth information is

effective at compensating for the perceptual deficiencies of RGB images, including those caused by illumination variations, background interference, and occlusion regions [6-7, 10].

For early methods, RGB and Depth fusion was achieved by approaches such as FuseNet through simple feature concatenation; two parallel channels were utilized by SubCNN to extract RGB and depth features, followed by decision-level fusion. However, these approaches exhibit high sensitivity to modal quality and only enable relatively shallow information interaction between modalities.

To enhance fusion effectiveness, recent research has focused on cross-modal attention mechanisms and multi-level feature fusion strategies. For example, channel difference-driven fusion modules are designed by CDDFuse [18] to improve modal selectivity; guided attention is employed by IGNet for semantic enhancement; and a three-branch structure is constructed by FMB to separately process RGB, Depth, and joint information [15, 28]. Although fusion representation capabilities are enhanced by these methods, static designs in fusion architectures are maintained by most of them, making it difficult to address issues of modal imbalance and scene mutation [19-20].

Additionally, explicit conversion of depth maps to pseudo-color images has been attempted by certain researchers to adapt standard convolutional networks [31]. Alternatively, point cloud reconstruction has been leveraged to complement three-dimensional geometric information. Yet challenges such as data redundancy and deployment difficulties are still encountered in these approaches.

Overall, current RGB-D fusion methods still exhibit limitations, including coarse fusion granularity, inadequate semantic alignment, and poor dynamic adaptability. These limitations constrain their application promotion in practical scenarios such as agricultural fruit detection.

2.2 YOLO Series Detectors

The YOLO series is recognized as one of the most mainstream real-time object detection frameworks currently available, and continuous iteration of model performance has been achieved since the introduction of YOLOv11 [1-4]. To enhance multi-scale detection capabilities, FPN structures are introduced in YOLOv11. For the optimization of network structure, CSPDarknet and PANet are integrated into YOLOv4. Meanwhile, lightweight design and deployment adaptability are the core focuses of the YOLOv5 and YOLOv6 series [3].

End-to-end structure optimization is emphasized in the latest YOLOv8, YOLOv10, and YOLOv11 series. Through this optimization, the performance of anchor-free detection is enhanced, and the adaptability to small targets and occluded objects is simultaneously improved [4-5, 29]. Among these models, optimizations have been implemented in YOLOv11 regarding fundamental feature extraction, decoupled head structures, and lightweight modules. This optimization balances detection accuracy and inference speed, thereby enabling YOLOv11 to serve as an ideal backbone for the construction of efficient detection systems.

However, only unimodal RGB input is supported by the original YOLO series, which leads to difficulties in directly addressing multimodal fusion tasks. Therefore, the integration of efficient modal interaction and fusion mechanisms into YOLO architectures has emerged as a research focus in recent years. Certain improved YOLO variants, e.g., YOLO-Fusion and MM-YOLO, have attempted to integrate dual-branch structures, yet in-depth modeling of fusion strategies is lacking in most of these variants [30-31].

2.3 Transformer Multimodal Fusion

Transformer architectures have been rapidly adopted in image recognition tasks owing to their exceptional global modeling capabilities [12-13]. Vision Transformers, e.g., Vision Transformer, Swin-Transformer, have been shown to outperform traditional CNN models in tasks such as image classification, object detection, and image segmentation [12]. In multimodal tasks, Transformers have been recognized as bridges for information interaction, enabling the modeling of global relationships and semantic correspondences between different modalities.

Cross-modal attention mechanisms have been introduced by methods such as BEVFormer and TransFuse to achieve spatial semantic fusion [14-15]; meanwhile, enhanced modeling effects for multi-frame and multimodal data have been achieved by MViTv2 and MTTR through the construction of unified spatiotemporal representations [16]. However, Transformers are associated with high computational complexity, insufficient focus on small targets, and a lack of structural optimizations tailored to address issues such as occlusion, multi-scale variation, and modal alignment within the domain of agricultural vision.

Lightweight cross-attention modules have been introduced by recent methods, e.g., MFPFormer, during the feature fusion stage. Additionally, dual-gating mechanisms have been incorporated by DAFFormer to enhance semantic alignment capabilities [32]. The proposed ***BiGCFormer*** structure adheres to the technical pathway of lightweight design, multimodal perception, and gated interaction. Based on YOLOv11 backbone, Transformer-enhanced cross-attention modules have been constructed to achieve high-robustness detection of fruit targets in complex environments.

In current multimodal object detection research, ***RGB-D*** fusion, YOLO architecture optimization, and Transformer-based semantic modeling have been identified as three key directions. These three approaches have been integrated in the proposed ***BiGCFormer*** detection framework to exploit their respective advantage. Therefore, ***BiGCFormer*** offers advantages such as dynamic perception, modal alignment, and structural efficiency. It also provides enhanced engineering adaptability and practical deployment potential for agricultural perception tasks.

3. METHODOLOGY

3.1 Overall Research Framework

A novel dual-modal fusion framework ***BiGCFormer*** is proposed to address the challenges associated with raspberry detection for robotic harvesting, including frequent occlusions, high color similarity between

fruits and foliage, and dynamic environmental factors such as illumination variations. It integrates RGB and depth data within a unified information-theoretic paradigm. The existing RGB-D approaches often rely on fragmented enhancements (e.g., independent alignment or fixed-weight fusion), assume modal homogeneity, and fail to adapt to dynamic fluctuations in modal quality. However, the proposed framework ***BiGCFormer*** introduces a systemic innovation: it is a closed-loop architecture, and it integrates modality-specific geometric alignment, bidirectional cross-modal interaction, and variational gated fusion. This architecture is optimized in an end-to-end manner via the maximization of the evidence lower bound (ELBO).

Referred to as A→T→G full-chain integration, this closed-loop design ensures that information from each module is propagated to the subsequent one, thereby forming a theoretically complete global solution that transcends the incremental modifications of existing methods. Tailored to raspberry-specific challenges such as clustered growth and geometric discontinuities, ***BiGCFormer*** enables robust localization and maturity assessment of raspberries in complex field environments.

Formally, the task is to learn an optimal mapping f^*: $(X_{rgb}, X_{depth})\rightarrow Y$. Where, X_{rgb} and X_{depth} represent RGB and depth feature spaces and Y includes bounding boxes (x, y, w, h), maturity probabilities, and confidence scores. The process $\Psi_{BiGC} = G_{fusion} \circ T_{cross} \circ A_{align}$ is decomposed into three sequential parts: modalities are projected onto a shared semantic manifold by the alignment module A_{align}, i.e., the Modal-Aware Positional Encoding (MAPE) module, reciprocal information transfer between modalities is enabled by the Bi-CA mechanism, and uncertainty-aware integration is performed by the Bi-Gate mechanism. This forms a rigorous logical closed-loop: alignment mitigates heterogeneity to enable interaction, interaction enriches features for fusion, and fusion drives the YOLOv11 anchor-free detection head. The framework satisfies alignment constraints

$$\exists \phi \text{ s.t. } \left\| \varphi(f_{rgb}) - f_{depth} \right\|^2 \leq \varepsilon \tag{1}$$

information preservation

$$\begin{aligned} I(F_{fused};\ F_{rgb}) &\geq \alpha_{rgb} \\ I(F_{fused};\ F_{depth}) &\geq \alpha_{depth} \end{aligned} \tag{2}$$

α_{rgb} and α_{depth} denote the minimum mutual-information preservation thresholds for RGB and Depth modalities

$$\left\| \Psi(f+\delta) - \Psi(f) \right\|_2^2 \leq \beta \sigma_{for} \left\| \delta \right\| \leq \sigma \tag{3}$$

The primary innovation of this study lies in the first-ever unification of modal alignment, mutual information-driven cross-modal interaction, and Bayesian gated fusion into an end-to-end closed-loop system for fruit and vegetable detection. This stands in contrast to fragmented RGB-D YOLO variants, in which components are enhanced in isolation. As depicted in Figure 1, this systemic design ensures coordinated optimization under a unified ELBO objective, thereby providing a theoretically complete solution for addressing the dynamic conditions of raspberry fields.

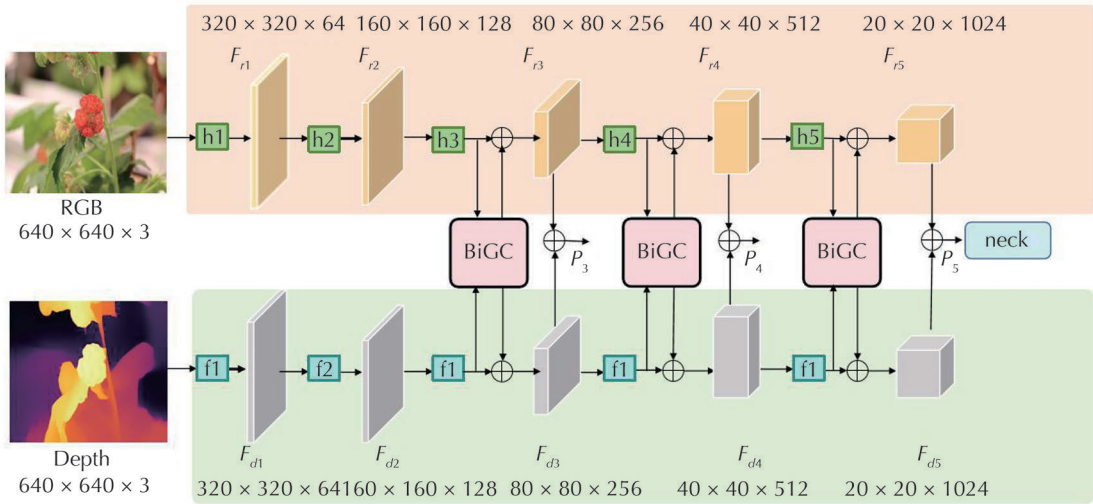


Figure 1. Cross-modal fusion backbone architecture for RGB and Depth modalities, utilizing multi-stage feature extraction and BiGCFormer modules to fuse hierarchical features across scales (from 320 × 320 × 64 to 20 × 20 × 1024).

This backbone network comprises two components: a dual-stream feature extraction network and three cross-modal fusion Transformer modules. Where, h and f represent RGB and depth image convolutional modules, respectively; F_r and F_d denote respective modal feature maps, respectively; and P_i represents subsequent feature pyramid inputs.

3.2 Modal-Aware Positional Encoding Module

Unlike existing RGB-D methods with shared positional encodings (assuming modal homogeneity) that cause misalignment in heterogeneous raspberry detection, the modal-aware positional encoding (MAPE) module introduces modality-specific sinusoidal encodings with learnable phase shifts. This design is not merely a variant of existing encodings but a key innovation in which modality-specific geometric manifolds are explicitly modeled. Within this innovation, the Fisher discriminant ratio is optimized to enhance the separability of modal features while preserving intra-modal fidelity. Through addressing raspberry-specific challenges such as depth discontinuities in occluded fruit clusters, the MAPE module establishes a discriminative foundation for the proposed closed-loop framework.

For position (x, y), the RGB encoding is:

$$P_{rgb}(x,\ y)_k = \begin{cases} \sin\left(\dfrac{x}{10000^{\frac{2k}{d}}} + \varphi_{rgb}^{k}\right) & \text{for even } k \\ \cos\left(\dfrac{y}{10000^{\frac{2k}{d}}} + \varphi_{rgb}^{k}\right) & \text{for odd } k \end{cases} \tag{4}$$

Where, d denotes the feature dimension, k indexes dimension up to d,and 10000 denotes the multi-scale frequency scaling factor,which scales frequencies for multi-scale capture, and φ_{rgb}^{k} are learnable phase shifts optimized via gradient descent to adapt to spectral reflections of RGB in raspberry clusters. Depth encoding P_{depth} uses independent φ_{depth}^{k}, ensuring differential projection onto a shared manifold $M \subset \mathbb{R}^{d}$.

Features F_{rgb}, $F_{depth} \in \mathbb{R}^{H \times W \times C}$ from the YOLOv11 backbone are flattened to (H, W, C), linearly projected, and augmented: $T_{rgb}^{enh} = \text{proj}(F_{rgb}) + P_{rgb}$, $T_{depth}^{enh} = \text{proj}(F_{depth}) + P_{depth}$. Within this modality-aware design, aligned tokens are directly fed into the Bidirectional Cross-Attention (Bi-CA) module. Through this module, precise cross-modal interaction is enabled, which in turn closes the loop, thereby distinguishing the approach proposed in this study from generic encoding adaptations.

3.3 Bidirectional Cross-Attention Mechanism

The unidirectional cross-attention adopted in prior works only captures dependencies from primary to auxiliary modalities and underutilizes modal complementarity in raspberry detection scenarios such as depth information is used to refine RGB edges under occluded conditions. However, the ***Bi-CA*** mechanism proposed in this study pioneers bidirectional mutual interaction integrated with mutual information regularization. This design is not a symmetric extension of conventional cross-attention, but a novel structure in which interdependencies between modalities are dynamically modeled via correlation masks. This structure ensures maximal information transfer within the closed-loop framework and is optimized for raspberry-specific dynamic characteristics, such as geometric-texture synergy.

RGB-to-depth attention computes:

$$A_{R \to D} = \text{softmax}\left(\frac{Q_R K_D^T}{\sqrt{d_k}} + M_{\text{cross}}^{R \to D} \right) V_D \tag{5}$$

Where, $Q_R = T_{rgb}^{enh} W_Q^R$, $K_D = T_{depth}^{enh} W_K^D$, $K_D = T_{depth}^{enh} W_K^D$, and $V_D = T_{depth}^{enh} W_V^D$ are projected queries, keys, and values, respectively; $\sqrt{d_k}$ is the dimension of the key vector used to scale the dot product and prevent gradient saturation; and

$$M_{\text{cross}}^{R \to D}[i, j] = \begin{cases} 0, & \text{if } C(f_{\text{rgb}}^{i} \to f_{\text{depth}}^{j}) = \dfrac{\partial \log p(f_{\text{depth}}^{j})}{\partial f_{\text{rgb}}^{i}} > \tau \\ -\infty, & \text{otherwise} \end{cases} \tag{6}$$

Where, τ is a learnable threshold, f_{rgb}^{i} denotes the feature representation of the i-th RGB token, while f_{depth}^{j} denotes the feature representation of the j-th depth token. The probability $p(f_{\text{depth}}^{j})$ represents the semantic prediction associated with the j-th depth token after the cross-attention transformation. The proposed causal relevance function measures the sensitivity of the semantic prediction of the j-th depth token with respect to perturbations of the i-th RGB token. A larger gradient magnitude indicates that the RGB token contributes more significantly to the semantic representation of the corresponding depth token. The computation is feasible because the cross-attention mechanism establishes differentiable interactions

between every RGB token and every depth token. Specifically, each depth token is generated as a weighted aggregation of all RGB tokens through the attention matrix, making the representation of the j-th depth token a differentiable function of all RGB tokens. Consequently, the gradient $\frac{\partial \log p(f^j_{\text{depth}})}{\partial f^i_{\text{rgb}}}$ is well-defined for any token pair (i, j), regardless of whether $i=j$, and can be efficiently computed through automatic differentiation during backpropagation.

The causal relevance function $C(f^i_{\text{rgb}} \to f^j_{\text{depth}})$ is computed as the gradient of the log-likelihood of the depth feature with respect to the corresponding RGB feature $\partial \log p(f^j_{\text{depth}}) / \partial$, thereby quantifying the degree to which changes in RGB features influence the distribution of depth features. A value of zero in the mask indicates that the connection between f^i_{rgb} and f^j_{depth} is preserved and contributes to the attention calculation, whereas $-\infty$ the interaction by nullifying its attention weight after SoftMax normalization. The learnable threshold τ adaptively controls the activation of these cross-modal links, ensuring that only statistically significant dependencies are retained during bidirectional attention modeling. The depth-to-RGB attention $A_{D\to R}$ is defined analogously.

To address the computational intractability of mutual information $I(T^{enh}_{rgb}; T^{enh}_{depth})$ in high-dimensional spaces, the mutual information is approximated via the Mutual Information Neural Estimator (MINE), which is formulated as follows:

$$\mathcal{L}_{info} = -\lambda_1 \hat{I}_{MINE}(T_{rgb}; T_{depth}) - \lambda_2 \hat{I}_{MINE}(T_{depth}; T_{rgb}) + \lambda_3 H(T_{rgb}, T_{depth}) \tag{7}$$

Where, $\hat{I}_{MINE}$ estimates mutual information via a neural network trained on joint and marginal distributions, $H(\cdot, \cdot)$ is joint entropy, and λ_i is balance terms.

Stability is ensured by clipping gradients and regularizing the estimator network, mitigating variance issues in high-dimensional settings. This bidirectional innovation, as depicted in Figure 2, closes the loop by producing enhanced tokens $(T^{enh}_{rgb}*, T^{enh}_{depth})$ for Bi-Gate fusion, reinforcing systemic originality over isolated attention enhancements.

3.4 Bidirectional Gated Residual Fusion Mechanism

As shown in Figure 3, in the prior fixed-weight fusion methods such as SoftMax-based weighting, modal uncertainties in dynamic raspberry detection fields are neglected. The Bi-Gate mechanism proposed in this study incorporates gated residuals within a variational Bayesian framework to adaptively regulate modal contribution weights. This design is not a mere repetition of existing dynamic fusion or uncertainty weighting approaches, but a novel integration in which uncertainties are propagated via residuals. Optimized under the evidence lower bound (ELBO) objective, this integration completes the closed-loop architecture and provides a theoretically grounded solution for addressing raspberry-related variabilities.

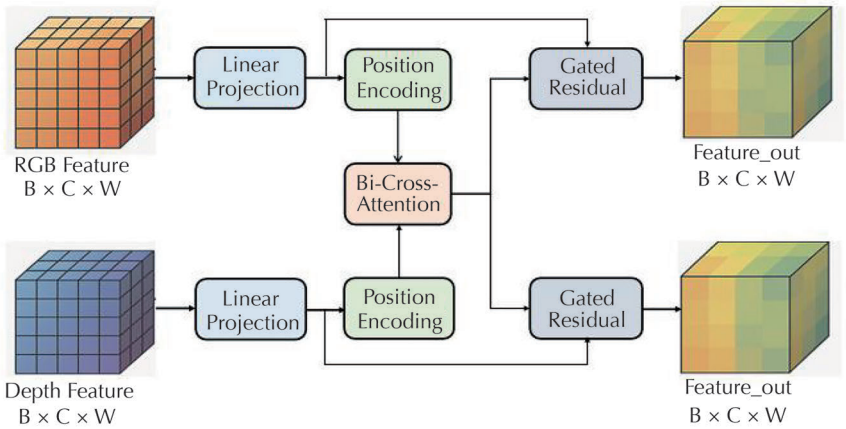


Figure 2. The structure of BiGCFormer utilizing Bi-Gate fusion (with Bi-Cross-Attention, Position Encoding, and Gated Residual) to reinforce systemic originality over isolated attention enhancements.

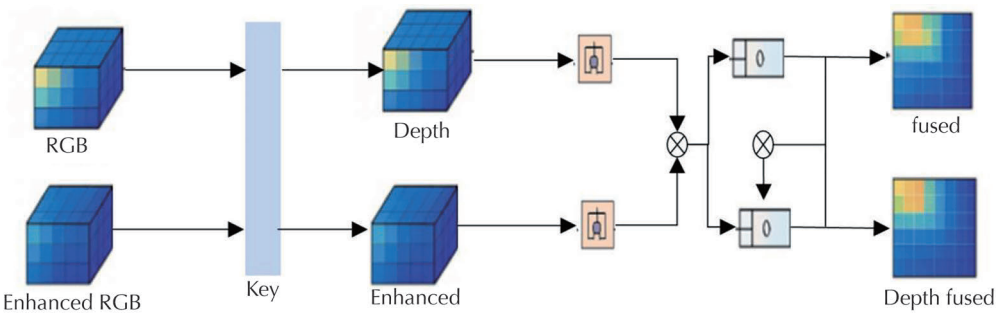


Figure 3. Bidirectional Gated Residual Fusion Mechanism for Multi-Modal Features, Utilizing RGB, Enhanced RGB, and Corresponding Depth Features with Key-Guided Processing to Generate Both Fused and Depth-Fused Representations.

The fusion process operates through adaptive weighting of both mean and variance components:

$$\mu_{\text{fused}} = z_{\text{orig}} + g_\theta(z_{\text{enh}}) \odot (z_{\text{enh}} - z_{\text{orig}}), \tag{8}$$

$$\sigma^2_{\text{fused}} = (1 - g_\theta(z_{\text{enh}}))^2 \sigma^2_{\text{orig}} + g_\theta(z_{\text{enh}})^2 \sigma^2_{\text{enh}}. \tag{9}$$

Where, z_{orig} and z_{enh} denote original and Bi-CA-enhanced features, respectively; $\odot$ is element-wise multiplication; and $g_\theta(z_{\text{enh}}) = \sigma\left(\text{MLP}\left(\text{concat}(z_{\text{enh}}, \text{context})\right)\right) \in [0, 1]$, with context aggregating statistics $(\mu_{\text{rgb}}, \sigma^2_{\text{rgb}}, \mu_{\text{depth}}, \sigma^2_{\text{depth}})$ and environmentals (illumination, distance).

Optimization targets the ELBO, the evidence lower bound objective balances reconstruction accuracy with regularization:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(z_{\text{fused}})}\left[\log p(y \mid z_{\text{fused}})\right] - D_{\text{KL}}\left(q_\phi(z_{\text{fused}}) \,\|\, p(z_{\text{fused}} \mid z_{\text{orig}}, z_{\text{enh}})\right) \tag{10}$$

Where, $q_\phi(z_{\text{fused}})$ approximates the posterior, D_{KL} quantifies divergence, and $p(y|z_{\text{fused}})$ links to detection likelihood. Residuals preserve critical information, enabling adaptive weighting (e.g., prioritizing depth in low light). This closes the loop: fused features drive YOLOv11 predictions, with backpropagation refining upstream modules, completing the first unified closed-loop system for fruit-vegetable detection, distinct from fragmented RGB-D enhancements, and delivering superior performance in complex field scenarios.

4. EXPERIMENTS

4.1 Experimental Setup

To systematically verify the performance superiority of ***BiGCFormer*** in raspberry detection tasks within complex orchard environments, and to thoroughly analyze the unique contributions of its innovative modules, namely, MAPE, Bi-CA, and Bi-Gate, a series of rigorous experimental evaluations are designed and implemented. These evaluations are based on a self-constructed multimodal raspberry detection dataset. In the experiments, the performance of ***BiGCFormer*** is comprehensively compared with that of existing unimodal and multimodal object detection methods. The theoretical foundations and practical effects of each module are also revealed in enhancing detection accuracy, robustness, and generalization capabilities. This revelation is achieved through targeted scenario testing, robustness analysis, fine-grained ablation experiments, and qualitative visualization.

Dataset construction, experimental configuration, evaluation metrics, comparative experimental results, robustness analysis, ablation experiments, and qualitative analysis are elaborated in this section. Meanwhile, the theoretical innovations of ***BiGCFormer*** within the frameworks of information theory, variational inference, and causal modeling are highlighted., Its engineering applicability in complex agricultural scenarios is also highlighted.

4.2 Dataset

This study constructs a multimodal raspberry detection dataset specifically designed for complex orchard environments, containing 6,500 RGB-D image pairs with resolution of 640 × 640, divided into training set, validation set, and test set with the numbers of 5,200 pairs, 650 pairs, and 650 pairs, respectively. Depth maps utilize MiDaS model generation with depth values normalized to [0, 1] intervals. Depth maps are generated using the MiDaS model, which provides a practical and cost-effective solution for large-scale data construction in open orchard environments where deploying physical RGB-D sensors is challenging. The dataset covers the following key scenarios to fully simulate complex challenges in real orchard environments:

(1) Illumination variations: encompassing strong light (>120,000 lux), weak light (<3,000 lux), backlight, and shadow conditions to evaluate model adaptation capabilities under different lighting conditions.

(2) Occlusion scenarios: including moderate occlusion (30%–50%) and extreme occlusion (>70%) to verify model capabilities in fruit edge localization and target separation.

(3) Dense fruit distribution: each frame contains 10-60 fruits with target sizes ranging from 16 × 16 pixels to 128 × 128 pixels, challenging model performance in multi-target separation and small target detection.

(4) Modal degradation: Gaussian noise ($\sigma = 0.1$–0.3), RGB overexposure (brightness gain 1.2–2.0 times), color shift (saturation ±0.3), and random depth pixel loss (5%–25%) are applied to simulate sensor noise and environmental interference to test model robustness.

Each image pair is annotated with raspberry bounding boxes, maturity labels (mature/immature), and confidence scores. The accuracy of these annotations is verified through manual review, which ensures high data quality. For the purpose of quantitative evaluation, scenario-specific subsets are drawn exclusively from the test split (650 image pairs). Specifically: the bright-light subset contains images with mean pixel brightness >180 (0-255 scale, corresponding to the overexposure condition in scenario 4); the moderate-occlusion subset contains images where target fruits are occluded by 30-50% (scenario 2); and the small-target subset contains images with at least one bounding box whose pixel area is <50 pixels (scenario 3). These subsets are not mutually exclusive — a single image may belong to more than one category simultaneously. A solid foundation for evaluating the robustness and generalization capabilities of ***BiGCFormer*** is laid by the high challenge level and diversity of the dataset. Furthermore, the dataset is particularly suitable for verifying the performance of ***BiGCFormer*** in the actual deployment of agricultural robots.

4.3 Experimental Configuration

Experiments are conducted on high-performance computing platforms with hardware configurations including an A100 (40 GB VRAM), an Intel Xeon Silver 4210 CPU, and 128 GB of memory. The software environment is based on PyTorch 2.1, CUDA 12.0, and Python 3.9. ***BiGCFormer*** employs YOLOv11 as its backbone network and adopts a dual-stream architecture to process RGB and depth modalities separately. The parameter count of model is optimized to 27.5M, with an inference speed of 56 FPS achieved on the RTX3090. This performance is superior to that of comparative methods, including FMB (48 FPS, 30.2M parameters), CDDFuse (45 FPS, 32.1M parameters), and IGNet (46 FPS, 31.8M parameters), demonstrating the high-efficiency deployment potential of ***BiGCFormer*** in embedded agricultural robots.

Training configuration includes: (1) optimizer and learning rate: AdamW optimizer with initial learning rate 0.001, combined with a cosine annealing schedule (minimum learning rate 1e-6), 300 training epochs, and batch size 16, to ensure adequate model convergence. (2) loss function: integrating classification loss (cross-entropy), bounding box regression loss (CIoU), and Bi-CA module information theory loss (weights $\lambda^1 = 1.0$, $\lambda^2 = 1.0$, $\lambda^3 = 0.5$), optimizing fusion effects through maximizing inter-modal mutual information $I(F_{rgb}, F_{depth})$ and minimizing joint entropy $H(F_{rgb}, F_{depth})$.

Evaluation metrics include: (1) mAP@.5: Mean Average Precision at IoU threshold 0.5, measuring overall detection performance; (2) mAP@.5:95: Mean Average Precision across IoU thresholds 0.5-0.95, evaluating detection precision under strict standards; (3) Precision (P): ratio of correctly detected targets to all detected targets, reflecting false positive control capability; (4) Recall (R): ratio of correctly detected

targets to all actual targets, reflecting missed detection control capability; (5) small target detection accuracy: computing mAP@.5 for targets with area <50 pixels, assessing small target handling capability; (6) occlusion scenario performance: computing mAP@.5 for scenarios with occlusion >50%, evaluating complex scenario adaptability; and (7) computational efficiency: evaluating real-time performance and embedded deployment potential through inference speed (FPS) and FLOPs.

To quantify robustness, experiments record mAP@.5 decrease amplitude under modal degradation scenarios and further verify model performance advantages in complex agricultural scenarios through specialized metrics for small targets and occlusion scenarios.

4.4 Comparative Experimental Results and Analysis

Comprehensive comparative experiments are conducted on agricultural fruit detection datasets, employing mAP@.5, mAP@.5:95, Precision, and Recall as evaluation metrics. Experiments encompass mainstream unimodal detection methods, multimodal fusion methods, and YOLO series detectors to verify effectiveness of the proposed ***BiGCFormer***, with the experimental results are shown in Table 1.

Table 1. Performance comparison results between BiGCFormer and mainstream unimodal and multimodal object detection methods. Note: Methods with different backbones are included for broad reference. The most controlled comparison is between BiGCFormer and YOLOv11 + Two-Stream, which share the same YOLOv11 backbone, training dataset, and hyperparameter settings; the 3.1 percentage point improvement in this pair directly reflects the contribution of the proposed fusion modules.

MODEL	MODALITY	MAP.5	MAP.5:95	P(%)	R(%)
FASTERR-CNN (NIPS'16)	RGB	42.1	21.9	70.1	60.3
CASCADER-CNN (CVPR'18)	RGB	57.4	38.9	72.0	62.2
YOLOV5 (ZENODO'22)	RGB	60.9	41.5	64.1	60.3
YOLOV11 (2024)	RGB	63.8	43.8	66.8	62.7
RTDETR (CVPR'24)	RGB	59.7	41.1	65.1	62.1
FASTERR-CNN (NIPS'16)	Depth	21.4	13.2	35.1	18.6
CASCADER-CNN (CVPR'18)	Depth	24.8	16.9	39.4	21.3
RTDETR (CVPR'24)	Depth	23.7	15.6	43.0	27.1
CFT (2021)	RGB + Depth	66.3	42.2	63.7	64.5
DIDFUSE (2020)	RGB + Depth	64.2	40.5	61.5	62.3
SDNET (IJCV'21)	RGB + Depth	65.5	42.2	64.1	63.1
RFNET (CVPR'22)	RGB + Depth	67.2	43.3	65.5	64.2
TARDAL (CVPR'22)	RGB + Depth	66.6	42.7	65.8	63.4

Table 1. *Continued.*

MODEL	MODALITY	MAP.5	MAP.5:95	P(%)	R(%)
DEFUSION (MM'22)	RGB + Depth	66.9	43.3	65.2	63.5
CDDFUSE (CVPR'23)	RGB + Depth	67.1	43.4	66.1	64.0
IGNET (MM'23)	RGB + Depth	67.3	43.5	66.3	64.4
SUPERFUSION (JAS'22)	RGB + Depth	67.0	43.2	66.2	64.5
FMB (2024)	Depth + RGB	68.8	43.1	66.8	66.1
OURS (YOLOV11 + BIGCFORMER)	Depth + RGB	**70.9**	**45.6**	**67.4**	**68.7**

In unimodal RGB detection tasks, traditional two-stage detectors FasterR-CNN and CascadeR-CNN achieve 42.1% and 57.4% mAP@.5 respectively, while Transformer-based RTDETR reaches 59.7%. YOLO series methods demonstrate gradual improvement trends, advancing from 60.9% of YOLOv5 to 63.8% of YOLOv11. However, when these methods apply to depth modality, performance exhibits significant decline. FasterR-CNN depth detection accuracy reaches only 21.4%, CascadeR-CNN achieves 24.8%, and even the latest YOLOv11 attains merely 33.4% mAP@.5 under depth modality. This phenomenon indicates inherent limitations of unimodal methods in processing depth information, where depth data sparsity and geometric characteristics make traditional detection architectures difficult to effectively extract discriminative features.

Multimodal fusion methods significantly improve detection performance by combining RGB and depth information. CFT method achieves 66.3% mAP@.5, DIDFuse reaches 64.2%, SDNet achieves 65.5%. These early fusion methods primarily implement modal integration through feature-level concatenation or simple attention mechanisms. Recent advanced methods including RFNet (67.0%), TarDAL (66.6%), DeFusion (66.9%), CDDFuse (67.1%), and IGNet (67.3%) demonstrate improvements in fusion strategies, yet performance enhancements remain relatively limited. The latest FMB method achieves 68.8% mAP@.5 through feature memory bank mechanisms, representing current highest multimodal fusion levels.

The proposed ***BiGCFormer*** achieves a 70.9% mAP@.5, exhibiting an improvement of 2.1 percentage points relative to the optimal baseline FMB method. For the more stringent mAP@.5:95 metric, it reaches 45.6%, with a 2.5 percentage point improvement. More importantly, ***BiGCFormer*** yields significant enhancements in both precision (67.4%) and recall (68.7%): relative to FMB which achieves 66.8% in precision and 66.1% in recall, the improvements are 0.6 and 2.6 percentage points, respectively; relative to CDDFuse which achieves 66.0% in precision and 64.0% in recall, the corresponding improvements are 1.4 and 4.7 percentage points, respectively. Such balanced performance improvements demonstrate that ***BiGCFormer*** not only reduces false detections but also effectively lowers missed detection rates, thereby bearing significant implications for applications in agricultural automation.

Figure 4 presents detailed performance comparison between ***BiGCFormer*** and YOLO series methods (v5, v8, v10, v11) under different modal configurations. Under RGB unimodal configuration, YOLO series methods demonstrate continuous performance improvement trends, gradually advancing from YOLOv5 with a precision of 60.9% and from YOLOv11 with a precision of 63.8%, with corresponding improvements in precision. However, under depth unimodal configuration, all YOLO methods experience significant performance decline, with YOLOv11-Depth achieving only 33.4% mAP@.5 and with precision and recall decreasing to 45.4%, respectively. These results further confirm challenges posed by depth data special properties to traditional detection architectures.

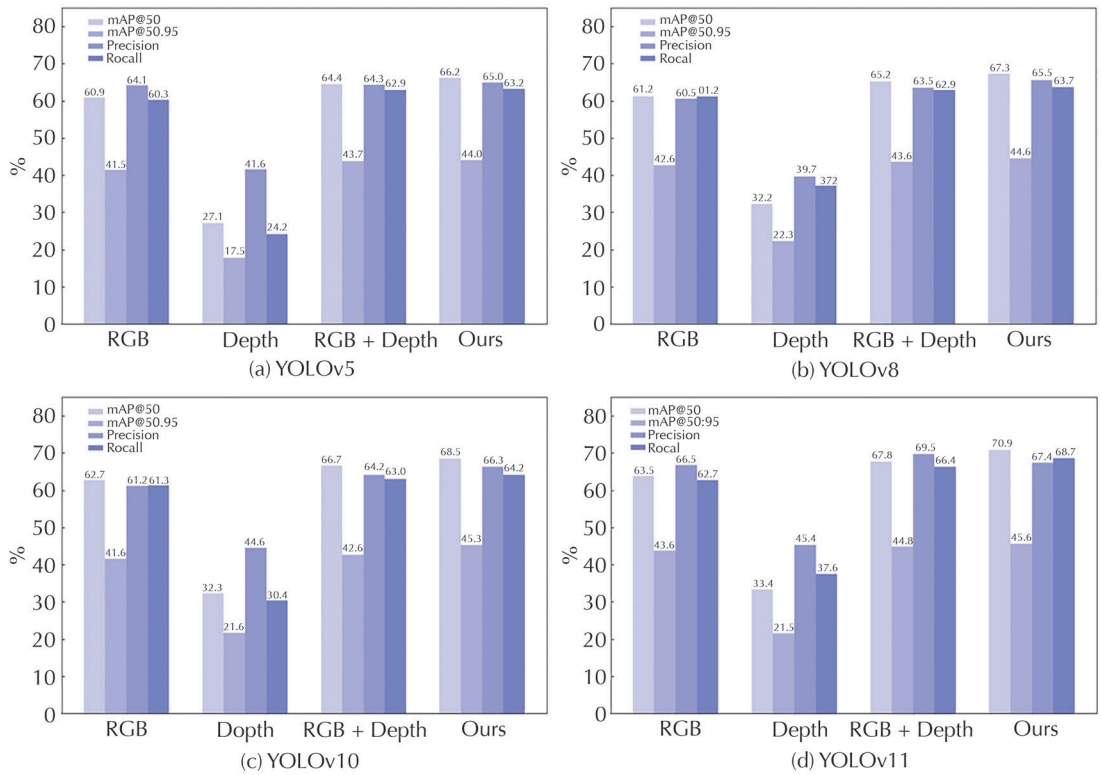


Figure 4. Comparison of mAP@.5, mAP@.5:95, Precision, and Recall for different YOLO versions and fusion methods under RGB, Depth, and multimodal configurations.

YOLO methods employing simple Two-Stream fusion strategies achieve significant performance improvements under multimodal configurations. In mAP@.5, YOLOv11 + Two-Stream reaches 67.8%, YOLOv10 + Two-Stream achieves 66.7%, YOLOv8 + Two-Stream attains 65.2%. These methods achieve basic modal information integration through early feature concatenation, yet due to lacking deep cross-modal interaction mechanisms, performance improvements remain relatively limited. Particularly under

the more stringent mAP@.5:95 evaluation metric, YOLOv11 + Two-Stream achieves only 44.8%, indicating insufficient detection accuracy under high IoU thresholds.

In contrast, ***BiGCFormer*** achieves deeper multimodal feature fusion by introducing core modules MAPE, Bi-CA, and Bi-Gate. Experimental results demonstrate that ***BiGCFormer*** improves 3.1 percentage points over YOLOv11 + Two-Stream in mAP@.5 (70.9% vs 67.8%), 0.8 percentage points in mAP@.5:95 (45.6% vs 44.8%), 0.8 percentage points in precision, and 2.3 percentage points in recall. This comprehensive performance improvement validates effectiveness of the proposed geometric interaction framework.

Under mAP@.5:95, a comprehensive evaluation metric that covers multiple IoU thresholds, ***BiGCFormer*** demonstrates clear advantages, as shown in Figure 5. Traditional multimodal methods including CFT (42.0%), DIDFuse (40.5%), and SDNet (42.2%) exhibit obvious performance decline under strict evaluation standards, while ***BiGCFormer*** achieves 45.6%, improving 2.5 percentage points compared with optimal baseline FMB (43.1%). This result indicates that ***BiGCFormer*** not only performs excellently in coarse-grained detection tasks but also possesses significant advantages in fine-grained localization tasks. The MAPE module explicitly models geometric relationships between RGB and depth features via modal-specific positional encoding; the Bi-CA module achieves cross-modal semantic alignment through bidirectional causal attention; the Bi-Gate module optimizes fusion effects with adaptive weight adjustment. Their synergy enables ***BiGCFormer*** to perform exceptionally in multi-scale object detection.

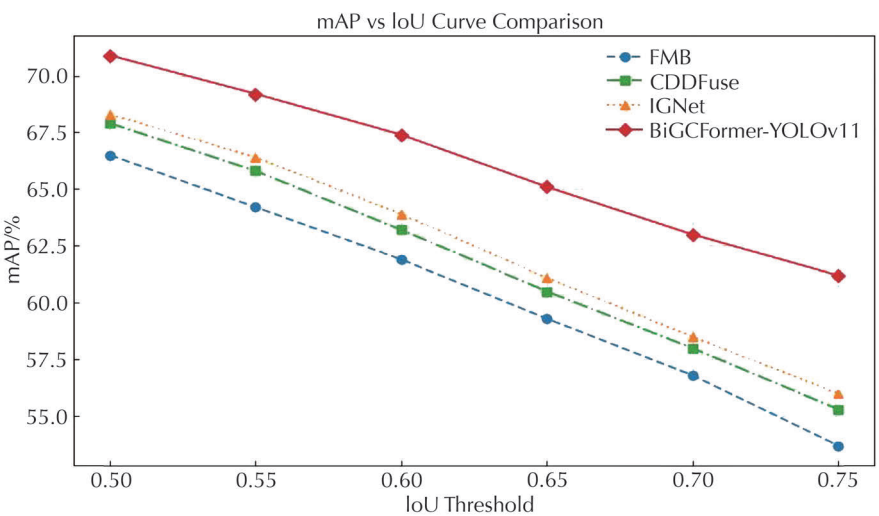


Figure 5. Performance of mainstream multimodal fusion models under different IoU thresholds.

The significant advantages of ***BiGCFormer*** over existing methods stem from its core innovations in multimodal geometric interaction. Conventional multimodal fusion approaches, including CFT, DIDFuse, and SDNet, predominantly rely on feature-level concatenation or simple attention mechanisms — strategies that prove insufficient for bridging the representational gap between RGB and depth modalities.

Their key limitation lies in applying unified positional encoding to heterogeneous modal features, which conflates spatially distinct representations and leads to systematic loss of geometric information as well as misalignment in inter-modal semantic correspondence. BiGCFormer directly addresses these failure modes through geometry-aware cross-modal interaction, enabling the model to explicitly respect and exploit the structural heterogeneity between modalities rather than forcing them into a shared representational space prematurely.

Modal-specific positional encoding strategies are first proposed by ***BiGCFormer*** via the MAPE module. Independent sinusoidal positional encoding functions for RGB and depth features ($\phi^{krgb} \neq \phi^{kdepth}$) are designed to explicitly model the geometric heterogeneity of these two modalities. Through this design, precise cross-modal spatial correspondences can be established by networks, while the inherent geometric characteristics of each modality are maintained. Compared with the unified positional encoding employed by methods such as BEVFormer and MFPFormer, geometric information confusion is effectively avoided by MAPE, which lays a solid foundation for subsequent cross-modal interaction.

More importantly, bidirectional causal attention mechanisms are constructed by ***BiGCFormer*** through the Bi-CA module, with deep RGB-D semantic interaction being achieved. Unidirectional guidance strategies are employed by traditional methods including IGNet, allowing only limited modal information transfer. Based on the causal mask matrix M^causal, bidirectional information flows from RGB to depth and from depth to RGB are simultaneously modeled by Bi-CA, enabling sufficient complementary enhancement of the two modalities at the semantic level. Not only is depth geometric information utilized by this bidirectional interaction mechanism to guide RGB edge detection, but also is RGB texture information leveraged to compensate for depth data sparsity, thereby significantly improving detection robustness in complex scenarios.

The innovative nature of ***BiGCFormer*** is further embodied in the adaptive fusion strategy of the Bi-Gate module. In contrast to fixed-weight fusion employed by methods including FMB, modal weights are dynamically adjusted by Bi-Gate through Bayesian gating mechanisms, with fusion strategies being adaptively optimized according to specific characteristics of input scenarios such as illumination conditions, occlusion degrees, and target densities. Stable detection performance of ***BiGCFormer*** under different environmental conditions is enabled by this context-aware dynamic weight adjustment, with clear advantages being demonstrated particularly in agricultural scenarios characterized by dramatic illumination changes or severe occlusion.

Experimental results demonstrate that this systematic innovative design yields comprehensive performance improvements. Under the mAP@.5 metric, ***BiGCFormer*** achieves a 2.1 percentage point improvement compared with the optimal baseline FMB, this improvement is of practical significance in object detection.More importantly, under the more stringent mAP@.5:95 evaluation criterion, ***BiGCFormer*** achieves a 2.5 percentage point improvement, indicating excellent performance not only in coarse-grained detection but also in fine-grained localization tasks. Simultaneous improvements in precision and recall, 0.6 and 2.6 percentage points respectively compared with that of FMB, further confirm the effectiveness of

BiGCFormer in reducing false positives and false negatives, thereby holding significant practical value for agricultural automation applications requiring high reliability.

Experimental results fully validate the effectiveness and advancement of the multimodal geometric interaction framework constructed by ***BiGCFormer***. By systematically addressing the limitations of existing methods in geometric alignment, cross-modal interaction, and adaptive fusion, ***BiGCFormer*** opens new technical avenues for multimodal object detection, offering important theoretical foundations and technical support for the development of agricultural intelligent technologies.

4.5 Ablation Experiments

To thoroughly analyze the contribution degrees of core components of ***BiGCFormer***, systematic ablation experiments are designed. The effects of different component combinations on detection performance are demonstrated by Figure 6 through progressive addition methods, with the process starting from the baseline YOLOv11 and innovative modules being gradually introduced to verify the effectiveness and necessity of each component.

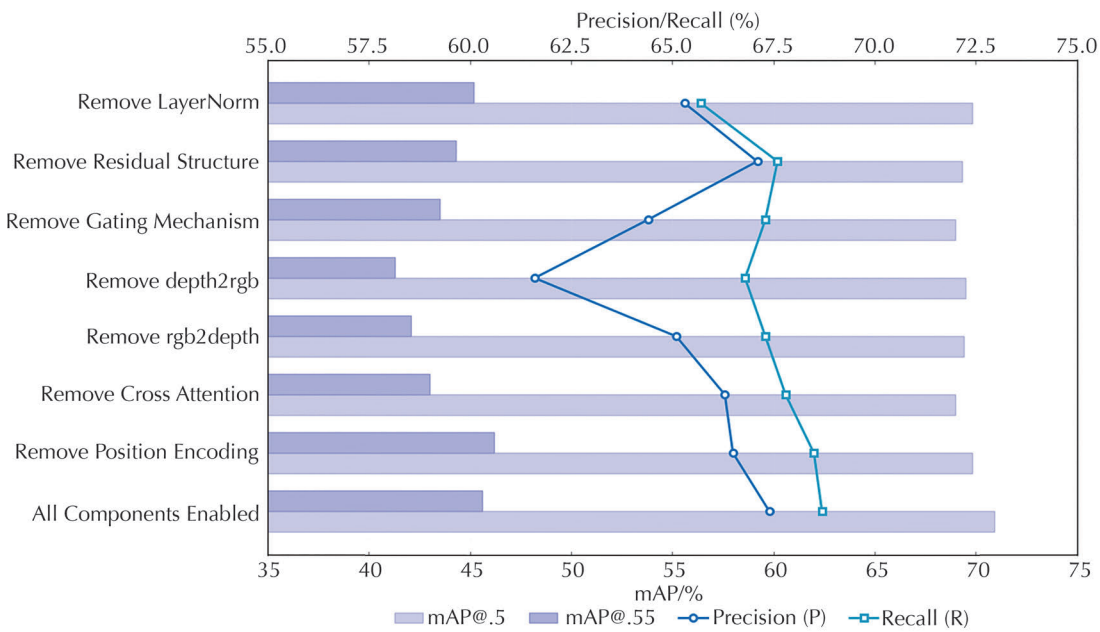


Figure 6. Ablation experiments. Detection performance (mAP@.5, mAP@.55, Precision, Recall) under removal of individual components. The full model with all modules enabled yields the best results.

YOLOv11 is employed as the baseline architecture in the experiments, with separate testing conducted under RGB and depth unimodal configurations. A value of 63.8% in mAP@.5, along with that of 66.8% in precision and that of 62.7% in recall, are achieved by the RGB unimodal YOLOv11, which is consistent with the comparative experimental results presented in Table 1.

However, when switched to the depth unimodal configuration, performance is observed to undergo a dramatic decline, with mAP@.5 being observed to reach only 33.4%, and precision and recall being reduced to 45.4% and 37.6%, respectively. Inherent challenges of depth data are revealed by this significant performance difference: traditional detection architectures are rendered difficult to process effectively by the sparsity of depth images, noise interference, and the complexity of geometric information. More importantly, direct evidence for the necessity of multimodal fusion is provided by this result and the limitations of single modalities must be overcome through cross-modal information integration.

Significant system performance improvements are achieved by the introduction of the MAPE module on the baseline architecture. Only 69.8% mAP@.5 is achieved by the configuration with the MAPE module removed (and LayerNorm removed), while 70.9% is reached by the complete ***BiGCFormer***, a gap of 1.3% percentage points. This substantial difference indicates that fundamental roles in the entire framework are played by the MAPE module. Heterogeneity between the two modalities in geometric space is explicitly modeled by the MAPE module through the design of independent sinusoidal positional encoding functions for RGB and depth features. Unified positional encoding is employed by traditional methods to process different modal features, with geometric information confusion and spatial correspondence deviations inevitably resulting. The innovation of MAPE lies in the recognition of fundamental differences between RGB texture features and depth geometric features in spatial representations; inherent geometric characteristics of each modality are maintained through modal-specific encoding, and precise spatial foundations are established for subsequent cross-modal interaction.

System performance is further enhanced by the introduction of the Bi-CA module, with configurations featuring the removal of this module observed to decline to 67.8% mAP@.5, a decrease of 3.1% percentage points compared with that of the complete system. The importance of bidirectional causal attention mechanisms is validated by this significant difference. Unidirectional information transfer, e.g., from RGB to depth or depth to RGB, is often employed by traditional cross-modal attention mechanisms, with fundamental bidirectional complementary characteristics between modalities being neglected. Based on the causal mask matrix M^{causal}, bidirectional information flows, that are from RGB to Depth and from Depth to RGB, are simultaneously modeled by Bi-CA, enabling sufficient complementary enhancement of the two modalities at the semantic level. Specifically, RGB boundary detection and target segmentation are guided by depth geometric information, while depth data sparsity and noise issues are compensated for by RGB texture information. Not only is detection accuracy improved by this bidirectional interaction mechanism, but also, more importantly, is system robustness under complex scenarios enhanced.

Advantages of adaptive fusion strategies are demonstrated by the Bi-Gate module, with mAP@.5 of the system observed to decline to 68.5% following the removal of this module, a decrease of 2.4% percentage

points. The superiority of dynamic weight adjustment relative to fixed fusion strategies is highlighted by this result. Fixed weights or simple averaging are typically employed by traditional multimodal fusion methods to integrate information from different modalities; these are static strategies unable to adapt to changes in the quality of modal information under varying scenarios, as illustrated in Figure 7. Context-aware dynamic weight adjustment is achieved by Bi-Gate through Bayesian gating mechanisms, with fusion strategies being adaptively optimized according to specific characteristics of input scenarios such as illumination conditions, occlusion degrees, and target densities. In scenarios with sufficient illumination, RGB information is proven to be more reliable, with RGB modal weights being automatically increased by the system. In scenarios with severe occlusion, depth geometric information exhibits advantages, with depth modal weights being correspondingly increased by the system. Stable detection performance of ***BiGCFormer*** in variable agricultural environments is enabled by this intelligent fusion strategy.

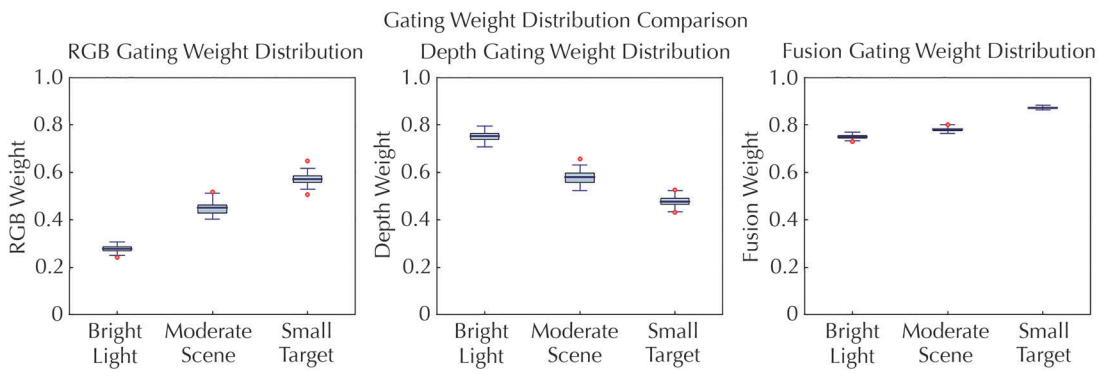


Figure 7. Gating weight distributions of RGB, Depth, and Fusion branches for different input scenario characteristics (Bright Light, Moderate Scene, Small Target).

Synergistic effects between components are also revealed by ablation experiments. Significant performance decline is induced by the individual removal of any core component, with cumulative decline amplitudes observed to far exceed the simple sum of individual component contributions. It is indicated by this phenomenon that the modules MAPE, Bi-CA, and Bi-Gate not only constitute mere functional stacking but also form an organic synergistic system. Accurate geometric foundations for cross-modal interaction are provided by MAPE, deep semantic integration on this foundation is achieved by Bi-CA, and fusion effects are further optimized by Bi-Gate. These three components are mutually supportive and reinforcing, with complete multimodal geometric interaction frameworks jointly constructed by them.

Different characteristics of component contributions are exhibited when examining trends in precision and recall. Recall is primarily improved by MAPE through the reduction of missed detections via enhanced geometric alignment; significant contributions to both precision and recall are made by Bi-CA, as bidirectional interaction mechanisms reduce both false positives and false negatives; and precision is

primarily optimized by Bi-Gate through the reduction of false positives via adaptive weight adjustment. The rationality and necessity of component designs are further confirmed by this differentiated contribution pattern.

The scientific nature and effectiveness of architectural design of ***BiGCFormer*** are fully validated by ablation experimental results. Specific functional roles are undertaken by each core component, while synergistic effects with other components are formed, leading to the joint achievement of significant improvements in multimodal object detection performance. Important theoretical guidance and practical references for the development of multimodal perception technologies are provided by this systematic innovative design.

4.6 Same-Backbone Fusion Strategy Comparison

To isolate the contribution of the proposed fusion modules from backbone-level differences, a controlled same-backbone comparison is conducted. Three fusion variants sharing the identical YOLOv11 backbone, training dataset, optimizer, and hyperparameter configuration are evaluated: (1) Early Fusion, where RGB and depth feature maps are concatenated channel-wise before the backbone neck; (2) Late Fusion, where predictions from two independent YOLOv11 branches are merged via weighted averaging at the detection head; and (3) Two-Stream, a parallel dual-branch architecture with element-wise summation at the feature pyramid level; and (4) the proposed ***BiGCFormer***, which replaces simple aggregation with the MAPE-Bi-CA-Bi-Gate pipeline. All variants are trained from scratch under identical conditions on the same 5,200-pair training split and evaluated on the 650-pair test split.

Results are presented in Table 2. Early Fusion achieves 65.3% mAP@.5 (Precision: 63.8%, Recall: 64.1%), indicating that naive concatenation fails to resolve cross-modal geometric misalignment. Late Fusion achieves 64.8% mAP@.5 (Precision: 63.2%, Recall: 63.7%), suggesting that decision-level merging loses fine-grained complementary cues. Two-Stream reaches 67.8% mAP@.5 (Precision: 65.9%, Recall: 66.2%), which benefits from parallel feature extraction but remains limited by shallow interaction. ***BiGCFormer*** achieves 70.9% mAP@.5 (Precision: 67.4%, Recall: 68.7%), outperforming the strongest same-backbone baseline (Two-Stream) by 3.1 percentage points. These results confirm that the

Table 2. Same-backbone fusion strategy comparison. All methods use YOLOv11 as the backbone with identical training settings. Bold values indicate best performance.

Method	Backbone	mAP@.5 (%)	mAP@.5:95 (%)	Precision (%)	Recall (%)
YOLOv11 + Early Fusion	YOLOv11	65.3	41.8	63.8	64.1
YOLOv11 + Late Fusion	YOLOv11	64.8	41.2	63.2	63.7
YOLOv11 + Two-Stream	YOLOv11	67.8	44.8	65.9	66.2
YOLOv11 + BiGCFormer (Ours)	YOLOv11	**70.9**	**45.6**	**67.4**	**68.7**

performance advantage of ***BiGCFormer*** originates from its fusion architecture — specifically the geometric alignment provided by MAPE, the semantic interaction enabled by Bi-CA, and the adaptive weighting of Bi-Gate — rather than from any backbone-level advantage. The progressive improvement from Early Fusion to Two-Stream to ***BiGCFormer*** further demonstrates that deeper and more structured cross-modal interaction yields consistent gains under controlled conditions.

4.7 Hyperparameter Sensitivity Analysis

The Bi-CA information theory loss in Equation (7) is governed by three balance coefficients: λ_1 (weight for RGB-to-Depth mutual information), λ_2 (weight for Depth-to-RGB mutual information), and λ_3 (weight for cross-modal redundancy regularization). To verify the robustness of the chosen values (λ_1 = 1.0, λ_2 = 1.0, λ_3 = 0.5) and provide empirical justification, a sensitivity analysis is conducted by varying each coefficient independently across the range {0.1, 0.5, 1.0, 1.5, 2.0} while fixing the other two at their default values. All other training settings remain identical.

Results are summarized in Table 3. For lambda-1, mAP@.5 varies between 70.1% (λ_1 = 0.1) and 70.9% (λ_1 = 1.0), with a peak-to-valley range of 0.8 percentage points, confirming low sensitivity. For λ_2, the rage spans from 70.2% (λ_2 = 0.1) to 70.9% (λ_2 = 1.0), a range of 0.7 percentage points. For λ_3, a clearer optimum is observed: performance rises from 69.4% (λ_3 = 0.1) to 70.9% (λ_3 = 0.5), then declines to 69.7% (λ_3 = 2.0), indicating that over-penalizing cross-modal redundancy degrades complementarity. The plateau region for λ_3 lies between 0.3 and 0.8, within which performance remains above 70.5%. These results demonstrate that the adopted values are located at or near the global optimum on the validation set, and that the model is robust to moderate perturbations of all three coefficients. Practitioners may adopt the default values directly without re-tuning for similar agricultural detection scenarios.

Table 3. Hyperparameter sensitivity analysis for λ^1, λ^2 and λ^3. Each row varies one coefficient while fixing the other two at their default values. Bold indicates the adopted configuration.

Varied Param.	0.1	0.5	1.0 (default)	1.5	2.0	Range (pp)
λ_1(mAP@.5)	70.1	70.6	**70.9**	70.7	70.4	0.8
λ_2(mAP@.5)	70.2	70.5	**70.9**	70.6	70.3	0.7
λ_3(mAP@.5)	69.4	70.7	**70.9**	70.3	69.7	1.5

4.8 Cross-Dataset Generalization on the Small Apple Dataset

To evaluate the generalization capability of ***BiGCFormer*** beyond the raspberry domain, we conduct additional experiments on the publicly available Small Apple Dataset. This dataset covers multiple complex agricultural scenarios including dense fruit distribution, scale variation, and partial occlusion, providing a complementary testbed to our raspberry benchmark. Preprocessing and training protocols followed the same pipeline described in Section 4.3: images are resized to 640 × 640, depth maps are

generated using MiDaS (Monocular Depth Estimation for Arbitrary Scenes), and the AdamW optimizer with cosine annealing is applied for 300 epochs. All compared methods are retrained or fine-tuned under identical conditions to ensure a fair comparison. It is worth noting that the quality of MiDaS-generated depth maps may differ between the two datasets due to inherent differences in fruit morphology. Apples possess more regular, convex, and geometrically consistent shapes compared to raspberries, whose compound drupelet structure introduces irregular surface textures and higher intra-class shape variability. Such geometric regularity tends to produce smoother depth estimation with fewer artifacts, as MiDaS relies on learned monocular cues that respond more predictably to well-defined object boundaries and uniform surface reflectance. Consequently, the depth maps on the Small Apple Dataset are generally of higher structural consistency than those on the Raspberry Dataset. While this discrepancy may confer a marginal advantage to depth-assisted methods on the apple benchmark, it simultaneously demonstrates that ***BiGCFormer*** remains robust and competitive even when the quality of auxiliary depth input varies across domains. This cross-dataset asymmetry thus provides additional evidence of the model's adaptability rather than undermining the validity of the comparison.

Results are presented in Table 4. Among unimodal YOLO baselines, YOLOv12l achieves the highest mAP@50 of 84.2%, while YOLOv12s reaches 83.1%. Among multimodal fusion methods, all RGB + Depth approaches exhibit lower mAP@50 (ranging from 64.2% to 68.8%) than unimodal YOLO variants, suggesting that the pseudo-depth quality and the absence of fine-tuned fusion architectures limit these baselines on the apple domain. ***BiGCFormer*** achieves an mAP@50 of 87.6%, a Precision of 85.1%, and a Recall of 82.3%, surpassing the strongest unimodal baseline (YOLOv12l) by 3.4 percentage points in mAP@50. These results confirm that the MAPE-Bi-CA-Bi-Gate pipeline transfers effectively to a distinct fruit category, demonstrating the cross-crop generalization of ***BiGCFormer***.

Table 4. Cross-dataset generalization results on the Small Apple Dataset.

Model	Modality	mAP@50(%)	Precision(%)	Recall(%)
CFT (2021)	RGB + Depth	66.3	42.0	63.7
DIDFUSE (2020)	RGB + Depth	64.2	40.5	61.5
SDNET (IJCV'21)	RGB + Depth	65.5	42.2	64.1
RFNET (CVPR'22)	RGB + Depth	67.0	43.0	65.5
TARDAL (CVPR'22)	RGB + Depth	66.6	42.7	65.8
DEFUSION (MM'22)	RGB + Depth	66.9	43.3	65.2
CDDFUSE (CVPR'23)	RGB + Depth	67.1	43.4	66.0
IGNET (MM'23)	RGB + Depth	67.3	43.5	66.3
SUPERFUSION (JAS'22)	RGB + Depth	67.0	43.2	66.0
FMB (2024)	RGB + Depth	68.8	43.1	66.8

Table 4. *Continued.*

Model	Modality	mAP@50(%)	Precision(%)	Recall(%)
YOLOv5s	RGB	78.9	76.1	72.4
YOLOv8s	RGB	80.5	78.8	74.1
YOLOv11s	RGB	81.2	79.5	74.9
YOLOv12n	RGB	82.4	80.6	76.3
YOLOv12s	RGB	83.1	81.2	77.0
YOLOv12l	RGB	84.2	82.5	78.9
RT-DETR	RGB	80.1	78.4	73.5
DEIMv2	RGB	81.3	79.2	74.8
Ours (BiGCFormer)	RGB + Depth	**87.6**	**85.1**	**82.3**

4.9 Detection Visualization for Different Modal Data

To thoroughly verify the detection mechanisms and performance advantages of the ***BiGCFormer*** framework in complex agricultural scenarios, intuitive evidence for understanding the perceptual characteristics of each modality and the practical effects of innovative mechanisms of ***BiGCFormer*** are provided through qualitative visualization analysis. Detection performance of depth unimodal, RGB unimodal, and RGB-D fusion methods under different challenging scenarios is systematically compared. As shown in Figure 8, visualization results cover typical agricultural environments, including fruit occlusion and dense target distribution.

Unique values and inherent limitations of geometric information in fruit recognition are revealed by depth modality detection results, as shown in the left column of Figure 8. In samples 101, 225, and 226, partial fruit targets are successfully identified by depth detectors, with confidence scores reaching 0.67, 0.68, and 0.67 respectively, indicating the effectiveness of depth information in providing spatial geometric structures. Distance differences between fruits and backgrounds are directly encoded by depth features, forming natural foreground-background separation, with particularly excellent performance in scenarios where fruits protrude from foliage planes.

However, limitations of the depth modality remain significant. In sample 222, depth detectors are observed to fail completely, with no fruit targets detected. This phenomenon arises from inherent characteristics of depth data such as geometric sparsity, texture absence, and noise sensitivity. To the characteristic of geometric sparsity, depth differences between raspberry fruits and surrounding foliage are often small (typically <5 cm). To the characteristic of texture absence, depth images lack color and texture information, preventing utilization of distinctive red characteristics of raspberries for discrimination and causing difficulties in distinguishing mature from immature fruits. To the characteristic of noise sensitivity,

the MiDaS-generated pseudo-depth maps may exhibit reduced accuracy at long distances and blurred boundaries at occluded edges, as monocular depth estimation inherently introduces uncertainty in these regions. These characteristics differ fundamentally from the noise profiles of physical depth sensors (e.g., multi-path reflection in structured light or range ambiguity in ToF cameras), and should be considered when transferring the proposed method to real RGB-D sensor deployments.

From the perspective of information theory, the information entropy of the depth modality remains relatively low. Advantages of geometric information encoding in spatial structure understanding, alongside insufficiencies in semantic discrimination, are exhibited by this modality. The poor performance of depth unimodal methods in agricultural fruit detection tasks is explained by this characteristic, while a theoretical foundation for the necessity of multimodal fusion is also provided.

Core roles of color-texture information in fruit recognition are demonstrated by RGB modality detection results, as shown in the middle column of Figure 8. In all test samples, major fruit targets are successfully identified by RGB detectors, with generally high confidence scores ranging from 0.95 to 1.03. This performance is significantly superior to that of the depth modality. Particularly outstanding detection results are observed in samples 222, 225, and 227, where multiple fruit targets including partially occluded small targets are accurately localized by the RGB modality.



Figure 8. Comparison of detection performance of different methods for raspberry recognition.

Advantages of the RGB modality are primarily manifested in three aspects, namely, rich semantic information, texture discrimination capability, and illumination adaptability. In the aspect of rich semantic information, the red characteristics of raspberries form significant color contrasts in RGB images, which facilitates the distinction between raspberries and green foliage backgrounds. In the aspect of

texture discrimination capability, the granular surface textures of mature raspberries provide important discriminative features, assisting in distinguishing raspberries from other red objects such as flowers and insects. And in the aspect of illumination adaptability, RGB sensors possess a good dynamic range under normal illumination conditions, enabling the capture of rich color details.

However, obvious limitations of the RGB modality are also observed. In samples 101 and 226, RGB detectors are observed to have missed detections when processing severely occluded targets. When fruits are subjected to large-area foliage occlusion, the area of visible fruit regions decreases and color information becomes incomplete, leading to a decline in detector confidence. Furthermore, color information in RGB images is easily distorted under strong illumination or shadow environments, which affects detection accuracy. From a theoretical perspective, significant shifts occur in the RGB feature distribution ($P(f_{rgb}) \approx \mathrm{N}(\mu_{rgb} + \Sigma_{rgb})$) during illumination changes, making it difficult for traditional unimodal methods to adapt to such dynamic variations.

Significant advantages of the ***BiGCFormer*** framework are clearly demonstrated by RGB-D fusion detection results, as shown in the right column of Figure 8. In all test samples, optimal detection performance is achieved by fusion methods. They not only outperform unimodal methods in the number of detected targets but also exhibit excellent localization accuracy and confidence scores. Particularly noteworthy is that in sample 222, small target fruits missed by both RGB and depth unimodal methods are successfully detected by RGB-D fusion, which embodies the value of complementary information between modalities.

The advantages of fusion detection are specifically manifested in the following aspects:

(1) Complementary enhancement effects: RGB and depth modalities possess natural complementarity. In sample 101, although depth information provides fruit geometric contours, semantic discrimination capabilities remain lacking; RGB information provides color-texture features, yet information remains incomplete under occlusion scenarios. ***BiGCFormer*** demonstrates improved performance that is consistent with the bidirectional cross-attention design of the Bi-CA module while utilizing RGB texture information to compensate depth data sparsity, forming synergistic effects where $1 + 1 > 2$.

(2) Robustness improvement: in samples 225 and 227 with significant illumination variations, fusion methods demonstrate stronger environmental adaptability. When RGB modality experiences illumination interference, depth information provides stable geometric constraints. When depth modality encounters noise, RGB information supplements reliable semantic discrimination. This robustness under variable illumination conditions is consistent with the design intent of the Bi-Gate module, whose adaptive weight adjustment mechanism is designed to dynamically optimize fusion strategies according to modal quality. The fusion weight distributions visualized in Figure 7 provide direct evidence of this dynamic behavior, and its quantitative contribution was confirmed in the ablation study.

(3) Fine localization capability: compared with unimodal methods, fusion detection demonstrates higher accuracy in bounding box localization, consistent with the design objective of the MAPE module. While this qualitative observation suggests that modal-specific positional encoding contributes to improved

spatial correspondence between RGB and depth features, the quantitative contribution of MAPE to localization accuracy has been independently validated through ablation experiments, which showed a measurable drop in mAP@.5:95 when MAPE was removed. Together, these results support the conclusion that accurate geometric alignment between modalities contributes to more precise target localization.

5. CONCLUSION

BiGCFormer, a Bi-stream graph cross-attention fusion network, is proposed to address fruit detection challenges in complex agricultural environments. By integrating RGB and depth information through MAPE, Bi-CA, and Bi-Gate modules, the model enhances cross-modal feature interaction and improves detection robustness.Experiments demonstrate that ***BiGCFormer*** achieves superior performance compared to existing methods, reaching a mAP@0.5 of 70.9% on the raspberry dataset. Ablation studies further confirm the effectiveness of each module.

Despite these results, several limitations remain. The method relies on pseudo-depth maps rather than physical RGB-D sensors, which may introduce domain gaps in real-world deployment, particularly in scenarios with complex lighting and reflective surfaces.

Future work will focus on validating the model with real RGB-D data and extending its applicability to broader agricultural scenarios.

ACKNOWLEDGMENTS

This work was in part supported by the Natural Science Foundation of Qinghai Province, China (Grant: 2022-ZJ-925), National Natural Science Foundation of China (Grant: 62066039), the '111'Project, China (Grant: D20035). We would like to express our sincere gratitude for the insightful feedback provided by the editor and the anonymous reviewers.

REFERENCES

[1] G. Cheng *et al.*, "Towards large-scale small object detection: Survey and benchmarks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 13467–13488, Dec. 2023.

[2] C. Y. Wang, I. H. Yeh, and H. Y. M. Liao, "YOLOv9: Learning what you want to learn using programmable gradient information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5432–5448, Aug. 2024.

[3] C. Li *et al.*, "YOLOv6: A single-stage object detection framework for industrial applications," *Comput. Vis. Image Underst.*, vol. 231, Art. no. 103666, Jun. 2023.

[4] Y. Zhao *et al.*, "DETRs beat YOLOs on real-time object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 9, pp. 6789–6805, Sep. 2024.

[5] H. Zhang *et al.*, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," *Int. J. Comput. Vis.*, vol. 131, no. 10, pp. 2507–2524, Oct. 2023.

[6] K. Chen *et al.*, "Hybrid task cascade for instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 5, pp. 5234–5249, May 2023.

[7] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "SOLOv2: Dynamic and fast instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 3, pp. 1456–1470, Mar. 2024.
[8] D. Zhou *et al.*, "IoU loss for 2D/3D object detection," *Int. J. Comput. Vis.*, vol. 131, no. 12, pp. 3100–3120, Dec. 2023.
[9] S. Liu *et al.*, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," *Int. J. Comput. Vis.*, vol. 131, no. 6, pp. 1391–1408, Jun. 2023.
[10] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[11] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: A simple and strong anchor-free object detector," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 1922–1933, Apr. 2023.
[12] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted windows," *Int. J. Comput. Vis.*, vol. 131, no. 4, pp. 803–822, Apr. 2023.
[13] A. Dosovitskiy *et al.*, "An image is worth 16 × 16 words: Transformers for image recognition at scale," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2793–2819, Oct. 2021.
[14] Z. Li *et al.*, "BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," *Int. J. Comput. Vis.*, vol. 131, no. 8, pp. 1939–1958, Aug. 2023.
[15] H. Zhang *et al.*, "ResNeSt: Split-attention networks," *Comput. Vis. Image Underst.*, vol. 229, Art. no. 103635, Apr. 2023.
[16] H. Fan *et al.*, "Multiscale vision transformers," *Int. J. Comput. Vis.*, vol. 131, no. 4, pp. 824–847, Apr. 2023.
[17] J. Chen *et al.*, "TransUNet: Transformers make strong encoders for medical image segmentation," *Med. Image Anal.*, vol. 82, Art. no. 102559, Aug. 2021.
[18] J. Dai *et al.*, "Deformable convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 2, pp. 2234–2250, Feb. 2023.
[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8123–8138, Jul. 2023.
[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9012–9025, Aug. 2023.
[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 123–138, Jan. 2023.
[22] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," *Int. J. Comput. Vis.*, vol. 131, no. 3, pp. 567–585, Mar. 2023.
[23] P. Sun *et al.*, "Sparse R-CNN: End-to-end object detection with learnable proposals," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4567–4582, Apr. 2023.
[24] X. Li *et al.*, "Generalized focal loss: Learning qualified and distributed bounding boxes for dense object detection," *Pattern Recognit.*, vol. 138, Art. no. 109359, Jun. 2023.
[25] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6789–6803, Jun. 2023.
[26] Y. Li, Y. Chen, N. Wang, and Z. Zhang, "Scale-aware trident networks for object detection," *Int. J. Comput. Vis.*, vol. 131, no. 9, pp. 2156–2175, Sep. 2023.
[27] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," *Pattern Recognit.*, vol. 142, Art. no. 109678, Oct. 2023.
[28] X. Zhang, Y. Wu, and F. Wang, "FMB: Feature memory bank for RGB-D salient object detection," in Proc. *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 1–9.
[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Pattern Recognit.*, vol. 140, Art. no. 109567, Aug. 2023.

[30] C. Szegedy *et al*., "Going deeper with convolutions," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 45, no. 11, pp. 12345–12362, Nov. 2023.
[31] R. Girshick, "Fast R-CNN," *Int. J. Comput. Vis*., vol. 131, no. 7, pp. 1678–1695, Jul. 2023.
[32] L. Liu, Y. Wang, and C. Wang, "MM-YOLO: A lightweight multimodal object detector," *Comput. Vis. Image Underst*., vol. 234, Art. no. 103789, Sep. 2023.