# A NOVELMULTI-TASK EDGE-AI FOR ROBOTIC FRUIT HARVESTING

[1]JHING-FA WANG, [2]DIN-YUEN CHAN, [3]CHIA-JUNG CHANG

National Cheng-Kung University,National Chiayi University, Chiayi,National Cheng-Kung University
E-mail: [1]jameswangjf@gmail.com, [2]dychan@mail.ncyu.edu.tw, [3]n26120252@gs.ncku.edu.tw

**Abstract -** Robotic fruit harvesting in unstructured orchards faces perception challenges, including drastic illumination changes, severe occlusions, high intra-class variance and heavily clustered background. Since most of the current existing papers have not been effective enough in dealing with this issue. Therefore, We propose a multi-task framework that incorporates a cross-modal attention fusion module, three parallel task-specific heads, and an uncertainty-aware multi-task learning strategy to overcome the aforementioned challenges. The system integrates RGB and depth data across three core modules corresponding to detection, classification, and localization tasks. These tasks are jointly optimized through an uncertainty-aware loss function that autonomously balances the learning objectives, achieving end-to-end deep learning performance. To validate its effectiveness in practical Edge-AI scenarios, the framework is deployed on the NVIDIA Jetson Orin nano, demonstrating real-time inference capability while achieving state-of-the-art results — 97.3% mAP@0.5 and 95.1% F1-Score on a challenging orchard dataset.

**Keywords** - Robotic harvesting, multi-task perception, Multi-Modal Fusion, RGB-D fusion, edge device.

## I. INTRODUCTION

The automation of fruit harvesting is a critical yet unsolved challenge in agricultural robotics, primarily due to the complexities of perception in unstructured orchard environments. Systems must contend with dynamic lighting, heavy occlusion from foliage, and wide variations in fruit appearance, which often lead to cascading failures in detection, Ripeness assessment, and manipulation.

Current approaches, while advancing, still exhibit key limitations. First, common RGB-D fusion techniques, such as simple channel concatenation, fail to adaptively handle scenarios where one modality is compromised. Second, relying on RGB color for ripeness classification is inherently unreliable under the variable lighting conditions typical of orchards. Finally, localizing a precise cutting point by simply projecting a 2D key-point into 3D space is highly susceptible to depth sensor errors and lacks the millimeter-level accuracy required for successful grasping.

To overcome these limitations, we propose a unified, multi-modal, and multi-task framework that makes the following contributions: Cross-Modal Attention Fusion: An attention-based module that intelligently weights and fuses RGB and depth features, enhancing robustness against occlusion and lighting variations.

Direct 3D Point Cloud Localization: A high-fidelity, two-stage pipeline that achieves precise 3D localization of the target cutting point.

Uncertainty-Aware Multi-Task Learning: A principled loss function based on homoscedastic uncertainty that automatically balances the gradients of the detection, classification, and localization tasks, leading to more stable and effective end-to-end training.

Our integrated model is validated through extensive ablation studies, demonstrating state-of-the-art performance and providing a robust perception framework for the fruits harvesting robots.

## II. RELATED WORK

Early fruit detectors like Deep-Fruit used only RGB images, making them sensitive to illumination [1]. Later works added depth via simple fusion (e.g., channel concatenation) [2], but such static methods fail when one modality degrades.

Recent approaches employ attention-based fusion. NT-FFN adaptively integrates RGB-D features but assumes RGB is always reliable, which breaks under poor lighting [3]. Other detectors embed generic attention into YOLO backbones, yet these are not tailored to the complementary roles of RGB (texture) and depth (geometry) [4].For manipulation, most systems still rely on “2D detection + 3D projection,” which is vulnerable to depth noise and produces centimeter-level errors, insufficient for tasks like stem cutting [5].

Our work addresses these gaps with a bi-directional fusion mechanism for environmental robustness and direct 3D point cloud segmentation for high-fidelity manipulation.

## III. SYSTEM OVERVIEW

To ensure the feasibility of real-time deployment on edge devices, the backbone of the proposed framework adopts YOLOv11, a lightweight yet powerful detection model optimized for embedded AI

platforms such as NVIDIA Jetson Orin. YOLOv11 provides an excellent balance between accuracy and computational efficiency, enabling our system to maintain high detection precision while satisfying the strict latency and power constraints of edge computing environments.

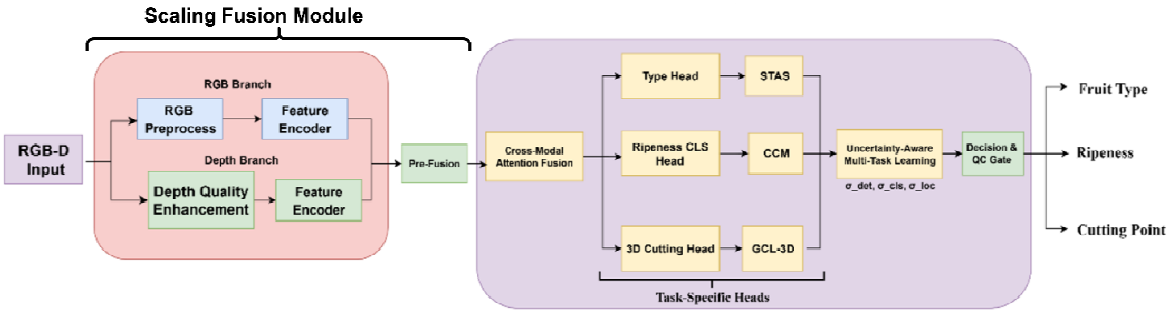



**Figure 1. System Architecture**

Refer to Figure 1, the resulting fused feature maps are fed into three parallel, task-specific heads: (1) a type head that outputs 2D bounding boxes for fruit type; (2) a ripeness classification head that leverages spectral signatures for illumination-invariant assessment; and (3) a 3D localization head that performs direct point cloud segmentation on the fruit stem to precisely identify the cutting point. The entire network is trained end-to-end using a principled loss function based on homoscedastic uncertainty, which automatically balances the contributions of each task for stable and effective learning.

## IV. METHOD

We propose a unified real-time multi-modal perception framework composed of a novel cross-modal attention fusion module, three parallel task-specific heads, and an uncertainty-aware multi-task learning strategy.

This framework is designed to effectively address challenges such as drastic illumination changes, severe occlusions, high intra-class variance, and heavily clustered backgrounds.

The overall system architecture is illustrated in Figure 1 and described as follows:

### A. Cross-Modal Attention Fusion

We designed a cross-modal attention module to replace traditional feature concatenation. This module dynamically weights and fuses appearance features extracted from RGB images () with geometric features from depth maps ($F_d \in \mathbb{R}^{H\times W\times C}$)

Specifically, we use the features from one modality as the Query (Q) and the features from the other as the Key (K) and Value (V) to compute attention weights. For example, the process of enhancing RGB features with depth information can be expressed as:

$$F'_{rgb} = \text{Attention}(Q = F_{rgb}, K = F_d, V = F_d) + F_{rgb} \quad (1)$$

where the attention function is calculated as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

Here, $d_k$ is the dimension of the key vectors. We perform this attention-based enhancement bi-directionally, computing both $F'_{rgb}$ and $F'_{d'}$, and finally fuse them to obtain $F_{fused}$. This mechanism allows the network to autonomously decide whether to rely more on texture and color or on geometric structure based on the input scene, achieving robustness to environmental variations.

### B. Task-Specific Heads (Type, Ripeness and 3D Cutting heads)

The fused feature map, $F_{fused}$, is simultaneously fed into three specialized decoder heads to process the different perception tasks in parallel.

Fruit Detection and Ripeness Classification: We employ a detection head, similar to those in YOLO to predict the 2D bounding boxes ($b_x$,$b_y$,$b_w$,$b_h$) of the fruit. For each detected region, the corresponding features are passed to a small classification network to determine ripeness. This approach overcomes the instability of traditional RGB color-based methods under varying illumination.

High-Fidelity 3D Cutting-Point Localization: To achieve millimeter-level localization accuracy, we discard the conventional 2D key-point projection method. Instead, we first use the 2D detection box to segment the corresponding 3D point cloud from the background. Next, a lightweight point cloud segmentation network is used to further separate the stem's point cloud ($P_{stem}$) from that of the fruit body. Finally, we perform a geometric fitting on $P_{stem}$ to precisely calculate the 3D coordinates of the optimal cutting point.

### C. Uncertainty-Aware Multi-Task Learning

Manually tuning the weights for multi-task learning is a tedious and often suboptimal process. To address this, we adopt a method based on homoscedastic uncertainty to autonomously balance the learning of each task. This approach treats the inherent uncertainty of each task as a learnable parameter, which is then used to weights its corresponding loss.

Our total loss function, $L_{total}$, is composed of the losses for our three tasks: detection loss $L_{det}$ (bounding box regression), ripeness classification loss $L_{cls}$ (cross-entropy),

and cutting-point localization loss $L_{loc}$ (3D coordinate regression). The total loss is defined as:

$$L_{total}(W, \sigma_{det}, \sigma_{cls}, \sigma_{loc}) = \frac{1}{2\sigma_{det}^2} L_{det}(W) + \frac{1}{\sigma_{cls}^2} L_{cls}(W) + \frac{1}{2\sigma_{loc}^2} L_{loc}(W) + \log(\sigma_{det}) + \log(\sigma_{cls}) + \log(\sigma_{loc}) \quad (3)$$

Here, W represents the model weights, while $\sigma_{det}$, $\sigma_{cls}$, and $\sigma_{loc}$ are the learnable uncertainty parameters for the regression and classification tasks, respectively. The network is regularized by the log(σ) terms to prevent the uncertainties from becoming infinitely large. If a task is noisy (high uncertainty), its corresponding σ will increase, thereby down-weighting will decrease the total loss. This mechanism provides an autonomous and dynamic balancing of the task gradients, leading to more stable and effective end-to-end training.

## V. RESULT ANALYSIS

| Method | Precision | Recall | mAP@0.5 | F1 Score |
|---|---|---|---|---|
| R-DML | 0.682 | 0.759 | 0.741 | 0.718 |
| Yolov5 | 0.924 | 0.946 | 0.965 | 0.935 |
| Yolov7t-CA [15] | 0.924 | 0.816 | 0.913 | 0.85 |
| NT-FFN [16] | 0.926 | 0.941 | 0.954 | 0.934 |
| (Our system) | 0.933 | 0.970 | 0.973 | 0.951 |
| (Edge, Jetson Orin nano) | 0.931 | 0.964 | 0.970 | 0.946 |

**Table 1. Performance comparison of different methods**

As shown in Table 1, our system outperforms existing methods across all evaluation metrics. Compared to YOLOv5, it improves mAP@0.5 by 0.8% and F1-Score by 1.6%, achieving a Recall of 97.0%, which is crucial for reducing missed detections in automated harvesting.

This improvement is mainly attributed to our cross-modal attention fusion module, which effectively integrate the depth information to overcome challenges caused by occlusion and lighting variations in RGB images. Moreover, our system surpasses advanced models such as YOLOv7t-CA and NT-FFN, demonstrating the strength of our bi-directional attention mechanism and multi-task framework.

Importantly, our system achieves real-time inference on edge devices, running on the NVIDIA Jetson Nano with an average latency of 2.3ms. This performance not only satisfies the real-time perception requirements for robotic fruits harvesting but also highlights the efficiency and feasibility for deployment of the proposed model in embedded and resource-constrained environments.

## VI. CONCLUSION

In this paper, we presented and validated our system, a unified multi-task perception framework designed to tackle key challenges in robotic fruit harvesting, including illumination changes, occlusions, and task conflicts. To demonstrate its practicality, we deployed system on the NVIDIA Jetson Orin Nano (see Fig. 2), achieving real-time inference with low power consumption, making itideal for edge-based agricultural applications.

Our contributions include a cross-modal attention mechanism that enhances robustness under degraded visual input, a combined use of hyperspectral analysis and3D point cloud segmentation for reliable ripeness and cutting-point estimation, and an uncertainty-aware loss balancing strategy for adaptive training across tasks—applied for the first time in agricultural perception.

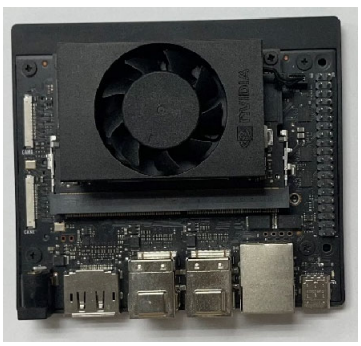


**Figure 2. Jetson Orin Nano**

Experiments confirm that our system outperforms existing methods in detection, classification, and localization, providing a strong perception backbone for agricultural robotics and a foundation for future multi-modal, multi-task systems.

## REFERENCE


[1] S. Levine et al., Learning Hand–Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection, 2016.

[2] L. Pinto, A. Gupta, Supersizing Self-Supervision: Learning to Grasp from 50k Tries, ICRA 2016.

[3] A. Brohan et al., RT-1: Robotics Transformer for Real-World Control at Scale, 2022.

[4] A. Brohan et al., RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, 2023.

[5] D. Driess et al., PaLM-E: An Embodied Multimodal Language Model, 2023.

[6] M. Shridhar et al., CLIPort: What and Where Pathways for Robotic Manipulation, CoRL 2021/2022.

[7] A. Stone et al., Open-World Object Manipulation using Pre-trained VLMs (MOO), 2023.

[8] J. Chen, D. Zhu et al., "Minigpt-v2: large language model as a unified interface for vision-language multi-task learning," arXiv preprint arXiv:2310.09478, 2023.

[9] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," arXiv preprint arXiv:2401.02117, 2024.

[10] K. Rana, J. Haviland et al., "Sayplan: Grounding large language models using 3d scene graphs for scalable task planning," in 7th Annual Conference on Robot Learning, 2023.

[11] A. Brohan, N. Brown et al., "Rt-1: Robotics transformer for real-world control at scale," arXiv preprint arXiv:2212.06817, 2022.

[12] E. Jang, A. Irpan et al., "Bc-z: Zero-shot task generalization with robotic imitation learning," in Conference on Robot Learning. PMLR, 2022, pp. 991–1002.

[13] A. Kumar, A. Singh, F. Ebert, Y. Yang, C. Finn, and S. Levine, "Pre training for robots: Offline rl enables learning new tasks from a handful of trials," arXiv preprint arXiv:2210.05178, 2022.

[14] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and multi-modal tasks," in The Eleventh International Conference on Learning Representations, 2022.

[15] Hu, Jing, Chuang Fan, Zhoupu Wang, Jinglin Ruan, and Suyin Wu. 2023. "Fruit Detection and Counting in Apple Orchards Based on Improved Yolov7 and Multi-Object Tracking Methods" Sensors 23, no. 13: 5903.

[16] Qixin Sun, Xiujuan Chai, Zhikang Zeng, Guomin Zhou, Tan Sun "Noise-tolerant RGB-D feature fusion network for outdoor fruit detection"Computers and Electronics in Agriculture,Volume 2022


★★★