ORIGINAL RESEARCH **Open access**

Lightweight Banana Detection and 3D Localization for Agricultural Robots Using Deep CNNs

Olivia Bennett[1*], Harry Collins[2], Jack Foster[1], Ethan Wright[3]

Abstract

Reliable identification and spatial positioning of fruit within natural settings is an essential capability for robotic harvesting systems to execute precise picking. Existing approaches for detecting and locating bananas, however, suffer from two principal drawbacks in real-world use: models contain an excessive number of parameters that complicate deployment, and their overall performance still requires enhancement. To address these challenges, an accurate yet lightweight framework for banana bunch detection and spatial positioning was developed and implemented on edge hardware. Initially, a Slim-Banana architecture was designed by modifying YOLOv8l. To reduce computational burden while maintaining strong performance, GSConv was incorporated into Slim-Banana to replace conventional convolution, along with grouped and spatial convolutions. At the same time, a cross-stage local network (GSCSP) module was constructed, reducing computational cost and structural complexity through a single-stage aggregation scheme. Subsequently, a RealSense depth sensor was paired with TOF technology to perform image alignment and three-dimensional banana positioning. Finally, the entire processing pipeline was deployed onto an Nvidia Orin NX edge device, and its operational performance, together with resource utilisation under actual working conditions, was comprehensively examined. Experimental outcomes show that detection precision, recall, mAP, and inference time are 0.947, 0.948, 0.98, and 113.6 ms, respectively; the network requires 4449 MiB of memory; and the mean localisation errors along the X-, Y-, and Z-axes are 13.47 mm, 12.87 mm, and 13.87 mm, respectively. To our knowledge, this is the first study to perform banana detection and localisation directly on edge devices. Results confirm that, relative to prior methods, the proposed approach delivers superior performance across complex orchard scenes, achieving both efficient and lightweight banana recognition and localisation.

Keywords Lightweight, Banana, Visual detection, 3D localization, YOLOv8

*Correspondence:
Olivia Bennett
olivia.bennett@nottingham.ac.uk
[1] Department of AI for Root Crop (Carrot & Parsnip) Digging and Sorting Robotics, Faculty of Agriculture, University of Nottingham, Nottingham, United Kingdom
[2] Department of Computer Vision for Blemish and Rot Detection, Faculty of Agriculture, University of Southampton, Southampton, United Kingdom
[3] Department of Machine Learning for Soil Compaction Detection, Faculty of Agriculture, Newcastle University, Newcastle, United Kingdom

Introduction

Bananas rank among the most economically significant agricultural products worldwide, with output reaching 127.3 million tons, making this the highest-yielding fruit crop on the planet [1, 2]. Harvesting is currently carried out largely by hand, a practice marked by low efficiency and elevated costs. As technology advances, interest in intelligent mechanical harvesting solutions continues to rise [3]. The visual perception system, sitting at the heart of agricultural robots tasked with harvesting, plays a critical role in enabling both reliable yield estimation and effective collection operations [4].

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Formulating an accurate, lightweight banana-detection and positioning method would substantially improve the effectiveness and precision of robotic fruit pickers [5]. That said, current banana identification techniques have shortcomings, including limited accuracy and slow detection speed, which pose a major bottleneck to the intelligent evolution of orchards. Consequently, engineering a lightweight and precise banana identification solution represents one of the central challenges in orchard intelligent management.

Scholars have devoted considerable research to fruit identification. Fu *et al.* [6] put forward a scheme combining support vector machines with banana local binary pattern features and histogram of oriented gradients for banana detection. That method delivered 89.63% accuracy with a mean execution duration of 1.325 s. He *et al.* [7] presented a deep bounding box regression forest (DBBRF)-based approach for detecting citrus green fruits, achieving an average recognition accuracy of 87.6% and a mean recognition time of 0.759 s. Chen *et al.* [8] described a machine vision-based method for ripe fruit detection and positioning, showing a recognition rate of 97.44% for ripe fruits and a mean processing time of 0.2966 s. Traditional machine learning techniques have seen some headway in fruit recognition. Still, owing to their reliance on hand-engineered features, they struggle with robustness issues when deployed in demanding orchard conditions.

With the rise of deep learning, a host of fruit detection studies rooted in it have emerged. Seeking to recognise banana fruits effectively, Fu *et al.* [9, 10] turned to YOLOv4 for precise detection of bananas across different scales. Their algorithm registered a detection rate of 99.29% alongside an average execution time of 0.171 s. Later, Fu *et al.* introduced a YOLO-Banana detection model recording a mean average precision (mAP) of 92.19%, a model footprint of 137 MB, and a single-frame detection time of 35.33 ms. Cai *et al.* [11] developed a YOLOV7-FM-based algorithm for banana pseudostem detection, reporting an average inference time of 8.0 ms per image and an AP of 81.45%. Neupane *et al.* [12] employed drone-captured imagery with Fast R-CNN to identify banana plants, reporting accuracies of 96.4%, 85.1%, and 75.8% across three detection zones. Wang *et al.* [13] developed a method for detecting young tomato fruits against similarly coloured backgrounds, built on an improved Faster R-CNN and an attention mechanism. The model reached a mAP of 98.46% with a per-image processing time of 0.084 s. Zheng *et al.* [14] developed a citrus detection network, Citrus-YOLO, achieving an F1 score of 89.07% and a mean absolute percentage error of 12.75%. Although deep learning models can achieve high fruit recognition accuracy, they generally require substantial computational power for both training and inference. These methods, which depend on high-spec hardware, struggle to meet the on-device deployment requirements of fruit-picking robots during real-world operations. Furthermore, most of these studies focus solely on fruit detection, whereas robotic harvesters must also obtain the target fruit's real-world spatial coordinates to perform picking.

When robots harvest fruit, the target's three-dimensional coordinates must be acquired. Current techniques rely predominantly on binocular vision, structured light, and TOF (time-of-flight) sensing for target localisation. Xiong *et al.* [15] advocated using binocular stereo matching to determine the picking point; their approach yielded a maximum depth error of 5.08 cm and a minimum of 1.96 cm. Wang [16] put forward a window-scaling-based matching localisation approach within a binocular vision framework. Across six distinct conditions, the average accurate recognition rate was 96.33%.

Nevertheless, this technique can fail to correctly locate fruit when the target vegetable is heavily occluded, particularly in more variable environments. The above studies realise target positioning through binocular vision. Yet this approach is vulnerable to lighting variations and relies heavily on ambient illumination, leading to degraded accuracy in complex settings. Moreover, binocular vision requires matching feature points across left and right image pairs, a process that entails significant computation and extended processing time. Where real-time throughput is demanded, this computational load becomes an operational limitation. The aforementioned drawbacks curtail the robustness of binocular vision-based methods in practical deployment scenarios. Zhou *et al.* [17] described a structured-light-based technique to estimate banana bunch weight in orchards and to pinpoint the stem centre point. While their method advanced orchard management meaningfully, it stopped short of addressing the demands of lightweight, real-time deployment on edge hardware. Milella *et al.* [18] estimated grape volume using the RealSense R200 depth camera. These investigations illustrate that structured light technology can facilitate three-dimensional object localisation. However, since structured

© 2024 The Author(s). | Published by Agro Intelligence Research Press

light depends on projected patterns, in orchard environments varying illumination can degrade the clarity and fidelity of those patterns, thereby compromising localisation accuracy. Yu *et al.* [19] used Kinect to capture RGB-D imagery for fruit recognition and localisation. Lin *et al.* [20] applied computer vision to detect oranges and perform their 3D positioning, reporting a depth-direction localisation error of 13.0 ± 3.0 mm. When compared with binocular vision and structured light methods, TOF-based localisation technology shows lower sensitivity to ambient illumination and is better suited for real-time measurement and dynamic scenes under diverse lighting conditions.

Building on earlier work, this paper introduces an efficient, lightweight pipeline for banana recognition and positioning that addresses the core limitations of existing techniques. This work refines the system's key components and ultimately delivers an end-to-end, cost-effective framework for fruit detection and localisation. To our knowledge, this represents the first study to realise banana detection and localisation directly on edge devices. Furthermore, relevant experimental evaluations are conducted, and the proposed method is benchmarked against other cutting-edge approaches. The primary contributions of this paper are as follows.

(1) A lightweight fruit detection algorithm, Slim-Banana, is introduced. It is compared with mainstream models, and the behaviour of the Slim-Banana network under front-lit, backlit, and occluded conditions is analysed.

(2) A RealSense depth sensor is incorporated to align RGB and depth frames, enabling 3D fruit localisation using TOF technology.

(3) The pipeline is deployed on the Nvidia Orin NX edge device, and an orchard field trial is designed and conducted to evaluate the model's on-device performance and resource consumption.

Materials and Methods

Image acquisition

The banana imagery used in this work was collected over two periods—October through November 2023 and June 2024—at a banana orchard operated by the Chinese Academy of Tropical Agricultural Sciences in Zhanjiang, Guangdong Province (21°10′N, 110°16′E). Photographs were taken by moving through the orchard and capturing each banana tree individually. A HUAWEI Mate 60 Pro camera, operated by hand, was positioned roughly 70–80 cm away from the fruit cluster for every shot. Pictures were stored as colour JPG files with a pixel resolution of 4032×3024; in total, 1131 images were compiled. A deliberate effort was made to ensure dataset variety by incorporating the effects of distinct illumination regimes and degrees of occlusion into the sensing pipeline. As depicted in **Figure 1**, the collection covers a range of real-world orchard scenarios, including direct sun, backlit conditions, heavy obstruction, and unobstructed views.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

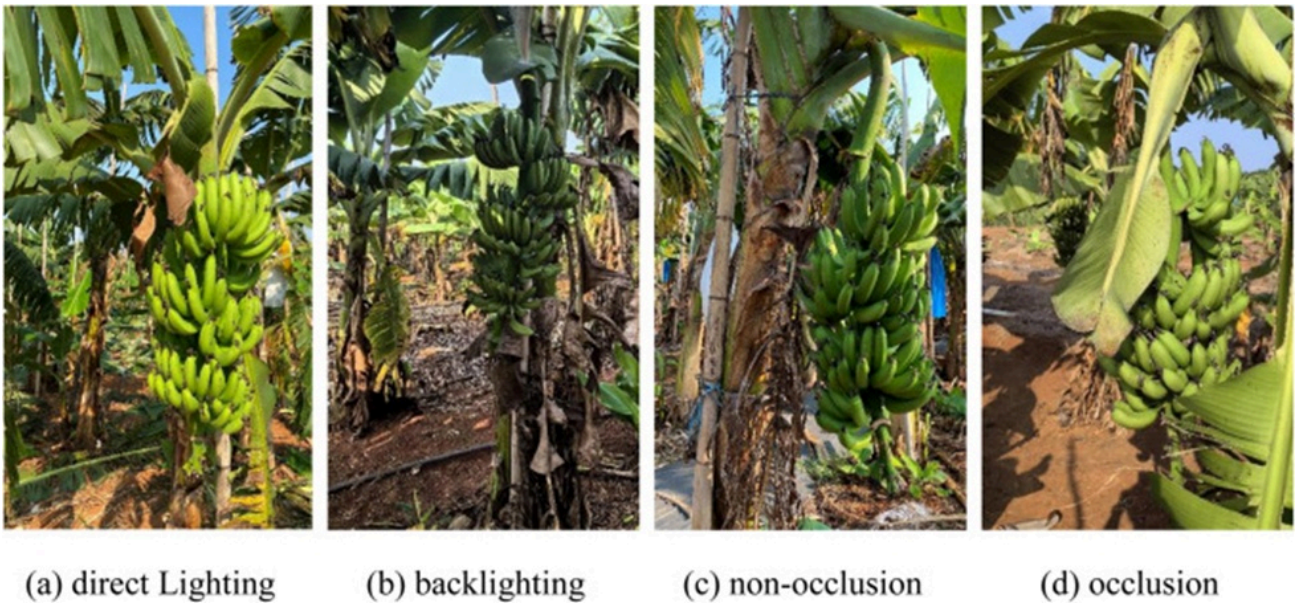

Figure 1. A picture of various fruit growing scenes in the natural environment of the orchard.

Data pre-processing

Dataset calibre plays a decisive role in shaping both the recognition model's training fidelity and its downstream utility. To maintain image clarity and representativeness, all 1131 banana photographs were carefully vetted and manually annotated using LabelImg, following the YOLO file format. To simultaneously support robust model training and meaningful validation, the image pool was split into training, validation, and test partitions at a 7:2:1 ratio. This yielded 791 training examples, 226 validation examples, and 114 test examples.

Banana visual detection method

YOLO v8 network

The perceptual backbone of a robotic banana harvester is its vision system; getting this right is critical for achieving both picking accuracy and operational speed. Current fruit detection techniques fall largely into two camps: single-stage regression-driven approaches, epitomised by the YOLO lineage [21, 22], and two-stage region-proposal-driven approaches, typified by the R-CNN family [23-26]. The two-stage paradigm is hampered by deployment difficulties and sluggish recognition rates, both of which stem from its complex anchor-box mechanics. Single-stage YOLO variants, on the other hand, perform class labelling and bounding box regression concurrently within a single unified network. This learner pipeline improves detection throughput and better aligns with real-time demands. In production environments, banana-picking robots demand a careful balance of recognition precision, model simplicity, and responsiveness. With these considerations in mind, the present study advances a banana recognition model that marries high accuracy with a lightweight footprint by refining the YOLOv8 architecture [27].

The YOLO (you only look once) family traces back to the original formulation by Redmon *et al.* [21]. Following multiple rounds of refinement, the series reached its eighth iteration in January 2023. Structurally, YOLOv8 is organised around three core blocks: a backbone for feature extraction, a neck for feature enhancement, and a detection head.

Backbone: This block stacks a succession of convolutional and deconvolutional layers, augmented with residual connections and bottleneck designs, to compress network scale while elevating representational power. The fundamental building unit here is the C2f module. Compared to the older C3 module in YOLOv5, C2f has a lower parameter count and

© 2024 The Author(s). | Published by Agro Intelligence Research Press

delivers richer feature extraction. YOLOv8 also weaves in a Spatial Pyramid Pooling Fast (SPPF) module to sharpen detection across objects of disparate sizes.

Neck: The Neck segment implements a PAN-FPN topology, weaving cross-layer fusion links between upward and downward sampling pathways. This bolsters the model's ability to merge features from objects at different scales. Concretely, upper network tiers gain access to broader contextual information, while lower tiers safeguard fine spatial details, together yielding optimised feature extraction.

Head: The decision-making centre of the detection model, the Head network generates final outputs. It comprises a detection head and a classification head. The detection head produces bounding-box predictions and associated confidence scores via a sequence of convolutional and deconvolutional operations. The classification head outputs class probability distributions, equipping YOLOv8 to tackle multi-class recognition tasks efficiently.

Slim-Banana network

To reconcile the competing demands of high precision and low model mass for banana fruit detection, this paper introduces Slim-Banana, a detector derived from an enhanced YOLOv8. As shown in **Figure 2**, the model's original neck is replaced with SlimNeck [28], a construct built on GSConv and VoV-GSCSP modules. In this arrangement, the standard backbone feeds directly into the SlimNeck structure. The modification accelerates runtime, reduces computational demand, and produces a model that is both lightweight and accurate—well-suited to the real-time constraints of commercial orchard operations.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

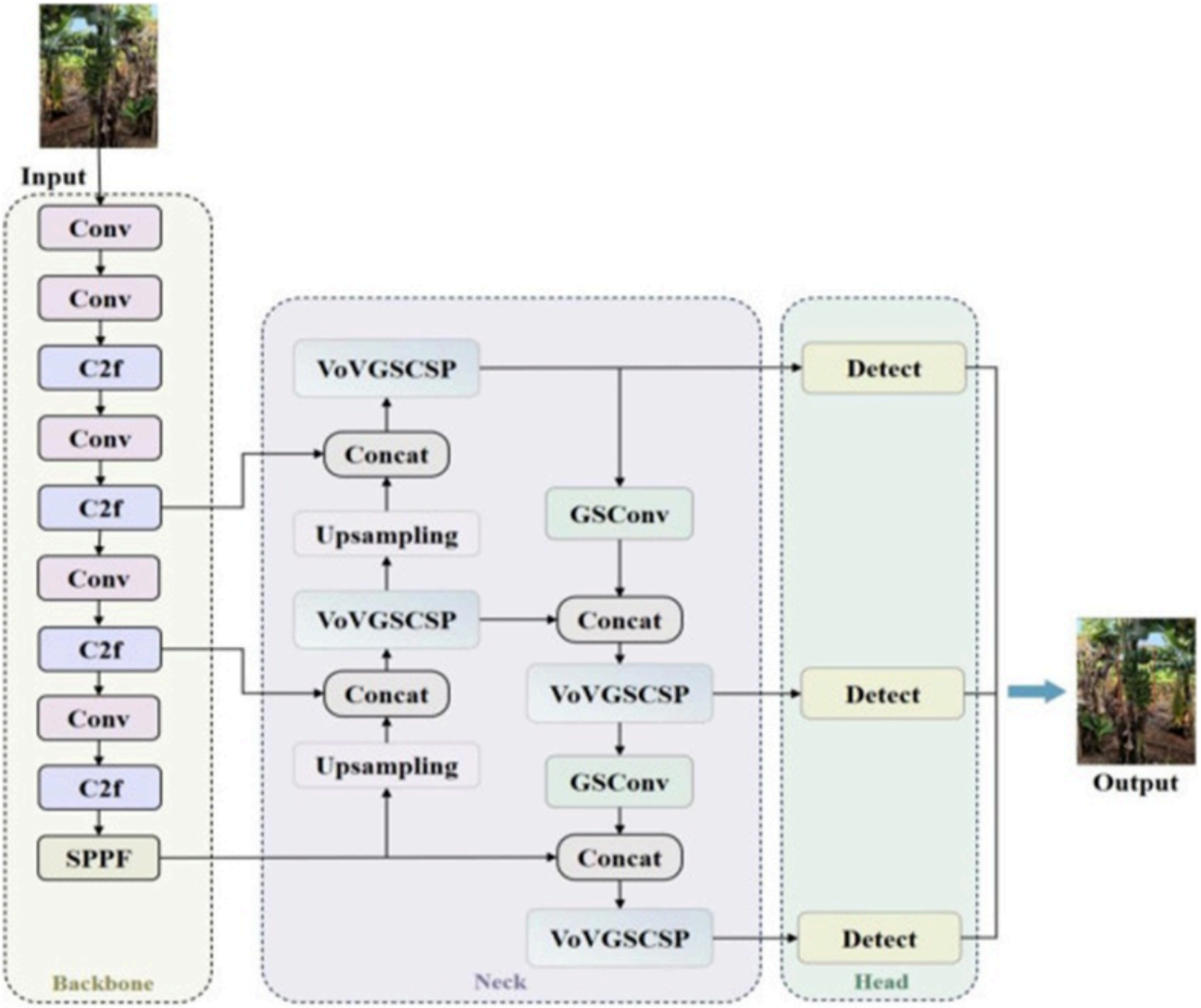

Figure 2. Overview of the Slim-Banana network, with SlimNeck replacing the original neck using GSConv and VoV-GSCSP modules.

Slimneck

During live harvesting tasks, the twin realities of limited on-farm computational resources and the necessity of edge-device deployment motivate the incorporation of GSConv (Grouped Shuffle Convolution) and Slimneck designs. This paper proposes a feature fusion module rooted in SlimNeck that simultaneously reduces model heft and improves detection accuracy, making it better suited to authentic orchard conditions.

Within a CNN, input imagery typically undergoes a progressive transformation as it passes through the Backbone: spatial detail is steadily traded for channel depth. As feature maps advance from one layer to the next, the spatial footprint of the input shrinks while the channel dimension expands. Each such transition risks discarding fragments of semantic information. Dense convolutions preserve these inter-channel connections to the greatest extent possible; sparse convolutions sever them. GSConv is a convolutional formulation specifically designed to preserve channel-wise connections. As shown in **Figure 3**, it curtails computation by marrying grouped convolution with spatial convolution. This design maintains feature-map diversity while improving computational efficiency. Deploying GSConv indiscriminately at every stage, however, would deepen the layer stack and inflate inference latency. Once feature maps enter the Neck, the

© 2024 The Author(s). | Published by Agro Intelligence Research Press

channel count has already peaked, and the spatial resolution has bottomed out; further transformations are unnecessary. It is therefore optimal to restrict GSConv to the Neck stage.

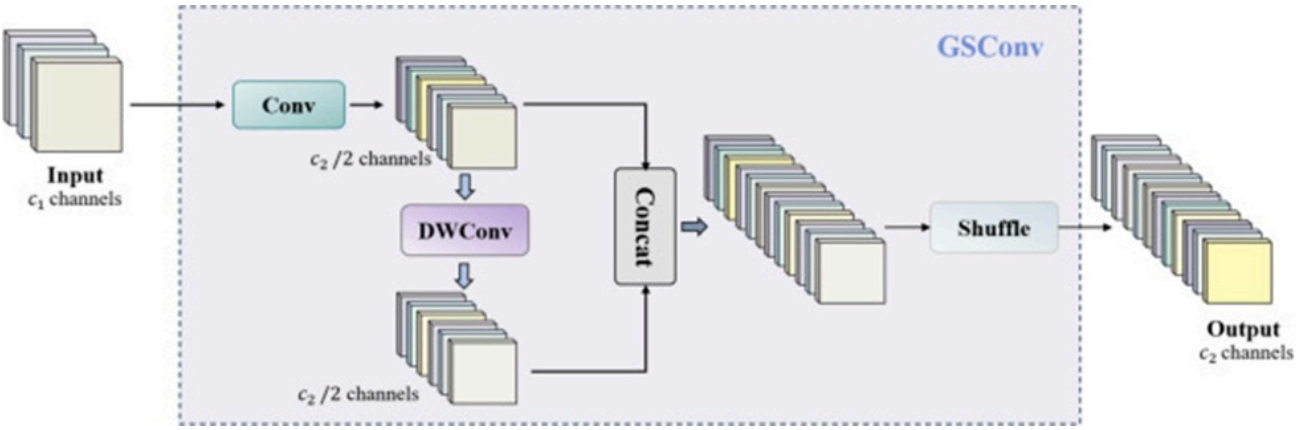

Figure 3. Overview of the GSConv module.

Along the same lines, a one-shot aggregation strategy is employed to construct the cross-level partial network (GSCSP) module, VoV-GSCSP, as illustrated in **Figure 4**. Embedding the VoV-GSCSP module reduces both computational load and architectural complexity while maintaining acceptable accuracy. Thus, in the present work, the standard CBS convolution blocks and C2F modules in the original model's neck are replaced with GSConv and VoV-GSCSP modules, which serve as feature-processing units.

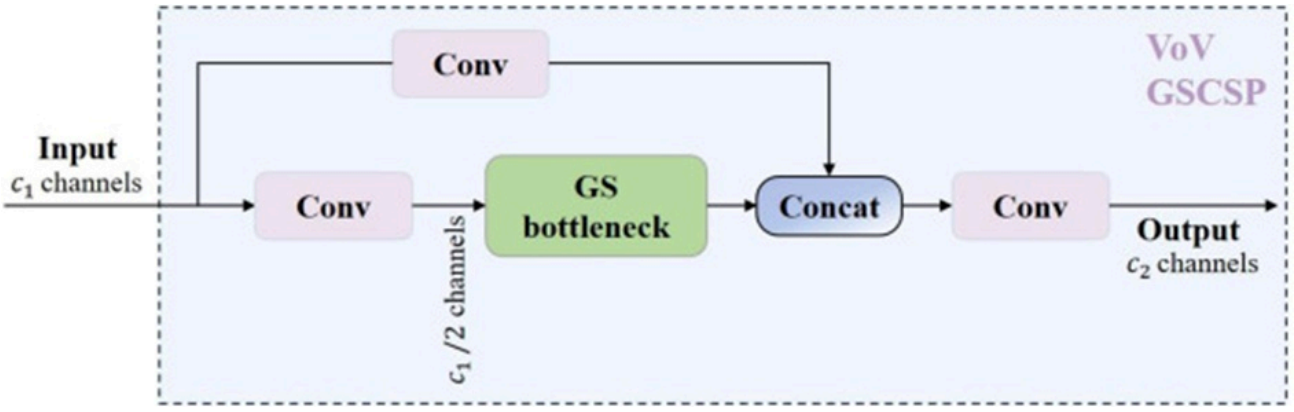

Figure 4. Overview of the VoV-GSCSP1 module.

Without sacrificing detection fidelity, the proposed approach reduces the model size and inference time. The resulting method is decidedly better matched to resource-constrained orchard environments and the practical requirement of terminal-device deployment.

3D localization method of the banana bunch

This paper leverages TOF (Time-of-Flight) technology to perform precise localisation of banana bunches. Using a three-dimensional imaging principle, distance is derived by clocking the interval between light emission and its return after reflection, from which the three-dimensional spatial coordinates of the banana fruit are extracted. The localisation pipeline is detailed below in two steps: capturing depth information and computing three-dimensional coordinates.

Depth information acquisition

© 2024 The Author(s). | Published by Agro Intelligence Research Press

A short-duration light pulse—typically from an invisible infrared or laser source—is fired from the TOF sensor. When this pulse intercepts a banana fruit, it is reflected and travels back towards the sensor along its incident path. A receiver on the sensor captures the returning pulse and, with the aid of an onboard high-precision timer, records the total round-trip interval t. The timing must resolve down to nanosecond or even picosecond granularity. Given the speed of light c, the separation between the banana and the sensor is calculated. The governing formula is given below:

$$d = \frac{c \cdot t}{2} \quad (1)$$

Here, d signifies the distance between the banana fruit and the sensor, c stands for the speed of light, and t is the full duration from pulse emission to return reception. By obtaining the distance at each pixel position, one can produce a two-dimensional image encoding depth values—the depth image. In such an image, each pixel's value corresponds to the distance from the surface point to the sensor. Compared with structured-light schemes and binocular stereo setups, TOF technology remains largely immune to ambient illumination and can operate stably in the orchard's highly variable lighting environment.

3D coordinate calculation

RGB imagery and depth data are generated by distinct sensors possessing dissimilar resolutions and optical perspectives. It is therefore essential to perform image alignment, guaranteeing that depth measurements and chromatic information correspond precisely within a shared coordinate reference, as depicted in **Figure 5**.

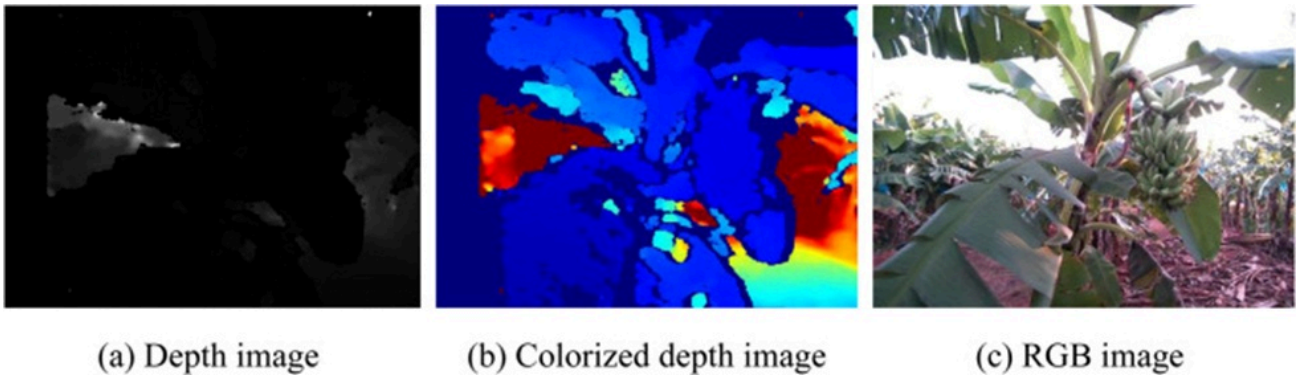

Figure 5. Images from different information sources.

The procedure begins with camera calibration, a step that relies on the camera's intrinsic coefficients (focal lengths fx and fy, and principal point offsets cx and cy) and its extrinsic parameters (rotation matrix R and translation vector T). The intrinsic coefficients encode the camera's internal optical characteristics, while the extrinsic parameters describe the spatial pose and orientation linking the two cameras. After calibration, the 3D point cloud is transferred from the depth sensor's reference frame to the RGB camera's reference frame. This geometric transformation applies the extrinsic terms R and T to remap every 3D point (X, Y, Z) originating in the depth camera frame into the RGB camera frame, as stipulated by **Eq. 2**:

$$\begin{bmatrix} X' \\ Y' \\ Z' \end{bmatrix} = R \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} + T \quad (2)$$

© 2024 The Author(s). | Published by Agro Intelligence Research Press

In this expression, (X′, Y′, Z′) denotes the point after relocation into the RGB camera coordinate frame. The concluding stage projects the transformed 3D point cloud onto the RGB sensor's 2D imaging surface. This projection exploits the RGB camera's intrinsic coefficients to convert the spatial triplets (X′, Y′, Z′) into pixel coordinates (u, v). The governing relationship is given below:

$$x' = f_x \frac{X'}{Z'} + c_x \quad (3)$$

$$y' = f_y \frac{Y'}{Z'} + c_y \quad (4)$$

In the notation above, x′, y′ signify positions on the resulting 2D image plane, whereas X′, Y′, Z′ indicate the 3D coordinates within the RGB camera frame. By executing the foregoing sequence, the depth sensor's 3D point cloud is registered with the RGB camera's 2D image, laying a trustworthy data foundation for the three-dimensional localisation and recognition tasks that follow.

With image registration complete, the two-dimensional bounding box returned by YOLO for each banana bunch must be combined with depth information to determine the cluster's exact 3D location. First, the RGB image is fed into the Slim-Banana architecture introduced earlier, producing a 2D bounding box around the bunch and allowing computation of its pixel-space centroid (xc, yc). Instead of trusting the depth reading at that lone centroid, a set of depth samples is drawn from multiple positions inside the bounding region. These samples are then ordered, and statistical outliers are removed to sidestep potential artefacts from the peduncle or background elements. Ultimately, the centroid coordinates (xc, yc) and the depth Z are fed into a computation that yields the three-dimensional spatial position of the bunch. The explicit formulation appears below:

$$\begin{cases} X_c = \frac{(x_c - c_x) \cdot Z}{f_x} \\ Y_c = \frac{(y_c - c_y) \cdot Z}{f_y} \\ Z_c = Z \end{cases} \quad (5)$$

Via this workflow, the precise three-dimensional location of a banana fruit can be inferred solely from its two-dimensional detection window. By marrying the Slim-Banana detector with TOF depth sensing, the technique advanced in this paper enables accurate 3D localisation of banana fruits amidst the complexities of orchard settings, equipping robotic harvesters with a dependable perceptual foundation.

Results and Discussion

Experimental environment

To ensure an impartial comparison of algorithmic merit, all trials reported here were executed on an identical hardware configuration and under matching hyperparameter settings. The computational setup consisted of: a 13th Gen Intel® Core i9-13900K CPU (3 GHz, 24 physical cores with 32 threads), an NVIDIA GeForce RTX 4090 GPU, Ubuntu 18.04 as the operating system, CUDA 11.1.74, OpenCV 4.8.0, and the PyTorch 2.0.1 deep learning toolkit.

(1) Parameter setting: Input image dimensions were standardised to 640 × 640 pixels, a choice that trims the computational footprint whilst preserving sufficient visual detail and accelerating training throughput. All models

© 2024 The Author(s). | Published by Agro Intelligence Research Press

underwent training and testing across 500 epochs, by which point performance metrics had plateaued. The batch size was fixed at 24, a value that reconciles GPU memory constraints with available computational capacity, permitting a balanced number of samples per optimisation step. The initial learning rate was set to 0.01, with momentum at 0.90, a pairing that promotes faster convergence and tempers fluctuations during stochastic gradient descent, resulting in a steadier, more efficient optimisation trajectory. A weight-decay coefficient of 0.0005 was used as a regularisation penalty to counteract overfitting.

(2) Training strategy: Throughout the training phase, the K-Means clustering procedure was harnessed to pinpoint the most suitable anchor box aspect ratios and to autonomously locate optimal cluster centroids, thereby steering the network's optimisation. Aiming to bolster generalisation capacity and resilience, an assortment of data augmentation operations was deployed, namely: Mosaic blending to enrich sample variety and scene complexity; Mixup interpolation, which fabricates synthetic training instances through convex combinations of image pairs and their labels; EMA (Exponential Moving Average) to temper parameter fluctuations and foster stability; HSV colour space perturbation, modifying saturation and brightness channels to handle diverse illumination and chromatic variation; and horizontal Flip transforms to enhance the model's sensitivity to bilateral symmetry.

Evaluation metrics

To provide a more comprehensive appraisal of model quality, this study uses precision, recall, mAP@0.5, F1 score, inference latency, and model footprint as yardsticks to assess the banana fruit detection network. The expressions defining precision, recall, mAP, and F1 are supplied below:

$$P = \frac{T_p}{T_p + F_p} \quad (6)$$

$$R = \frac{T_p}{T_p + F_N} \quad (7)$$

$$mAP = \frac{\sum_{i=1}^{C} AP_i}{C} \quad (8)$$

$$F1 = \frac{2 \times P \times R}{P + R} \quad (9)$$

Here, Tp counts genuine positive instances correctly flagged as positive; Fp counts genuine negative instances erroneously flagged as positive; and FN counts genuine positive instances incorrectly flagged as negative—equivalently, the tally of overlooked banana bunches (false negatives). These three quantities jointly govern the computed values of P and R. Precision, recall, mAP, and the F1 measure collectively offer critical insight into a detection model's capabilities. AP corresponds to the integral beneath the Precision-Recall curve, summarising the model's overall detection quality. Inference latency captures the time required to process a single banana photograph and serves as an indicator of computational throughput, carrying significant weight in deployment contexts. Model_Size quantifies the storage volume occupied by the model parameters.

Comparison with the different networks

To substantiate the value of our contribution, the Slim-Banana model developed in this work was benchmarked against multiple contemporary architectures using the held-out test partition. The findings are collated in **Table 1**, which juxtaposes the test outcomes of our detector with those of other leading methods on the fruit detection problem.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

Evaluation spans a broad spectrum of criteria, among them Precision, Recall, mAP@0.5, inference speed, and storage size. Inference speed and storage size are explicitly included to gauge the model's suitability for real-time operation and its lightweight nature.

Table 1. Comparison of the detection performance of different networks.

Pattern	Model_Size/MB	Inference time/ms	mAP	Recall	Precision
YOLOv8l	87.7	8.0	0.967	0.941	0.934
YOLOv8x	136.8	9.5	0.958	0.892	0.93
YOLOv8m	52.1	6.7	0.951	0.871	0.975
YOLOv8l-vanillanet	109.6	8.8	0.935	0.88	0.96
Slim-Banana	70.9	6.6	0.978	0.94	0.956

For a thorough head-to-head assessment, Slim-Banana was compared against members of the YOLOv8 series (specifically YOLOv8m, YOLOv8l, and YOLOv8x) as well as YOLOv8l-vanillanet [29], a variant built upon a VanillaNet backbone.

When evaluated on the test set, the Slim-Banana detector posted precision, recall, mAP, per-frame inference latency, and model size figures of 0.956, 0.94, 0.978, 6.6 ms, and 70.9 MB, respectively. Stacked against the baseline YOLOv8l, Slim-Banana delivered a 1.10% uplift in mAP, shaved 1.4 ms off per-frame inference time, and reduced the model footprint by 16.8 MB. Compared with several alternative YOLOv8 instantiations, the detector proposed here improves fruit detection fidelity by 1.1%-4.3%. Notably, alongside these accuracy gains, Slim-Banana achieves a compact form factor, reducing inference duration by 0.1-2.9 ms, thereby simplifying installation on resource-constrained end devices. This performance profile is attributable to Slim-Banana's substitution of GSConv for standard convolution (SC); GSConv reduces the operation count by fusing grouped convolution with spatial convolution, thereby shrinking the overall model's bulk while maintaining high detection quality. Additionally, a cross-stage local network (GSCSP) block, engineered via a single-stage aggregation paradigm, reduces both computational expense and architectural complexity without eroding accuracy.

Figure 6 provides a visual comparison of detection outputs from the refined model and YOLOv8l under occlusion, weak illumination, and similar conditions. The side-by-side examples reveal that the enhanced detector yields markedly better recognition fidelity. In comparison with YOLOv8l, it delineates and pinpoints fruit with greater fidelity, suppressing spurious detections and superfluous bounding boxes. Within occluded settings, YOLOv8l tends to generate multiple overlapping or incorrectly judged fruit boundaries, eroding detection quality. Slim-Banana, through algorithmic refinements and strengthened feature extraction, reliably spots partially hidden clusters, lowering both false alarm and miss rates. Under taxing illumination—be it harsh glare, deep shade, or dim conditions—YOLOv8l may miss fruit entirely, whereas Slim-Banana continues to detect them effectively. Slim-Banana demonstrates greater adaptability and resilience, reliably detecting banana bunches across a wide range of operational conditions.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

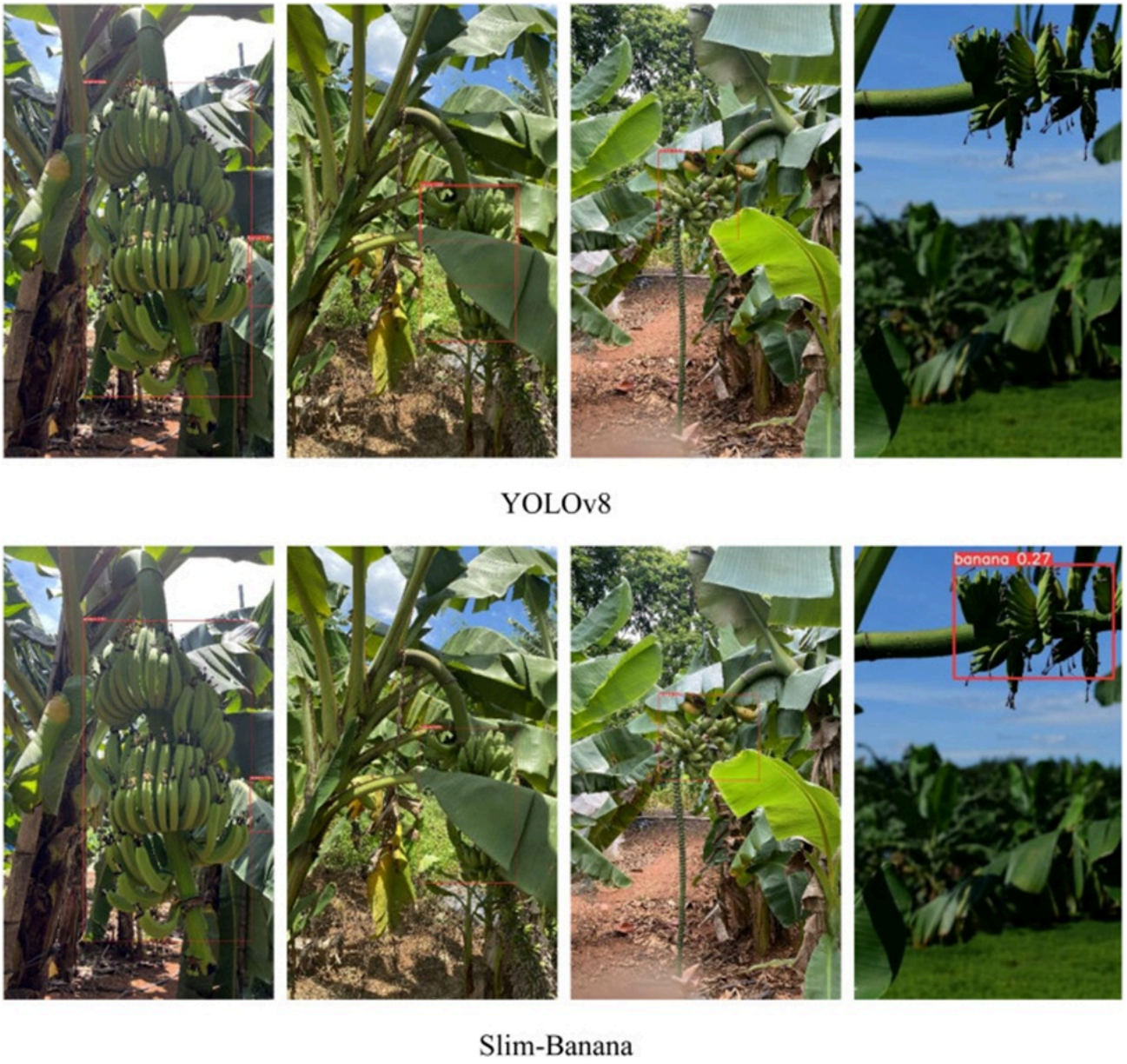

Figure 6. Comparison before and after YOLOv8 network improvement.

Field deployment experiment

Pipeline deployment on Nvidia Orin NX

In genuine orchard environments, constrained hardware resources and spotty network access pose obstacles to real-time fruit recognition and analysis. Offloading the model onto edge computing hardware to realise end-to-end banana fruit detection and localisation thus ensures dependable perception outputs despite resource scarcity, thereby boosting the overall system's dependability. We ported the pipeline onto an Nvidia Orin NX edge module and contrasted its behaviour before versus after transfer to assess its efficacy across a variety of real orchard conditions.

As shown in **Table 2**, Slim-Banana achieved a precision of 0.947, a recall of 0.948, and an mAP of 0.98 on a high-performance desktop workstation. Relative to its pre-deployment figures, mAP increased by 0.2% after deployment. Nevertheless, because the edge device offers considerably less computational horsepower, inference latency stretched to 113.6 ms after deployment. This elongation in processing time reflects the reality of executing inference on a resource-

© 2024 The Author(s). | Published by Agro Intelligence Research Press

constrained edge platform. Although latency reached approximately 0.114 s, it remains sufficiently brisk for practical orchard workflows and does not meaningfully degrade the system's real-time performance or the operator's experience. Concurrently, the merits of edge computing persist, as it enables data processing locally and on-site recognition within the orchard.

Table 2. Comparison of model performance before and after deployment.

Patterns	Display memory	Inference time/ms	mAP	Recall	Precision
Slim-Banana	4449MiB	6.6	0.978	0.94	0.956
Slim-Banana _ Orin		113.6	0.98	0.948	0.947

Performance evaluation in different scenarios

Owing to the natural growth patterns of banana plants, foliage and fruit clusters frequently interleave in commercial orchards, a difficulty compounded by varied illumination phenomena such as direct sunlight, cast shadows, and strong backlighting—all of which impose stricter demands on the perception system. To more rigorously probe the model's behaviour under real conditions, the present work partitions field scenarios into four distinct types—direct lighting, backlighting, occluded, and non-occluded—and scrutinises model outputs within each category. This granular taxonomy enables an in-depth examination of the model's true performance and flexibility across diverse, challenging environmental states, with findings summarised in **Table 3**.

Table 3. Comparison of model performance in different scenarios.

Scenes	F1	mAP	Recall	Precision
Direct Lighting	0.978	0.989	0.965	0.991
Backlighting	0.913	0.967	0.905	0.921
Occlusion	0.927	0.966	0.93	0.925
Non-occlusion	0.977	0.991	0.96	0.995

Reviewing **Table 3** reveals that under the orchard's fluctuating illumination patterns, the model's recognition precision, recall, mAP, and F1 in direct-lighting scenes stand at 0.991, 0.965, 0.989, and 0.978, respectively. When placed alongside backlit settings, these indicators rose by 7%, 6%, 2.2%, and 6.5%, respectively. In scenes free of occlusion, the model's precision, recall, mAP, and F1 reach 0.995, 0.96, 0.991, and 0.977, respectively. Contrasted with occluded situations, these metrics increased by 7%, 3%, 2.5%, and 5%, respectively. A more searching analysis indicates that in genuine field conditions, obstructions and adverse lighting—backlighting being a prime example—can hamper the model's detection capacity. Occlusion may partially shroud the target object, prompting the model to mischaracterise fruit boundaries or to overlook the fruit completely. Challenging light, such as backlighting, can wash out image contrast, blurring the distinction between foreground objects and their surroundings and consequently impairing recognition precision. The Slim-Banana model presented in this paper incorporates a cross-stage partial network module realised via a one-shot aggregation scheme, which facilitates productive information exchange among feature representations at different depths. This architectural choice enables the model to maintain high recognition accuracy even under taxing conditions such as occlusion and backlighting, achieving 0.966 and 0.967, respectively. **Figure 7** shows Slim-Banana's detection performance across different scenario types, confirming that the model reliably identifies banana fruits across a broad spectrum of field conditions.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

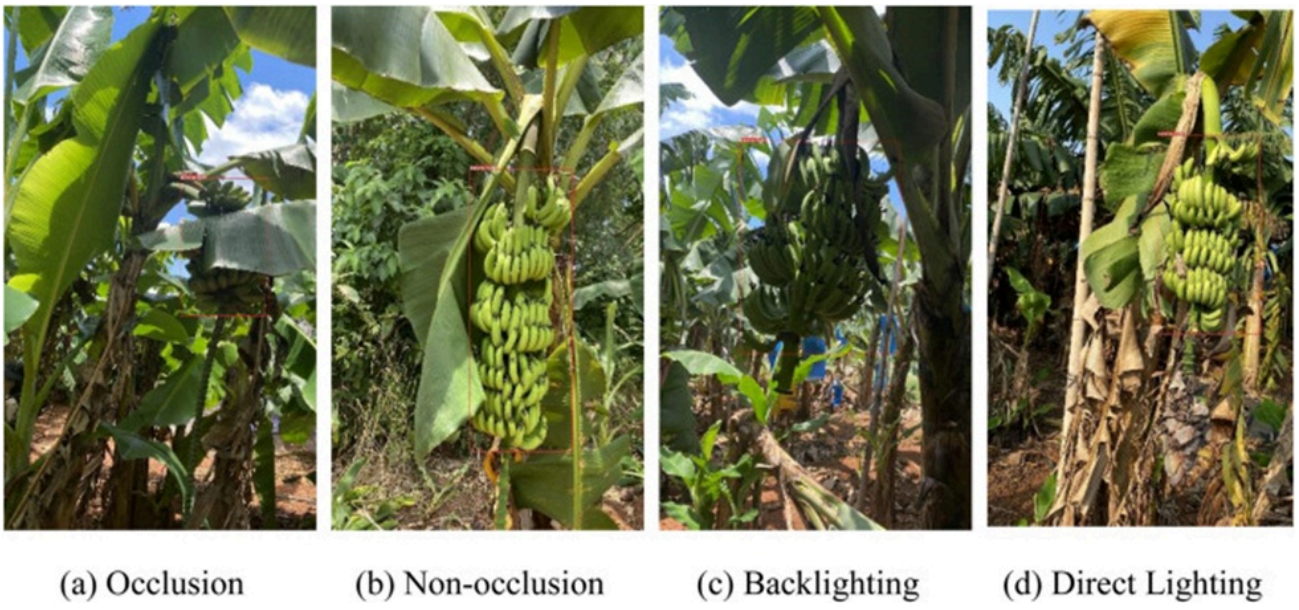

Figure 7. Recognition renderings in different scenarios.

Three-dimensional localization of fruits

This investigation sets forth a methodology for pinpointing fruit coordinates with precision in open-field conditions. The procedure starts with camera calibration to establish correspondence between the 2D detection outputs from Slim-Banana and the depth measurements from the sensing hardware. Slim-Banana provides bounding rectangles that enclose individual banana clusters in the scene, fixing their positions on the image plane. Meanwhile, the depth sensor delivers per-pixel range readings, which help gauge the physical distance from the camera to each banana. The area enclosed by Slim-Banana's predicted bounding rectangle is subsequently marked as the region of interest (ROI), and the depth values falling within that region are extracted from the accompanying depth map. In the final step, the camera's intrinsic calibration terms are combined with the depth readings to transform the banana's image-space coordinates into real-world spatial coordinates, thus recovering its three-dimensional position.

The hardware rig is depicted in **Figure 8** and consists of a Realsense D455 unit, a Jetson Orin NX edge processor, a visual monitor, and a laser distance meter. The Realsense D455 is mounted atop a tripod to ensure consistent, vibration-free capture of both photographic frames and depth data. The laser distance meter measures the fruit's actual 3D coordinates, thereby yielding authoritative reference values for validation. The Jetson Orin NX edge processor ingests the visual and depth feeds from the Realsense D455 and executes the Slim-Banana detection routine onboard. The monitor renders detection and localisation outputs as they are generated. Experimental observations indicate that the depth sensor's usable measurement range is 0.3-1.2 m; fruit within this range can be localised reliably. The comprehensive set of empirical results is tabulated in **Table 4**.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

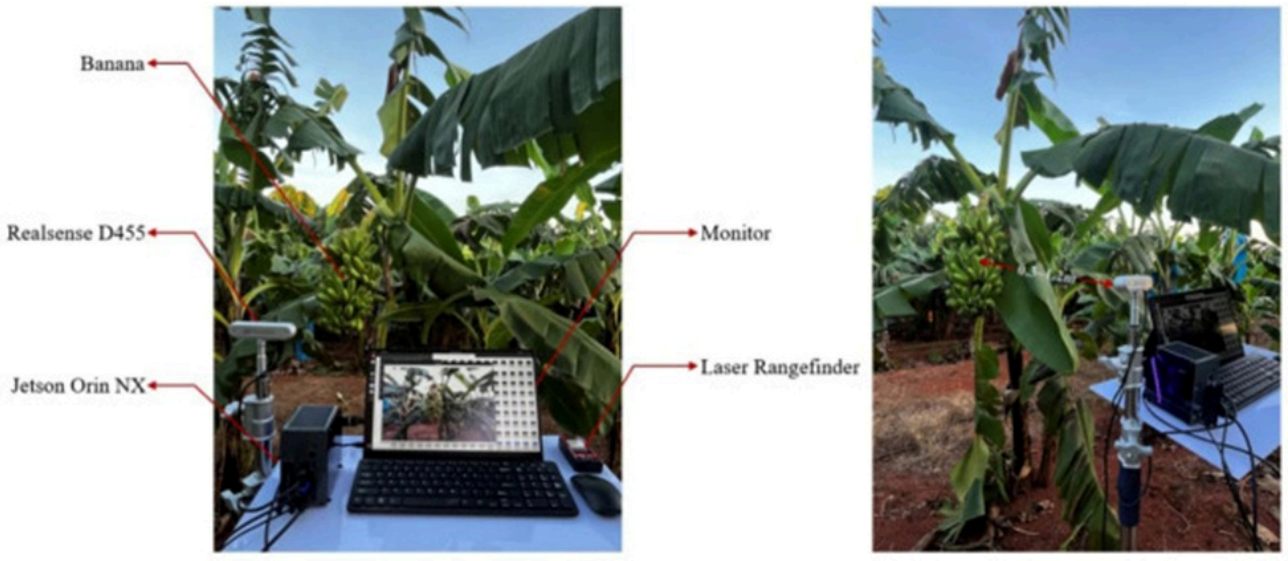

Figure 8. Equipment working diagram. The platform includes a RealSense D455 for depth sensing, an edge computing device Jetson Orin NX, a display screen for real-time monitoring, and a laser rangefinder for distance measurement.

Table 4. Three-dimensional coordinate measurements of bananas in the field.

No	Coordinate error (mm)	Real 3D coordinates (m)	Detection 3D coordinates (m)
1	(3, 6, 17)	(0.116, 0.048, 0.287)	(0.113, 0.042, 0.304)
2	(13, 7, 21)	(0.061, 0.042, 0.326)	(0.074, 0.035, 0.347)
3	(5, 9, 10)	(−0.029, 0.040, 0.432)	(−0.024, 0.049, 0.422)
4	(13, 20, 10)	(−0.096, 0.080, 0.423)	(−0.109, 0.060, 0.433)
5	(2, 12, 13)	(−0,042, 0.071, 0.436)	(−0.040, 0.059, 0.449)
6	(11, 9, 1)	(0.248, 0.122, 0.486)	(0.237, 0.113, 0.487)
7	(21, 7, 6)	(0.216, 0.112, 0.583)	(0.195, 0.119, 0.589)
8	(6, 16, 3)	(−0.109, 0.150, 0.608)	(−0.115, 0.134, 0.605)
9	(15, 4, 16)	(0.083, 0.193, 0.688)	(0.068, 0.189, 0.704)
10	(13, 11, 11)	(0.314, 0.208, 0.784)	(0.301, 0.197, 0.795)
11	(19, 18, 7)	(0.229, 0.071, 0.890)	(0.248, 0.053, 0.897)
12	(26, 15, 22)	(−0.303, 0.201, 0.952)	(−0.277, 0.186, 0.974)
13	(19, 31, 15)	(−0.037, 0.123, 1.075)	(−0.018, 0.092, 1.090)
14	(24, 21, 33)	(−0.122, 0.067, 1.158)	(−0.098, 0.088, 1.191)
15	(12, 7, 23)	(0.138, 0.875, 1.174)	(0.126, 0.868, 1.197)

While running experiments, we noticed that whenever the sensor-to-fruit spacing contracted below 0.3 m, the system lost the ability to determine position reliably; conversely, when that spacing stretched beyond 1.2 m, measurement errors swelled dramatically, compromising localisation quality. Based on these observations, we settled on 0.3-1.2 m as the

© 2024 The Author(s). | Published by Agro Intelligence Research Press

optimal operating band for spatial positioning. Reviewing the entries in **Table 4**, the largest recorded deviation along the X-axis is 26 mm, with the smallest at 2 mm; the Y-axis registers a peak deviation of 31 mm and a trough of 4 mm; and the Z-axis spans a maximum discrepancy of 33 mm down to a minimum of 1 mm. Averaged across measurements, the errors amount to 13.47 mm, 12.87 mm, and 13.87 mm for the three axes, respectively. Although a degree of inaccuracy is unavoidable, these magnitudes are tolerable within the practical working envelope.

Potential contributors to these deviations include the sensor's inherent resolution limits, the irregular geometry and spatial attitude of individual fruits, and variability in ambient lighting. Nevertheless, taken as a whole, the system delivers trustworthy fruit localisation so long as it remains within the prescribed distance window. Additionally, as illustrated in **Figure 9**, the system demonstrates the ability to locate fruit across diverse illumination conditions, further supporting its real-world utility and operational stability.

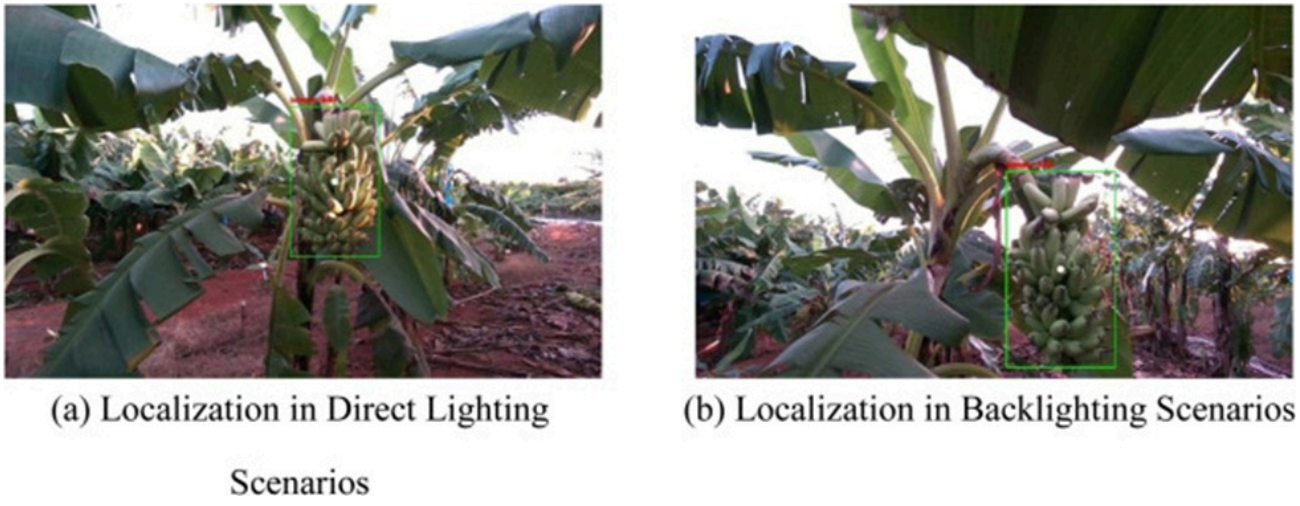

Figure 9. Banana 3D localization.

This paper describes Slim-Banana, a lightweight pipeline for fruit detection and spatial positioning, designed to address the challenges of banana identification and localisation under the shifting light conditions encountered in orchard environments. We reworked the YOLOv8l network architecture by embedding the GSConv convolutional block into the VoV-GSCSP module, yielding a compact model that does not compromise detection fidelity. Additionally, we transferred this model onto an Nvidia Orin NX edge computing platform and scrutinised its on-device performance, including inference throughput and resource demands. By integrating TOF sensing, we achieved high-precision 3D localisation of bananas, and field evaluations confirmed the system's practical viability.

Numerous research teams have lately made substantial headway in fruit detection and spatial localisation. With respect to banana detection specifically, Fu *et al.* [9, 10], Cai *et al.* [11], and Neupane *et al.* [12] deployed deep neural architectures to identify bananas and their pseudostems, underscoring the considerable potential of deep learning for fruit perception. That said, these solutions often come with bulky parameter sets and are cumbersome to deploy—factors that may restrict their real-time suitability in operational settings. Our Slim-Banana network, by contrast, not only upgrades the YOLOv8l topology—lifting detection accuracy across varying light conditions—but also sharply curtails computational overhead and resource usage once fielded. Regarding banana localisation, Zhou *et al.* [17] used structured light to perform 3D positioning of bananas, achieving a mean spatial error of 22.875 mm. Structured light approaches depend on projected illumination; however, fluctuations in ambient light can erode their localisation precision. In comparison, our TOF-driven banana positioning method exhibited sharper accuracy, quicker response, and stronger generalisation in the conducted trials.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

The originality of this work lies in its pairing of lightweight, high-efficiency detection with a demonstration of real-orchard deployability through edge-device implementation. In parallel, the incorporation of TOF depth sensing provided three-dimensional fruit coordinates, yielding a more end-to-end, operationally ready solution for robotic harvesting platforms. Although our system delivered strong results in terms of detection accuracy and responsiveness, the experimental evidence also exposed certain weaknesses. First, while Slim-Banana performed strongly under variable lighting and partial obstruction, detection rates slipped during extreme illumination episodes—intense backlighting being a prime case—particularly when banana fruits were heavily concealed, or when the sensor approached too close to the target, at which juncture localisation failed due to intrinsic sensor limitations. Second, after deployment, while the Nvidia Orin NX hardware proved adequate for sustaining live detection, a slight deceleration in inference speed surfaced when the pipeline was tasked with processing large image batches concurrently. Upcoming research will aim to address these deficiencies and further refine the algorithm to enhance its ruggedness in extreme operating environments.

Conclusion

This paper advances Slim-Banana, a lightweight banana detection architecture derived from an enhanced YOLOv8, which is deployed on edge computing hardware and integrated with RealSense sensors for fruit spatial localisation. The principal takeaways are condensed below:

(1) This paper introduces the Slim-Banana model, which achieves an effective trade-off between high detection precision and real-time execution by incorporating modules such as SlimNeck and GSConv. When set against the pre-modification baseline, the mAP metric improved by 1.10% to 97.8%; single-frame inference latency decreased by 1.4 ms; and the model's on-disk footprint decreased by 16.8 MB. Benchmarked against contemporary mainstream architectures, experimental results confirm that Slim-Banana substantially improves both accuracy and speed in the banana recognition task.

(2) Integrated with a RealSense depth sensor to register RGB imagery with depth maps, this paper proposes a banana fruit localisation framework that recovers three-dimensional fruit coordinates via TOF technology. Trials verified the system's positioning fidelity in outdoor field conditions, yielding mean errors of 13.47 mm, 12.87 mm, and 13.87 mm along the three axes, respectively. Within the defined working range, dependable fruit localisation is attainable.

(3) Slim-Banana was ported to Nvidia Orin NX edge hardware; orchard field tests were devised and conducted to measure the model's on-site performance and resource footprint in a genuine orchard context, and an analysis of the model's behaviour under varying illumination and occlusion regimes was undertaken. Experimental results reveal that mAP reaches 98.0%, inference latency is approximately 0.114 s, and GPU memory consumption is 4449 MiB. These outcomes verify that the model remains capable of delivering high-accuracy, computationally efficient inference on edge computing devices in complex environments, further corroborating the practicality of the methodology proposed in this paper.

Acknowledgements

This work was supported by the Key Research and Development Program of Hainan (Grant number ZDYF2024XDNY153), the National Natural Science Foundation of China (Grant number 32401709), the Open Competition Program of Top Ten Critical Priorities of Agricultural Science and Technology Innovation for the 14th Five-Year Plan of Guangdong Province (Grant number 2022SDZG03), and the Central Public-interest Scientific Institution Basal Research Fund for the Chinese Academy of Tropical Agricultural Sciences (Grant number 1630132024001). The authors wish to thank the anonymous reviewers for their useful comments on this paper.

Conflict of interest

© 2024 The Author(s). | Published by Agro Intelligence Research Press

None

Financial support

None

Ethics statement

The research presented in this study was conducted in accordance with ethical standards and guidelines for research involving technology and agricultural practices. No human or animal subjects were involved in this study. The research methodologies used, including data collection and analysis, were carried out with full respect for the environment and agricultural practices in the study area.

Received: 15 Apr 2023 Revised: 20 Jun 2023 Accepted: 06 Jul 2023
Published online: 25 November 2024

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Panigrahi N, Thompson AJ, Zubelzu S, Knox JW. Identifying opportunities to improve management of water stress in banana production. Sci Hortic. 2021;276:109735.

2. Ploetz RC, Kema GHJ, Ma LJ. Impact of diseases on export and smallholder production of banana. Annu Rev Phytopathol. 2015;53(1):269-88.

3. Chen T, Zhang R, Zhu L, Zhang S, Li X. A method of fast segmentation for banana stalk exploited lightweight multi-feature fusion deep neural network. Machines. 2021;9(3):66.

4. Fu L, Wu F, Zou X, Jiang Y, Lin J, Yang Z, et al. Fast detection of banana bunches and stalks in the natural environment based on deep learning. Comput Electron Agric. 2022;194:106800. https://doi.org/10.1016/j.compag.2022.106800.

5. Kootstra G, Wang X, Blok PM, Hemming J, van Henten E. Selective harvesting robotics: current research, trends, and future directions. Curr Robot Rep. 2021;2(1):95-104.

6. Fu L, Duan J, Zou X, Lin G, Song S, Ji B, et al. Banana detection based on color and texture features in the natural environment. Comput Electron Agric. 2019;167:105057.

7. He Z, Xiong J, Chen S, Li Z, Chen S, Zhong Z, et al. A method of green citrus detection based on a deep bounding box regression forest. Biosyst Eng. 2020;193:206-15. https://doi.org/10.1016/j.biosystemseng.2020.03.001.

8. Chen J, Wu J, Wang Z, Qiang H, Cai G, Tan C, et al. Detecting ripe fruits under natural occlusion and illumination conditions. Comput Electron Agric. 2021;190:106450. https://doi.org/10.1016/j.compag.2021.106450.

9. Fu L, Duan J, Zou X, Lin J, Zhao L, Li J. Fast and accurate detection of banana fruits in complex background orchards. IEEE Access. 2020;8:196835-46. https://doi.org/10.1109/ACCESS.2020.3029215.

10. Fu L, Yang Z, Wu F, Zou X, Lin J, Cao Y, et al. YOLO-banana: a lightweight neural network for rapid detection of banana bunches and stalks in the natural environment. Agronomy. 2022;12:391.

11. Cai L, Liang J, Xu X, Duan J, Yang Z. Banana pseudostem visual detection method based on improved YOLOV7 detection algorithm. Agronomy. 2023;13:999. https://doi.org/10.3390/agronomy13040999.

12. Neupane B, Horanont T, Hung ND, Vadrevu KP. Deep learning based banana plant detection and counting using high-resolution red-green-blue (RGB) images collected from unmanned aerial vehicle (UAV). PLoS One. 2019;14(10):e0223906.

13. Wang P, Niu T, He D. Tomato young fruits detection method under near color background based on improved Faster R-CNN with attention mechanism. Agriculture. 2021;11:1059. https://doi.org/10.3390/agriculture11111059.

14. Zheng Z, Xiong J, Wang X, Li Z, Huang Q, Chen H, et al. An efficient online citrus counting system for large-scale unstructured orchards based on the unmanned aerial vehicle. J Field Robot. 2022;40(3):552-73. https://doi.org/10.1002/rob.22147.

15. Xiong J, He Z, Lin R, Liu Z, Bu R, Yang Z, et al. Visual positioning technology of picking robots for dynamic litchi clusters with disturbance. Comput Electron Agric. 2018;151:226-37.

16. Wang C, Luo T, Zhao L, Tang Y, Zou X. Window zooming–based localization algorithm of fruit and vegetable for harvesting robot. IEEE Access. 2019;7:103639-49.

17. Zhou L, Yang Z, Deng F, Zhang J, Xiao Q, Fu L, et al. Banana bunch weight estimation and stalk central point localization in banana orchards based on RGB-D images. Agronomy. 2024;14(6):1123.

18. Milella A, Marani R, Petitti A, Reina G. In-field high throughput grapevine phenotyping with a consumer-grade depth camera. Comput Electron Agric. 2019;156:293-306.

19. Yu L, Xiong J, Fang X, Yang Z, Chen Y, Lin X, et al. A litchi fruit recognition method in a natural environment using RGB-D images. Biosyst Eng. 2021;204:50-63.

20. Lin G, Tang Y, Zou X, Li J, Xiong J. In-field citrus detection and localisation based on RGB-D image analysis. Biosyst Eng. 2019;186:34-44.

21. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

22. Wang CY, Bochkovskiy A, Liao HYM. YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

23. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2014:580-7.

24. Girshick R. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision (ICCV). IEEE; 2015. p. 1440-8.

© 2024 The Author(s). | Published by Agro Intelligence Research Press

25. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell. 2017;39(6):1137-49.

26. He K, Gkioxari G, Dollár P, Girshick R. Mask R-CNN. arXiv e-prints. 2017;arXiv:1703.06870. https://doi.org/10.48550/arXiv.1703.06870.

27. Jocher G, Chaurasia A, Qiu J. Ultralytics YOLO. 2023. https://github.com/ultralytics/ultralytics.

28. Li H, Li J, Wei H, Liu Z, Zhan Z, Ren Q, et al. Slim-neck by GSConv: a better design paradigm of detector architectures for autonomous vehicles. 2022.

29. Chen H, Wang Y, Guo J, Tao D. VanillaNet: the power of minimalism in deep learning. Advances in Neural Information Processing Systems (NeurIPS). 2024;36.

© 2024 The Author(s). | Published by Agro Intelligence Research Press