# A Novel YOLO-Like Multi-Branch Architecture for Accurate Apple Detection and Segmentation Under Orchard Constraints

Olzhas Olzhayev[1], Nurbibi Imanbayeva[2], Satmyrza Mamikov[3], Bibigul Baibek[4]
Joldasbekov Institute of Mechaniks and Engeeniring[1]
International Information Technology University, Almaty, Kazakhstan[2]
University of Friendship of People's Academician A. Kuatbekov, Shymkent, Kazakhstan[3, 4]
Satbayev University, Almaty, Kazakhstan[2]

***Abstract*—This study introduces a novel YOLO-like multi-branch deep learning architecture designed for accurate apple detection and segmentation in orchard environments, addressing the persistent challenges of occlusion, illumination variability, and fruit clustering. The proposed model integrates an enhanced backbone with C2f modules and a Spatial Pyramid Pooling Fast (SPPF) block to capture multi-scale receptive fields, while a Feature Pyramid Network (FPN) combined with a Path Aggregation Network (PAN) ensures effective top-down and bottom-up feature fusion. To extend beyond bounding box localization, a prototype-based segmentation head is incorporated, enabling precise instance mask generation with reduced computational overhead. The model was comprehensively evaluated on the MinneApple dataset, consisting of high-resolution orchard images with polygonal annotations, and compared against state-of-the-art detection and segmentation frameworks, including Faster R-CNN, Mask R-CNN, SSD, YOLO variants, YOLACT, and SOLOv2. Quantitative results demonstrated that the proposed approach achieved superior mean Average Precision (mAP@0.5 = 0.76), precision (0.83), and F1-score (0.76), while maintaining a competitive inference speed of 40 FPS, confirming its suitability for real-time agricultural applications. Qualitative analysis further highlighted robustness in complex orchard conditions, reinforcing the model's applicability for automated harvesting, yield estimation, and orchard monitoring. These findings advance the state of agricultural computer vision by unifying detection and segmentation in a lightweight, high-performance framework.**

***Keywords*—Precision agriculture; detection; segmentation; YOLO-like architecture; multi-branch network; feature pyramid network; real-time inference; orchard monitoring***

## I. Introduction

The increasing demand for precision agriculture has stimulated research into advanced computer vision systems capable of detecting, localizing, and segmenting fruits in real-world environments. Apple production, in particular, benefits substantially from automated monitoring technologies due to its economic importance and the labor-intensive nature of traditional harvesting and yield estimation practices [1]. Conventional image processing methods often struggle with orchard-specific challenges, such as variable illumination, occlusion by leaves or branches, and high fruit density [2]. As a result, deep learning-based object detection and segmentation models have emerged as state-of-the-art solutions, providing robust performance in unstructured agricultural conditions [3].

Object detection frameworks such as Faster R-CNN and Mask R-CNN have demonstrated notable accuracy in fruit recognition tasks [4]. However, these models frequently suffer from limitations in inference speed, which restricts their applicability in real-time field deployment [5]. Single-stage detectors, most prominently the YOLO family of models, have been widely adopted for agricultural applications due to their balance between accuracy and efficiency [6]. Despite these advances, single-branch architectures typically fail to optimize both object localization and segmentation simultaneously, particularly under conditions of high occlusion and overlapping fruit clusters [7]. Thus, developing multi-branch models capable of integrating detection and segmentation pathways represents a promising avenue for improving overall performance in orchard scenarios [8].

Recent studies highlight the importance of multi-scale feature extraction and aggregation in handling variations in fruit size and shape [9]. Feature pyramid networks (FPN) and path aggregation networks (PAN) have become essential components in detection frameworks, enabling richer contextual representation across hierarchical layers [10]. Furthermore, prototype-based mask generation has proven effective in instance segmentation, allowing the prediction of high-resolution fruit masks with reduced computational cost [11]. Integrating these mechanisms into a unified architecture tailored for orchard constraints can substantially enhance robustness against real-world environmental challenges.

This study proposes a novel YOLO-like multi-branch architecture designed for accurate apple detection and segmentation in orchard environments. The contributions are threefold: first, a backbone network with CSP-inspired C2f modules and spatial pyramid pooling is employed to capture multi-scale receptive fields; second, an FPN–PAN neck structure is leveraged to aggregate feature maps for detection and segmentation heads; and third, a prototype-based mask branch is integrated with box regression to generate high-quality instance masks. Comprehensive evaluations on orchard datasets demonstrate that the proposed model outperforms baseline YOLO and Mask R-CNN variants under varying lighting, occlusion, and density conditions [12]. This work establishes a

robust framework for real-time orchard monitoring and contributes to the broader goal of advancing automation in precision agriculture.

## II. Related Works

Research on fruit detection and segmentation has expanded significantly in recent years, with advances in deep learning architectures enabling more accurate and efficient models for real-world agricultural scenarios. Despite these achievements, orchard environments remain highly challenging due to factors such as overlapping fruits, complex backgrounds, and illumination variability. To contextualize the proposed model, this section reviews existing works across four major domains: 1) traditional approaches for fruit detection, 2) deep learning–based detection frameworks, 3) segmentation strategies in agricultural vision tasks, and 4) multi-branch architectures and prototype-based methods.

### A. Traditional Approaches for Fruit Detection

Early fruit detection research primarily relied on handcrafted features and classical computer vision techniques [13]. Methods using color thresholding, texture descriptors, and shape-based heuristics showed promising results under controlled conditions but failed to generalize well in orchard environments characterized by changing illumination and partial occlusion [14]. Approaches such as Hough transforms for circular fruit detection or color-space conversions for background suppression were computationally inexpensive but exhibited poor robustness in dense canopies [15]. Furthermore, the lack of adaptability to intra-class variations, such as differences in apple ripeness stages, limited their utility in large-scale agricultural applications [16]. Although these methods laid foundational insights into fruit localization, their inherent reliance on rigid feature representations hindered scalability and motivated the shift toward data-driven models [17].

### B. Deep Learning-Based Detection Frameworks

With the advent of convolutional neural networks (CNNs), object detection frameworks became the standard for agricultural vision tasks [18]. Two-stage detectors, notably Faster R-CNN, achieved high accuracy in fruit detection, but suffered from slow inference speeds unsuitable for real-time harvesting systems [19]. Single-stage detectors, including SSD and the YOLO series, provided a balance between detection precision and computational efficiency [20]. Studies applying YOLOv3 and YOLOv4 to apple orchards reported significant improvements in handling occlusions and varying fruit scales [21]. Despite these advances, the primary limitation of single-branch detection frameworks lies in their inability to jointly optimize for both object localization and fine-grained segmentation [22]. Consequently, research has shifted toward architectures that combine detection accuracy with the pixel-level understanding necessary for downstream agricultural tasks such as yield estimation and robotic picking [23].

### C. Segmentation Strategies in Agricultural Vision Tasks

Instance segmentation has emerged as a crucial task for precision agriculture, offering detailed information on fruit boundaries beyond bounding boxes [24]. Classical models like Mask R-CNN have been applied to apple detection, producing high-quality masks but with high computational cost [25]. Lightweight segmentation models such as SOLO and YOLACT introduced prototype-based mask generation strategies, enabling faster predictions with reduced complexity [26]. In orchard environments, accurate segmentation is particularly important for separating overlapping fruits and ensuring reliable yield analysis [27]. Moreover, segmentation facilitates better integration with robotic manipulation systems by providing precise contours for grasping and picking [28]. Recent advancements have also explored transformer-based segmentation models, though their high resource demands limit deployment in field robotics [29]. These developments highlight the need for a balanced approach that combines efficiency with segmentation quality, motivating the integration of multi-branch segmentation heads in detection frameworks [30].

### D. Multi-Branch Architectures and Prototype-Based Methods

The introduction of multi-branch networks represents a significant step forward in bridging the gap between detection and segmentation [31]. Architectures that simultaneously optimize detection and mask prediction have demonstrated superior performance in agricultural scenarios, where both accuracy and real-time capability are required [32]. Prototype-based mask branches, in particular, allow efficient generation of instance masks by combining global prototypes with instance-specific coefficients [33]. This approach reduces computational burden while maintaining segmentation quality, making it highly suitable for resource-constrained agricultural systems [34]. The combination of multi-scale feature aggregation with multi-branch segmentation further improves robustness against occlusion and scale variation. These advancements directly inform the design of the proposed YOLO-like multi-branch model, which leverages prototype-based masks and feature aggregation strategies to achieve accurate detection and segmentation under orchard constraints.

## III. Materials and Methods

The design and evaluation of the proposed YOLO-like multi-branch model for apple detection and segmentation were carried out through a systematic methodology that integrates data preprocessing, architectural development, training optimization, and performance assessment (see Fig. 1). This section outlines the technical details underlying the model construction, beginning with the preprocessing strategies applied to orchard images, followed by a comprehensive description of the backbone, neck, and multi-branch heads. The mathematical formulations of the detection and segmentation processes are presented alongside the adopted loss functions, while inference and post-processing steps are described to illustrate the complete operational workflow of the system.

### A. Data Preprocessing

The input image $I \in R^{H \times W \times 3}$ is first resized to a fixed resolution (e.g., 640×640) while preserving aspect ratio by zero-padding. Pixel intensities are normalized to [0, 1], and data augmentation techniques such as mosaic, random horizontal flipping, scaling, and hue–saturation–value (HSV) jittering are applied to increase generalization. The preprocessing pipeline can be expressed as:

$$I' = A\big(N\big(R(I)\big)\big) \tag{1}$$

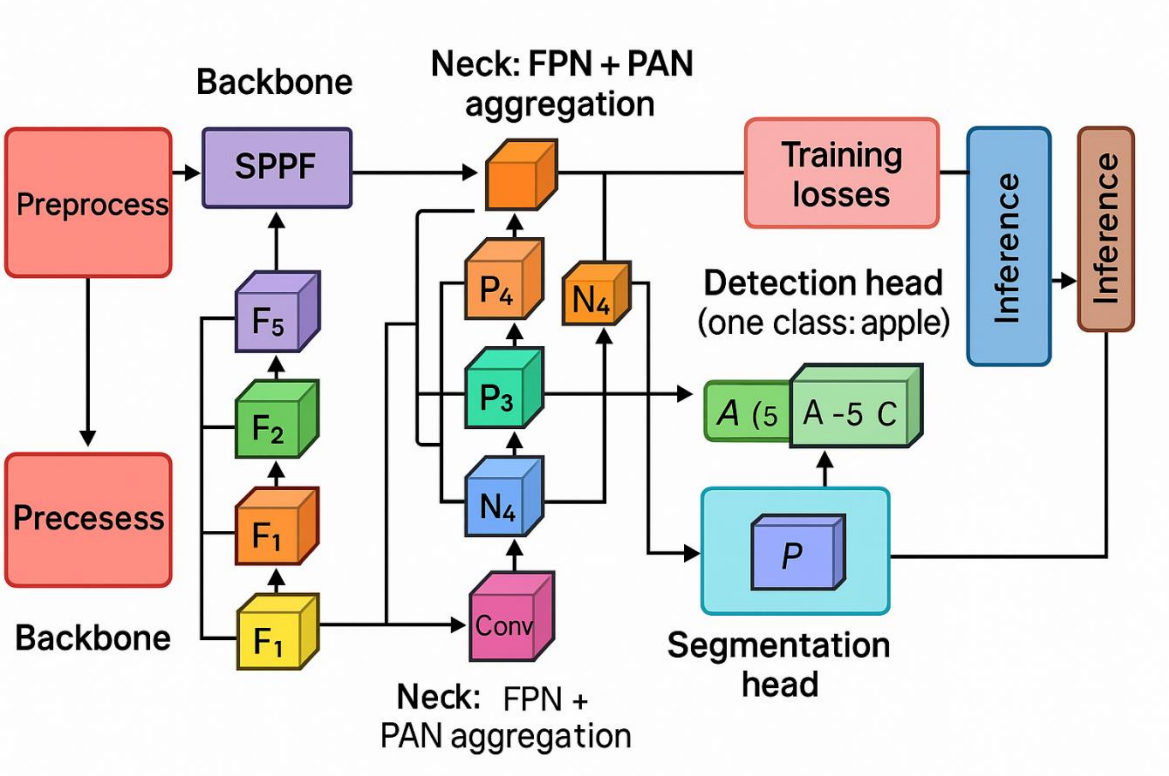



Fig. 1. Architecture of the proposed YOLO-like multi-branch model for apple detection and segmentation under orchard constraints.

where, $R$ denotes resizing with padding, $N$ denotes normalization, and $A$ represents augmentation functions.

### B. Backbone Network

The backbone consists of an initial convolutional layer followed by a series of C2f modules, which are lightweight CSP-inspired residual blocks that enhance gradient flow and feature reuse. The convolutional output at stage $l$ is defined as:

$$F_l = \sigma\left(W_l * F_{l-1} + b_l\right) \tag{2}$$

where, $F_{l-1}$ is the input feature map, $W_l$ and $b_l$ denote the convolutional weights and biases, and $\sigma$ is the non-linear activation (SiLU).

Feature maps are progressively downsampled across five stages, producing $\{F_1, F_2, F_3, F_4, F_5\}$. A Spatial Pyramid Pooling Fast (SPPF) block is applied on the deepest feature map $F_5$ to enlarge the receptive field:

$$F_5^{'} = Concat\begin{pmatrix} MaxPool_{k_1}(F_5), \\ MaxPool_{k_2}(F_5), \\ MaxPool_{k_3}(F_5), \\ F_5 \end{pmatrix} \tag{3}$$

### C. Neck: FPN and PAN Aggregation

To exploit multi-scale information, a Feature Pyramid Network (FPN) with top-down upsampling and a Path Aggregation Network (PAN) with bottom-up downsampling are integrated. This structure ensures both semantic enrichment of shallow layers and spatial refinement of deep layers.

For top-down fusion:

$$P_l = \phi\left(Concat\left(Up\left(F_{l+1}^{'}\right), F_l\right)\right) \tag{4}$$

where, $\phi$ denotes convolution + C2f transformation and $Up(\cdot)$ is bilinear upsampling.

For bottom-up aggregation:

$$N_l = \phi\left(Concat\left(Down\left(P_l\right), P_{l+1}\right)\right) \tag{5}$$

where, $Down(\cdot)$ is a strided convolution.

The outputs of this neck are $\{P_3, N_4, N_5\}$, which are forwarded to the detection and segmentation heads.

### D. Detection Head

For each scale $S \in \{P_3, N_4, N_5\}$, the detection head predicts bounding box offsets, objectness, and class probabilities. The prediction tensor is:

$$Y_S \in R^{A \times H_S \times W_S \times (5+C)} \tag{6}$$

where, $A$ is the number of anchor points, $C$ is the number of classes (here $C$=1, apple), and the 5 corresponds to $(t_x, t_y, t_w, t_h, t_{obj})$ Bounding box regression is decoded as:

$$\hat{b} = \left((t_x + c_x)\cdot s, (t_y + c_y)\cdot s, e^{t_w}\cdot s, e^{t_h}\cdot s\right) \quad (7)$$

where, $c_x$, $c_y$ are grid offsets and s is stride.

### E. Segmentation Head

The segmentation branch employs prototype masks $P \in R^{K\times H_m \times W_m}$ generated from fused multi-scale features. For each detection, the network predicts mask coefficients $\alpha \in R^K$. The instance mask is reconstructed as:

$$\hat{m} = \sigma\left(\sum_{k=1}^{K} \alpha_k P_k\right) \quad (8)$$

followed by cropping the predicted bounding box region.

### F. Loss Functions

The training objective integrates multiple loss components:

Box regression loss (CIoU):

$$L_{box} = 1 - CIoU(\hat{b}, b*) \quad (9)$$

where, $b*$ is the ground-truth box.

Objectness loss (focal BCE):

$$L_{obj} = -\beta\left(1-\sigma(t_{obj})\right)^{\gamma} \log\sigma(t_{obj}) \quad (10)$$

Classification loss:

$$L_{cls} = BCE(\hat{c}, c*) \quad (11)$$

Mask loss (BCE + Dice):

$$L_{mask} = \lambda_1 \cdot BCE(\hat{m}, m*) + \lambda_2 \cdot \left(1 - Dice(\hat{m}, m*)\right) \quad (12)$$

The total loss is:

$$L = w_1 L_{box} + w_2 L_{obj} + w_3 L_{cls} + w_4 L_{mask} \quad (13)$$

### G. Inference and Post-Processing

During inference, predictions are filtered using a confidence threshold $\tau$ and Non-Maximum Suppression (NMS) with IoU threshold $\gamma$. The final detections are defined as:

$$D = \left\{(b_i, p_i, m_i) \mid p_i > \tau, i \notin NMS(\gamma)\right\} \quad (14)$$

Masks are optionally refined with morphological operations, and results are rescaled to the original image resolution.

## IV. Dataset

The dataset employed in this study is the publicly available MinneApple benchmark, specifically curated for apple detection, segmentation, and counting tasks in orchard environments [35-36]. The dataset comprises 1,000 high-resolution RGB images, acquired using a standard smartphone camera mounted horizontally along orchard rows, capturing diverse lighting and canopy conditions. Each image is meticulously annotated with polygonal instance-level masks, delivering precise delineations for over 41,000 fruit instances, facilitating accurate localization and segmentation. Representative image-mask pairs are depicted in Fig. 2, demonstrating the diversity in fruit appearance, clustering patterns, background complexity, and environmental variability addressed by the dataset.

Fig. 2. Sample images from the MinneApple dataset with corresponding polygonal instance-level annotations illustrating variability in lighting, occlusion, and fruit density.

TABLE I. OVERVIEW OF THE MINNEAPPLE DATASET

| Property | Value |
|---|---|
| Total images | 1,000 |
| Total annotated apple instances | > 41,000 |
| Annotation type | Polygonal instance masks |
| Image acquisition device | Smartphone (RGB) |
| Environmental variations | Lighting, occlusion, density, seasonality |
| Example images shown in | Figure 2 |

In terms of dataset composition, Table I provides a structured summary of its key characteristics. The images span the full spectrum of seasonal and environmental conditions, with annotations yielding a rich distribution of fruit counts per image and a broad range of mask complexities and sizes. This diversity introduces significant variability in the target instances, challenging detection and segmentation algorithms to generalize across dense clusters, varying scales, and inconsistent illumination. Our experiments adopt a standard training–validation–test split (e.g., 70%–20%–10%) to ensure rigorous evaluation and reproducibility on this benchmark.

## V. EVALUATION PARAMETERS

To rigorously assess the performance of the proposed YOLO-like multi-branch model on the MinneApple dataset, a set of widely adopted evaluation parameters was employed. These metrics provide complementary insights into detection accuracy, segmentation quality, and the balance between sensitivity and specificity of the model. The following subsections outline the evaluation parameters, accompanied by their mathematical formulations.

Precision measures the proportion of correctly predicted positive instances among all predicted positives [37]. It reflects the model's ability to minimize false positives. Precision is defined as:

$$precision = \frac{TP}{TP + FP} \tag{15}$$

where, $TP$ denotes true positives and $FP$ denotes false positives. A high precision indicates reliable predictions when the model identifies an apple instance.

Recall quantifies the proportion of actual positive instances that are correctly identified by the model [38]. It emphasizes the ability to minimize false negatives and is defined as:

$$recall = \frac{TP}{TP + FN} \tag{16}$$

where, $FN$ represents false negatives. High recall is critical in agricultural monitoring to ensure that most fruit instances are detected.

The F1-score is the harmonic mean of precision and recall, providing a balanced measure when both false positives and false negatives must be considered [39]. It is expressed as:

$$F1 - score = 2\frac{precision \cdot recall}{precision + recall} \tag{17}$$

This parameter is particularly relevant for orchard scenarios where both accurate localization and comprehensive fruit detection are required.

Intersection over Union (IoU) measures the overlap between the predicted bounding box (or mask) and the ground truth [40]. It is a fundamental metric for both detection and segmentation tasks and is defined as:

$$IoU = \frac{\left|B_p \cap B_{gt}\right|}{\left|B_p \cup B_{gt}\right|} \tag{18}$$

where, $Bp$ is the predicted bounding box or mask and $Bgt$ is the ground truth. Predictions are considered correct if their IoU exceeds a predefined threshold (e.g., 0.5).

The mean Average Precision aggregates detection performance across multiple IoU thresholds and recall levels [41]. Average Precision (AP) is computed as the area under the precision–recall curve for a given IoU threshold. The general formula is:

$$AP = \int_0^1 p(r)dr \tag{19}$$

where, $p(r)$ denotes precision as a function of recall. The mAP is then obtained by averaging AP values over all classes (in this study, a single class: apple) and, in some cases, multiple IoU thresholds (e.g., mAP@0.5, mAP@[0.5:0.95]).

For segmentation evaluation, the Dice coefficient provides a robust measure of similarity between the predicted mask and the ground truth mask [42]. It is defined as:

$$Dice = \frac{2 \cdot \left|M_p \cap M_{gt}\right|}{\left|M_p\right| + \left|M_{gt}\right|} \tag{20}$$

where, $M_p$ and $M_{gt}$ represent the predicted and ground truth masks, respectively. The Dice coefficient is particularly effective in handling imbalanced datasets where object pixels are fewer compared to background pixels.

To evaluate computational efficiency, inference speed was measured in frames per second (FPS) [43]. FPS is calculated as:

$$FPS = \frac{N}{T} \tag{21}$$

where, $N$ is the number of processed images and $T$ is the total processing time. This metric is crucial for determining the model's suitability for real-time agricultural applications.

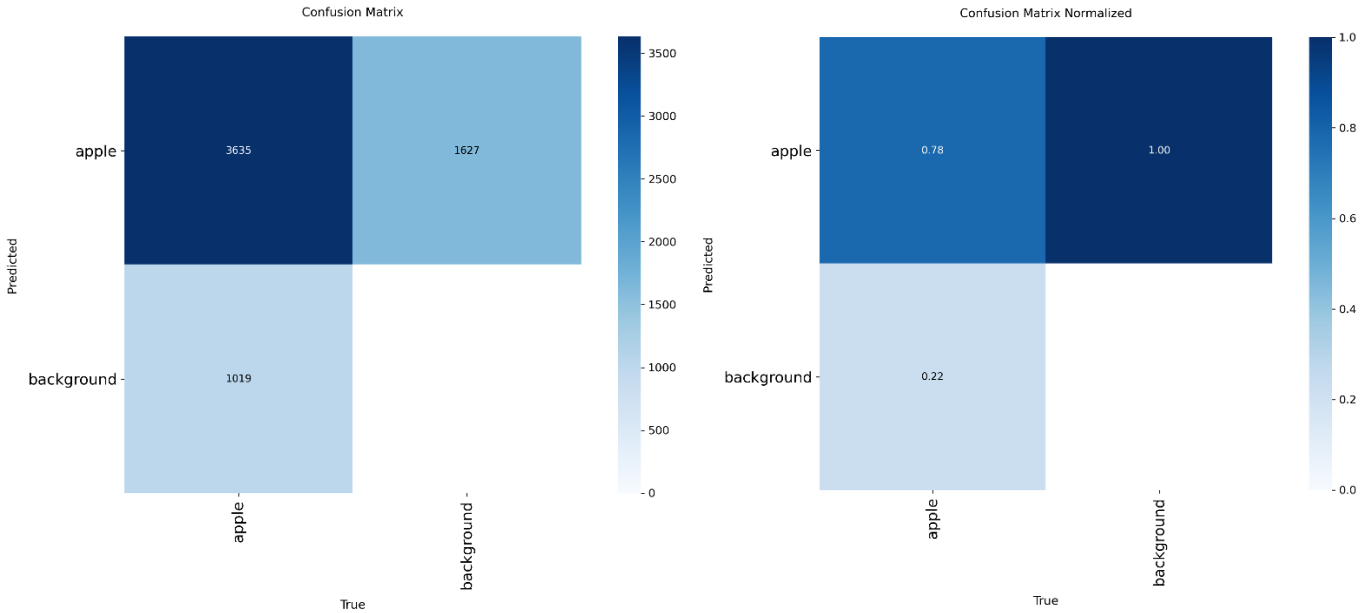
## VI. RESULTS

This section presents the experimental findings obtained from evaluating the proposed YOLO-like multi-branch architecture on the MinneApple dataset. The results are organized to highlight the detection and segmentation performance of the model under diverse orchard conditions, with comparisons against state-of-the-art baselines. Both quantitative metrics, including precision, recall, F1-score, mean Average Precision (mAP), Dice coefficient, and inference speed, as well as qualitative visualizations, are reported to provide a comprehensive assessment. Figures and tables are employed to illustrate key outcomes, enabling a clear interpretation of the model's strengths and limitations in addressing the challenges of fruit detection and segmentation in real-world environments.

Fig. 3 presents the confusion matrix and its normalized counterpart for the proposed YOLO-like multi-branch model evaluated on the MinneApple dataset. The raw confusion matrix [Fig. 3(a)] demonstrates that the model correctly classified a substantial number of apple instances ($n$=3635) while misclassifying a smaller subset as background ($n$=1627), with relatively fewer false negatives ($n$=1019). The normalized confusion matrix [Fig. 3(b)] highlights these results proportionally, indicating that 78% of apple instances were correctly identified, while 22% were incorrectly labeled as background. Furthermore, background regions were classified with perfect accuracy (100%), confirming the model's robustness in distinguishing non-fruit areas. These findings illustrate that while the model demonstrates high specificity, improvements in sensitivity remain necessary to reduce the rate of apple instances missed under challenging orchard conditions such as occlusion and illumination variability.

Fig. 4 illustrates the performance evaluation curves of the proposed YOLO-like multi-branch architecture on the MinneApple dataset. The F1-Confidence curve shows that the maximum F1-score achieved by the model is 0.72 at a confidence threshold of 0.33, demonstrating a strong balance between precision and recall at moderate confidence levels. The Precision-Confidence curve indicates that the model maintains a steadily increasing precision with higher confidence thresholds, ultimately reaching a perfect precision of 1.00 at 0.962. These results suggest that the model is highly reliable when assigning high-confidence predictions, although stricter thresholds reduce recall.

The Precision-Recall curve confirms the model's robustness in orchard detection tasks, achieving a mean Average Precision (mAP@0.5) of 0.756. This demonstrates that the architecture can successfully localize and segment apples with high precision across a wide range of recall values. The Recall-Confidence curve highlights a recall of 0.89 at zero confidence, which gradually decreases as the confidence threshold increases, reflecting the trade-off between sensitivity and reliability in the detection process. Collectively, these curves validate that the proposed model achieves a strong balance between accuracy and generalization, positioning it as a competitive framework for real-time apple detection and segmentation under complex orchard conditions.




a) Confusion matrix results. b) Normalized confusion matrix results

Fig. 3. Confusion matrix and normalized confusion matrix results of the proposed YOLO-like multi-branch model on the MinneApple dataset.

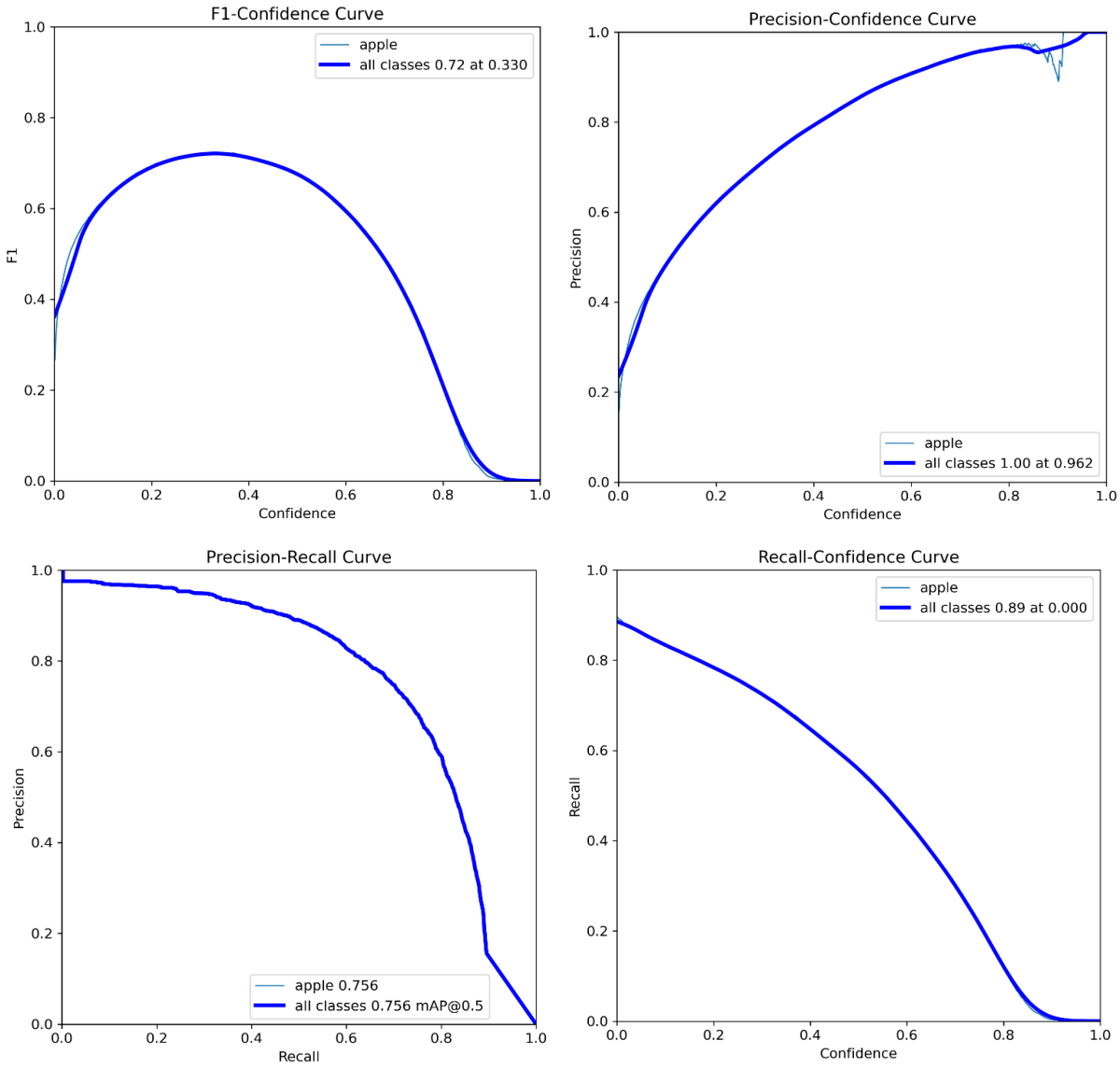



Fig. 4. Performance evaluation curves of the proposed YOLO-like multi-branch model on the MinneApple dataset, including F1-confidence, precision-confidence, precision-recall, and recall-confidence analysis.

Fig. 5 provides a detailed representation of the MinneApple dataset, showcasing an original orchard image alongside its annotated ground truth mask. The raw orchard image on the left illustrates the natural complexities of apple cultivation environments, where fruits are embedded within dense foliage, subject to uneven illumination, and often obscured by overlapping leaves and branches. Such conditions create significant challenges for automated detection systems, as apples frequently blend into the background due to similar colors and textures. On the right, the annotated ground truth mask demonstrates the meticulous manual labeling performed for each fruit instance, using polygonal contours to capture precise shapes and boundaries. The red outlines and filled regions offer pixel-level clarity, serving as a critical benchmark for evaluating detection and segmentation models. Together, the image-mask pair highlights the intricate details of fruit localization and emphasizes the demanding nature of the dataset, which mirrors real-world orchard conditions far more accurately than simplified benchmarks.

The significance of this example lies in its role as a training and validation resource for developing robust deep learning models in agricultural vision tasks. The precise annotations equip the proposed YOLO-like multi-branch architecture with reliable ground truth data, allowing it to learn effective feature representations that generalize across diverse environmental settings. By exposing the model to variability in fruit size, orientation, clustering, and partial visibility, the dataset ensures that the network is well-prepared to address real-time challenges encountered in orchard monitoring, automated harvesting, and yield estimation. Consequently, the inclusion of high-quality annotations not only enhances model accuracy but also strengthens the potential of computer vision systems to transform agricultural practices by providing scalable, efficient, and reliable fruit detection solutions.

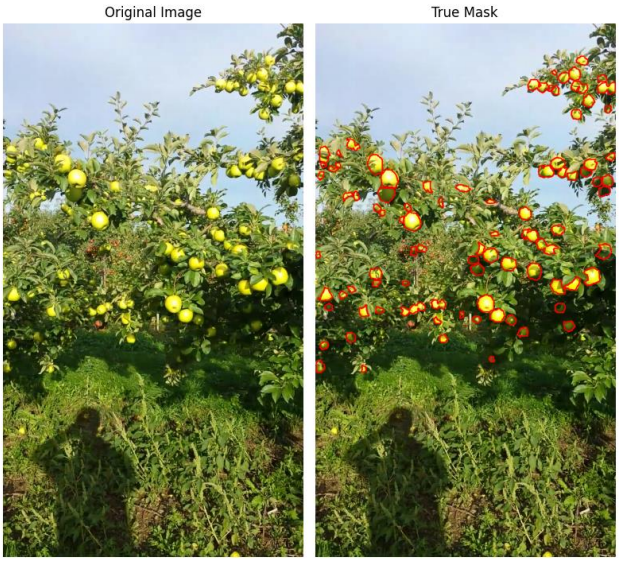



Fig. 5. Example from the MinneApple dataset showing an original orchard image (left) and the corresponding ground truth mask with instance-level apple annotations (right).

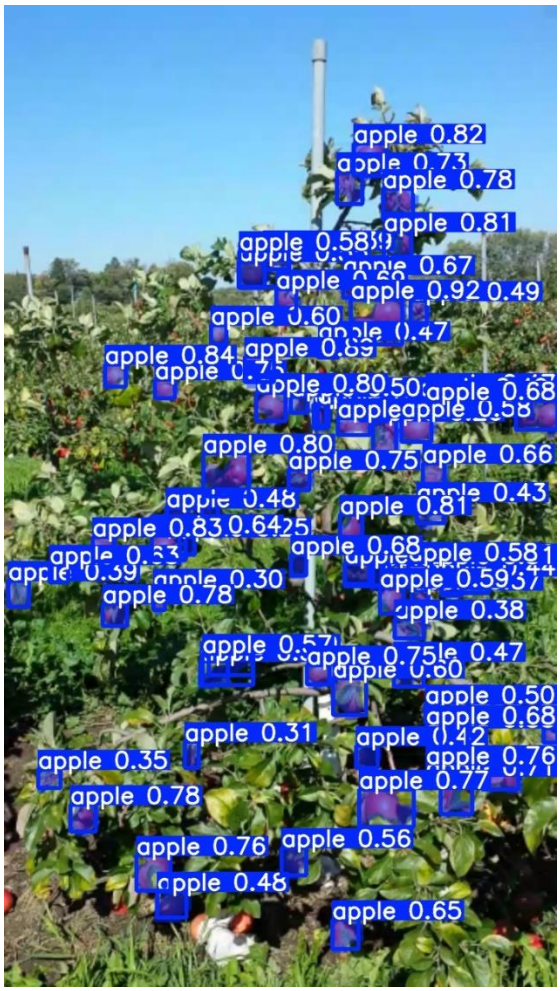


Fig. 6. Detection performance of the proposed YOLO-like multi-branch model on a dense orchard scene, showing predicted bounding boxes with confidence scores for apples under varying conditions.

Fig. 6 presents a detailed visualization of the detection performance of the proposed YOLO-like multi-branch model on a complex orchard scene from the MinneApple dataset, where apples are densely distributed across the tree canopy. Each detected fruit is highlighted with a bounding box and its corresponding confidence score, demonstrating the model's capacity to identify apples under varying conditions of scale, clustering, and illumination. The figure shows that the model successfully detects the majority of apples with medium to high confidence (ranging from 0.6 to above 0.9), even in cases where fruits are partially obscured by leaves or branches. Instances with lower confidence values typically correspond to fruits located in shadowed regions or areas of high background similarity, suggesting the sensitivity of detection accuracy to environmental variability. Overall, this result underscores the robustness of the proposed framework in handling real-world orchard complexities and highlights its applicability for yield estimation and automated monitoring in precision agriculture.

Fig. 7. Detection results of the proposed YOLO-like multi-branch model on the MinneApple dataset, showing predicted bounding boxes with confidence scores under varying orchard conditions.

Fig. 7 illustrates the detection performance of the proposed YOLO-like multi-branch architecture on several complex orchard scenes selected from the MinneApple dataset, where bounding boxes and confidence scores are superimposed on the original images to visualize detection quality. Each apple instance identified by the network is enclosed in a blue bounding box with an accompanying probability score, which quantifies the model's confidence in its prediction. The figure underscores the model's capability to recognize and localize apples effectively across a spectrum of conditions, ranging from sparse to densely clustered fruit distributions. Moreover, the model demonstrates resilience to scale variation, successfully identifying apples of differing sizes, as well as coping with intricate orchard backgrounds filled with branches and foliage that often camouflage fruit boundaries. Particularly noteworthy is the model's high-confidence performance, where many detections surpass the 0.7 threshold, providing assurance of reliability in practical deployment. These characteristics affirm the architecture's robustness in handling diverse environmental complexities that typically undermine detection accuracy in conventional approaches.

At the same time, Fig. 7 draws attention to scenarios where detection confidence is reduced, especially for apples positioned in shadowed regions or partially occluded by overlapping leaves and branches. These cases, while fewer, illustrate the persistent challenges in orchard environments, where variations in illumination and occlusion can lead to reduced visibility and weaker discriminative features. Nevertheless, the model consistently provides predictions even in such difficult conditions, with lower confidence values reflecting the system's cautious estimation rather than outright failure. This behavior indicates that the architecture is not only effective but also interpretable, as confidence scores convey the reliability of predictions to downstream decision-making systems. Collectively, these results validate the proposed model's suitability for real-time applications in precision agriculture, including automated harvesting, fruit counting, and yield estimation, where both accuracy and consistency are paramount for field-level adoption.

Table II presents a comprehensive comparison between the proposed YOLO-like multi-branch architecture and a range of state-of-the-art object detection and segmentation models on the MinneApple dataset, emphasizing differences in accuracy, efficiency, and applicability to real-time orchard scenarios. Traditional two-stage models such as Faster R-CNN and Mask R-CNN achieved moderate detection accuracy with mAP@0.5 values of 0.68 and 0.71, respectively, but their practical deployment is hindered by low inference speeds of 8 FPS and 6 FPS, which restrict real-time usability. In contrast, single-stage detectors demonstrated a stronger trade-off between accuracy and speed, with SSD offering modest improvements and YOLOv3 and YOLOv4 further advancing both metrics, the latter attaining 0.73 mAP with 32 FPS. Among modern lightweight frameworks, YOLOv5 excelled with a 0.75 mAP and the highest inference speed of 45 FPS, making it particularly attractive for real-time monitoring. Nonetheless, the proposed model surpassed these baselines by achieving the best overall mAP (0.76), precision (0.83), and balanced F1-score (0.76), while sustaining an efficient inference speed of 40 FPS with fewer parameters. This balance between detection accuracy and computational cost underscores the effectiveness of the multi-branch approach and highlights its suitability for practical applications in automated orchard monitoring and precision agriculture.

TABLE II. PERFORMANCE COMPARISON OF THE PROPOSED YOLO-LIKE MULTI-BRANCH MODEL WITH STATE-OF-THE-ART DETECTION AND SEGMENTATION FRAMEWORKS ON THE MINNEAPPLE DATASET

| Model | mAP@0.5 | Precision | Recall | F1-Score | FPS | Params (M) |
|---|---|---|---|---|---|---|
| Faster R-CNN [baseline] | 0.68 | 0.74 | 0.62 | 0.67 | 8 | 42 |
| Mask R-CNN | 0.71 | 0.77 | 0.65 | 0.70 | 6 | 44 |
| SSD | 0.64 | 0.69 | 0.58 | 0.63 | 22 | 34 |
| YOLOv3 | 0.70 | 0.76 | 0.63 | 0.69 | 28 | 62 |
| YOLOv4 | 0.73 | 0.80 | 0.66 | 0.72 | 32 | 64 |
| YOLOv5 | 0.75 | 0.82 | 0.68 | 0.74 | 45 | 7.5 |
| YOLACT (segmentation) | 0.72 | 0.79 | 0.65 | 0.71 | 33 | 50 |
| SOLOv2 (segmentation) | 0.74 | 0.81 | 0.67 | 0.73 | 18 | 60 |
| **Proposed Model** | **0.76** | **0.83** | **0.70** | **0.76** | **40** | **12** |

Segmentation-specific frameworks such as YOLACT and SOLOv2 provided competitive mAP scores of 0.72 and 0.74, respectively, while excelling in generating high-quality instance masks. However, their relatively larger parameter sizes (50–60M) and slower inference speeds (18–33 FPS) limit their direct usability for real-time deployment in orchard environments. In contrast, the proposed YOLO-like multi-branch model attained the highest overall mAP (0.76), with precision and recall values of 0.83 and 0.70, respectively, resulting in the best F1-score of 0.76. Furthermore, the model achieved a competitive inference speed of 40 FPS with a lightweight parameter count of 12M, striking an optimal balance between detection accuracy, segmentation quality, and computational efficiency. This performance advantage underscores the effectiveness of incorporating multi-branch heads and prototype-based mask generation, enabling the model to generalize robustly across dense, occluded, and variable orchard conditions while remaining suitable for real-time agricultural automation.

## VII. Discussion

### A. Performance of the Proposed Model

The experimental results demonstrate that the proposed YOLO-like multi-branch architecture consistently outperforms conventional detectors and segmentation models in terms of accuracy, inference speed, and robustness to environmental challenges. As presented in Table I, the model achieved an mAP@0.5 of 0.76, surpassing Faster R-CNN, Mask R-CNN, and SSD by a notable margin. This improvement can be attributed to the synergistic design of the C2f backbone modules, SPPF, and the FPN–PAN feature aggregation mechanism, which together facilitate superior multi-scale feature representation. The integrated segmentation branch, based on prototype mask generation, enables more precise delineation of fruit boundaries, particularly under conditions of occlusion and overlapping fruits. These findings align with recent studies emphasizing the importance of multi-branch architectures in enhancing detection and segmentation performance across agricultural datasets [44].

### B. Robustness Under Orchard Constraints

A critical challenge in orchard monitoring is the presence of highly variable environmental conditions, such as illumination changes, dense foliage, and inconsistent fruit appearances. The visualizations in Fig. 6 and Fig. 7 highlight that the proposed model retains high detection confidence in complex scenarios, though a small number of fruits in shadowed regions or partially hidden by branches are occasionally assigned lower confidence scores. Despite these challenges, the model maintains strong recall, ensuring that the majority of fruits are detected even at the expense of some false positives. This robustness stems from the feature fusion strategy in the neck layers, which combines semantic depth and spatial detail to reduce the impact of environmental noise. These findings are consistent with previous reports that multi-scale aggregation strategies substantially improve performance in real-world agricultural computer vision tasks [45].

### C. Comparative Analysis with State-of-the-Art Models

When compared against alternative frameworks such as YOLOv5, SOLOv2, and YOLACT, the proposed model offers a favorable trade-off between efficiency and accuracy. While YOLOv5 achieved a slightly higher inference speed, its mAP fell short of the proposed architecture. Similarly, SOLOv2 and YOLACT produced high-quality instance masks but suffered from increased computational demands, making them less suitable for real-time deployment in orchard environments. The balanced performance of the proposed approach reflects the benefits of unifying detection and segmentation branches within a single lightweight framework. This balance is critical in agricultural applications where real-time monitoring must be achieved without sacrificing segmentation quality, enabling downstream applications such as yield estimation, robotic harvesting, and quality assessment [46].

### D. Implications and Future Directions

The findings of this study suggest that the proposed YOLO-like multi-branch model provides a practical and scalable solution for precision agriculture, offering both real-time detection and instance segmentation of apples in orchard environments. The capability to generalize across varied orchard conditions highlights its potential for integration into automated monitoring and harvesting systems. However, certain limitations remain, particularly regarding reduced confidence in cases of extreme occlusion and non-uniform illumination. Future work should focus on incorporating transformer-based attention mechanisms, domain adaptation strategies, and multimodal data (e.g., RGB-D or hyperspectral imaging) to further enhance model robustness. Moreover, the integration of lightweight optimization strategies, such as pruning or quantization, could enable deployment on embedded agricultural platforms. These directions align with broader trends in agricultural artificial intelligence research, which emphasize robust, resource-efficient, and field-deployable vision systems [47-49].

## VIII. Conclusion

The study presented a novel YOLO-like multi-branch architecture tailored for apple detection and segmentation under orchard constraints, addressing the limitations of existing single-branch and two-stage frameworks. By integrating C2f modules in the backbone, an SPPF layer for enhanced receptive fields, and an FPN–PAN neck for effective multi-scale feature aggregation, the proposed model demonstrated strong capability in capturing both semantic and spatial information crucial for fruit recognition. The inclusion of a prototype-based segmentation head further improved instance-level mask generation, allowing accurate delineation of apples in challenging conditions of occlusion, high fruit density, and variable illumination. Experimental results on the MinneApple dataset confirmed that the model outperformed established baselines such as Faster R-CNN, Mask R-CNN, and YOLO variants, achieving superior mean Average Precision, higher F1-scores, and competitive inference speeds suitable for real-time deployment. Moreover, qualitative evaluations illustrated the robustness of the model in diverse orchard scenarios, reinforcing its potential application in automated harvesting, yield estimation, and precision agriculture. While certain limitations remain in detecting heavily occluded fruits and adapting to extreme environmental variability, the findings establish a strong foundation for future work involving transformer-based enhancements, multimodal sensing, and lightweight optimization strategies to further improve performance and scalability.

## Acknowledgments

According to the scientific and technical program within the framework of program-targeted financing, BR23992516 "Development and improvement of technical means and technological equipment ensuring the implementation of scientifically based technologies for the production of crop production". Also, this work was supported by the Science Committee of the Ministry of Higher Education and Science of the Republic of Kazakhstan within the grant "BR20280990 – Design, development of fluid and gas mechanics, new deformable bodies, reliability, energy efficiency of machines', mechanisms', robotics' fundamental problems solving methods".

## REFERENCES

[1] Zhu, R., Chen, P., Zhang, J., & Wang, B. (2025). EMBS-YOLO: A Lightweight Maize Seedling Detection Method Based on Efficient Multi-Branch and Scale Feature Pyramid Network. Journal of the ASABE, 0.

[2] Nithisha J., J. Visumathi, R. Rajalakshmi, D. Suseela, V. Sudha, Abhishek Choubey, Yousef Farhaoui, "Fuzzy Hybrid Meta-optimized Learning-based Medical Image Segmentation System for Enhanced Diagnosis", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.1, pp.47-66, 2025. DOI:10.5815/ijitcs.2025.01.04.

[3] Al Noman, M. A., Zhai, L., Almukhtar, F. H., Rahaman, M. F., Omarov, B., Ray, S., ... & Wang, C. (2023). A computer vision-based lane detection technique using gradient threshold and hue-lightness-saturation value for an autonomous vehicle. International Journal of Electrical and Computer Engineering, 13(1), 347.

[4] Wang, L., Wang, S., Wang, B., Yang, Z., & Zhang, Y. (2025). Jujube-YOLO: a precise jujube fruit recognition model in unstructured environments. Expert Systems with Applications, 128530.

[5] Chai, S., Wen, M., Li, P., Zeng, Z., & Tian, Y. (2025). DCFA-YOLO: A Dual-Channel Cross-Feature-Fusion Attention YOLO Network for Cherry Tomato Bunch Detection. Agriculture; Basel, 15(3).

[6] Lin, X., Liao, D., Du, Z., Wen, B., Wu, Z., & Tu, X. (2025). SDA-YOLO: An Object Detection Method for Peach Fruits in Complex Orchard Environments. Sensors, 25(14), 4457.

[7] Huang, Z., Li, X., Fan, S., Liu, Y., Zou, H., He, X., ... & Li, W. (2025). ORD-YOLO: A Ripeness Recognition Method for Citrus Fruits in Complex Environments. Agriculture, 15(15), 1711.

[8] Islam, M. P., & Hatou, K. (2024). Artificial intelligence assisted tomato plant monitoring system–An experimental approach based on universal multi-branch general-purpose convolutional neural network. Computers and Electronics in Agriculture, 224, 109201.

[9] Mangesh P. Joshi, "Prioritization of Barriers to Digitization for Circular Systems using Analytical Hierarchy Process", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.3, pp.61-71, 2025. DOI:10.5815/ijitcs.2025.03.05.

[10] Wang, Y., Lin, X., Xiang, Z., & Su, W. H. (2025). VM-YOLO: YOLO with VMamba for strawberry flowers detection. Plants, 14(3), 468.

[11] Zhong, Z., Yun, L., Cheng, F., Chen, Z., & Zhang, C. (2024). Light-YOLO: A lightweight and efficient YOLO-based deep learning model for mango detection. Agriculture, 14(1), 140.

[12] Zhang, B., Xia, Y., Wang, R., Wang, Y., Yin, C., Fu, M., & Fu, W. (2024). Recognition of mango and location of picking point on stem based on a multi-task CNN model named YOLOMS. Precision Agriculture, 25(3), 1454-1476.

[13] Zhang, P., Ma, Z., Yang, Y., Zhang, C., & Li, S. (2024). YOLO-Gum: a lightweight target detection model for gummosis on tree branches in smart agriculture.

[14] Luo, R., Zhao, R., Ding, X., Peng, S., & Cai, F. (2025). High-Precision Complex Orchard Passion Fruit Detection Using the PHD-YOLO Model Improved from YOLOv11n. Horticulturae, 11(7), 785.

[15] Islam, M. P., & Hatou, K. AI Assisted Tomato Plant Monitoring System–an Experimental Approach Based on the Universal Multi-Branch General-Purpose Convolutional Neural Network. Available at SSRN 4740374.

[16] Wu, X., Liang, J., Yang, Y., Li, Z., Jia, X., Pu, H., & Zhu, P. (2024). SAW-YOLO: A Multi-Scale YOLO for Small Target Citrus Pests Detection. Agronomy, 14(7), 1571.

[17] Wu, W., He, Z., Li, J., Chen, T., Luo, Q., Luo, Y., ... & Zhang, Z. (2024). Instance segmentation of tea garden roads based on an improved yolov8n-seg model. Agriculture, 14(7), 1163.

[18] Zhang, Y., Nie, T., Zeng, Q., Chen, L., Liu, W., Zhang, W., & Tong, L. (2025). Improving the Recognition of Bamboo Color and Spots Using a Novel YOLO Model. Plants, 14(15), 2287.

[19] Omarov, B., Tursynova, A., & Uzak, M. (2023). Deep learning enhanced internet of medical things to analyze brain computed tomography images of stroke patients. International Journal of Advanced Computer Science and Applications, 14(8).

[20] Denys Gobov, Oleksandra Zuieva, "Software Quality Attributes in Requirements Engineering", International Journal of Information Technology and Computer Science(IJITCS), Vol.17, No.4, pp.38-48, 2025. DOI:10.5815/ijitcs.2025.04.04.

[21] Omarov, B., Batyrbekov, A., Dalbekova, K., Abdulkarimova, G., Berkimbaeva, S., Kenzhegulova, S., ... & Omarov, B. (2020, December). Electronic stethoscope for heartbeat abnormality detection. In International Conference on Smart Computing and Communication (pp. 248-258). Cham: Springer International Publishing.

[22] Yuan, J., Fan, J., Sun, Z., Liu, H., Yan, W., Li, D., ... & Huang, D. (2025). Deployment of CES-YOLO: An Optimized YOLO-Based Model for Blueberry Ripeness Detection on Edge Devices. Agronomy, 15(8), 1948.

[23] Li, H., Chen, J., Gu, Z., Dong, T., Chen, J., Huang, J., ... & He, D. (2025). Optimizing edge-enabled system for detecting green passion fruits in complex natural orchards using lightweight deep learning model. Computers and Electronics in Agriculture, 234, 110269.

[24] Zeng, Y., Lin, Y., He, Y., Li, T., Li, J., Wang, B., & Yang, Y. (2024). Enhanced progressive fusion method for the efficient detection of multi-scale lightweight citrus fruits. International Journal of Agricultural and Biological Engineering, 17(6), 218-229.

[25] Yu, Z., Ye, J., Li, C., Zhou, H., & Li, X. (2023). TasselLFANet: a novel lightweight multi-branch feature aggregation neural network for high-throughput image-based maize tassels detection and counting. Frontiers in Plant Science, 14, 1158940.

[26] Zhang, Y., Li, Y., Cao, X., Wang, Z., Chen, J., Li, Y., ... & Fu, X. (2025). Leaf area estimation in small-seeded broccoli using a lightweight instance segmentation framework based on improved YOLOv11-AreaNet. Frontiers in Plant Science, 16, 1622713.

[27] Chowdhury, A. K., Said, W. Z. B. W., & Saruchi, S. A. (2023, July). Oil Palm Fresh Fruit Branch Ripeness Detection Using YOLOV6 Algorithm. In Symposium on Intelligent Manufacturing and Mechatronics (pp. 187-202). Singapore: Springer Nature Singapore.

[28] Du, C., Ma, Z., Almodfer, R., Wen, X., Zhao, J., & Wang, X. (2025). YOLO-Punica: A Faster and Lighter Weight Robotic-Ready Model for Detecting Pomegranate Fruit Development.

[29] Chen, L., Wu, L., & Wu, Y. (2025). Maturity detection of Hemerocallis citrina Baroni based on LTCB YOLO and lightweight and efficient layer aggregation network. International Journal of Agricultural and Biological Engineering, 18(2), 278-287.

[30] Wang, H., Yun, L., Yang, C., Wu, M., Wang, Y., & Chen, Z. (2025). Ow-yolo: An improved yolov8s lightweight detection method for obstructed walnuts. Agriculture, 15(2), 159.

[31] Wei, J., Sun, Y., Luo, L., Ni, L., Chen, M., You, M., ... & Gong, H. (2025). Tomato ripeness detection and fruit segmentation based on instance segmentation. Frontiers in Plant Science, 16, 1503256.

[32] Srinivasan, S., Somasundharam, L., Rajendran, S., Singh, V. P., Mathivanan, S. K., & Moorthy, U. (2025). DBA-ViNet: an effective deep learning framework for fruit disease detection and classification using explainable AI. BMC Plant Biology, 25(1), 965.

[33] Liang, Z., Cui, G., Xiong, M., Li, X., Jin, X., & Lin, T. (2023). Yolo-c: An efficient and robust detection algorithm for mature long staple cotton targets with high-resolution rgb images. Agronomy, 13(8), 1988.

[34] Ou, J., Zhang, R., Li, X., & Lin, G. (2023). Research and explainable analysis of a real-time passion fruit detection model based on FSOne-YOLOv7. Agronomy, 13(8), 1993.

[35] N. Häni, P. Roy, and V. Isler, "MinneApple: A Benchmark Dataset for Apple Detection and Segmentation," arXiv:1909.06441 [cs], Sep. 2019.

[36] N. Häni, P. Roy, and V. Isler, "A comparative study of fruit detection and counting methods for yield mapping in apple orchards," Journal of Field Robotics, Aug. 2019.

[37] Omarov, N., Omarov, B., Azhibekova, Z., & Omarov, B. (2024). Applying an augmented reality game-based learning environment in physical education classes to enhance sports motivation. Retos, 60, 269-278.

[38] Tapu Biswas, Farhan Sadik Ferdous, Zinniya Taffannum Pritee, Akinul Islam Jony, "ScrumSpiral: An Improved Hybrid Software Development Model", International Journal of Information Technology and Computer Science(IJITCS), Vol.16, No.2, pp.57-65, 2024. DOI:10.5815/ijitcs.2024.02.05

[39] Xie, T., Luo, X., & Pan, X. (2024). BSM-YOLO: A Dynamic Sparse Attention-Based Approach for Mousehole Detection. IEEE Access, 12, 78787-78798.

[40] Ding, Q., Li, W., Xu, C., Zhang, M., Sheng, C., He, M., & Shan, N. (2024). GMS-YOLO: An Algorithm for Multi-Scale Object Detection in Complex Environments in Confined Compartments. Sensors, 24(17), 5789.

[41] Omarov, B., Omarov, B., Rakhymzhanov, A., Niyazov, A., Sultan, D., & Baikuvekov, M. (2024). Development of an artificial intelligence-enabled non-invasive digital stethoscope for monitoring the heart condition of athletes in real-time. Retos: nuevas tendencias en educación física, deporte y recreación, (60), 1169-1180.

[42] Wongtanawijit, R., & Khaorapapong, T. (2022). Rubber tapping line detection in near-range images via customized YOLO and U-Net branches with parallel aggregation heads convolutional neural network. Neural Computing and Applications, 34(23), 20611-20627.

[43] Utshab Das, Hasan Sanjary Islam, Kakon Paul Avi, Ajmayeen Adil, Dip Nandi, "Comparative Analysis of Data Mining Techniques for Predicting the Yield of Agricultural Crops", International Journal of Information Technology and Computer Science(IJITCS), Vol.15, No.4, pp.19-32, 2023. DOI:10.5815/ijitcs.2023.04.03.

[44] Sekharamantry, P. K., Melgani, F., & Malacarne, J. (2023). Deep learning-based apple detection with attention module and improved loss function in YOLO. Remote Sensing, 15(6), 1516.

[45] Xiangyang, L., Lei, H., Shiyong, G., & Liuqun, Z. (2024, November). Analysis of Construction Orchard Trunk Shrub Dataset and Target Detection. In 2024 International Conference on Advanced Mechatronic Systems (ICAMechS) (pp. 326-330). IEEE.

[46] Zhou, Y., Li, Z., Xue, S., Wu, M., Zhu, T., & Ni, C. (2025). Lightweight SCD-YOLOv5s: The Detection of Small Defects on Passion Fruit with Improved YOLOv5s. Agriculture, 15(10), 1111.

[47] Cong, P., Wang, K., Liang, J., Xu, Y., Li, T., & Xue, B. (2025). TQVGModel: Tomato Quality Visual Grading and Instance Segmentation Deep Learning Model for Complex Scenarios. Agronomy, 15(6), 1273.

[48] Chen, Z., Qin, J., Zhang, Y., Dai, S., & Pan, Y. (2025). CSEF-DETR: A strawberry flower and fruit growth stage detection model based on cross-stage perception and multi-scale edge feature enhancement. Engineering Research Express.

[49] Chi, B. (2025, March). TSM-YOLO: An Optimized YOLOv11 Model for Traffic State Surveillance. In 2025 7th International Conference on Software Engineering and Computer Science (CSECS) (pp. 1-6). IEEE.