frontiers | Frontiers in Mechanical Engineering

TYPE Original Research
PUBLISHED 21 January 2026
DOI 10.3389/fmech.2026.1741396

OPEN ACCESS

EDITED BY
Hamid Reza Karimi,
Polytechnic University of Milan, Italy

REVIEWED BY
Yingxing Jiang,
Jiangsu University, China
Peng Huo,
Inner Mongolia Agricultural University, China

*CORRESPONDENCE
Yingwu Xu,
godfatherwww@163.com

RECEIVED 07 November 2025
REVISED 04 January 2026
ACCEPTED 05 January 2026
PUBLISHED 21 January 2026

CITATION
Xu Y (2026) Harvesting target positioning and robotic arm obstacle avoidance algorithm based on improved YOLOv8 and BIT*. *Front. Mech. Eng.* 12:1741396. doi: 10.3389/fmech.2026.1741396

Harvesting target positioning and robotic arm obstacle avoidance algorithm based on improved YOLOv8 and BIT*

Yingwu Xu*

Anqing Vocational and Technical College, Anqing, China

Introduction: To address the core challenges of inaccurate fruit occlusion localization and inefficient robotic arm dynamic obstacle avoidance in complex, unstructured agricultural environments, this study proposes an integrated algorithm for harvesting.
Methods: The proposed algorithm is built upon an improved YOLOv8 model and the BIT* planner. The YOLOv8 model was enhanced by introducing the Swin Transformer module to improve multi-scale feature fusion and global context modeling. The BIT* planner was integrated with a BiLSTM network to endow it with dynamic obstacle prediction capabilities, thereby constructing a unified architecture for visual perception and motion planning.
Results: Experimental results demonstrated that the algorithm achieved real-time performance with a processing frame rate of 32.7 fps and an inference time of 32.6 ms for target localization, with a localization error standard deviation as low as 1.70 mm. In obstacle avoidance planning, it achieved a balance with manipulator energy consumption of 124.58 J, while controlling the computational load and memory resource consumption per task to 22.7 GFlops and 187 MB, respectively.
Discussion: This approach provides a high-precision, low-energy-consumption cooperative control solution for agricultural harvesting robots, advancing the practical application of automated fruit and vegetable harvesting.

KEYWORDS

agriculture, automated harvesting, BIT*, robotic arm, YOLOv8

1 Background

Harvesting is one of the most labor-intensive and time-consuming steps in the production of fruits and vegetables. Its level of automation and intelligence directly impacts production efficiency, cost control, and industrial upgrading (Liu and Liu, 2024). Therefore, developing efficient, precise, and autonomous intelligent harvesting robot systems holds significant practical and economic value for freeing up labor, advancing agricultural modernization, and ensuring food security (Zhou et al., 2022). Among these, the precise target localization of the perception module and the dexterous obstacle avoidance path planning of the execution module represent two critical technological bottlenecks determining system performance (Zeeshan and Aized, 2023). In complex, unstructured natural field environments, harvesting targets (such as fruits) are frequently disrupted by factors including variable lighting, foliage obstruction, similar colors and textures, variable scales, and overlapping clusters. This places extremely high demands on the robustness and accuracy of visual detection algorithms (Panduranga et al.,

2024). Simultaneously, when executing grasping tasks, robotic arms must navigate dense, intertwined crop branches to plan collision-free, highly efficient trajectories. Any planning failure or delay may lead to task interruption or crop damage. The performance and dependability of path planning algorithms in real-time are severely hampered by this (Droukas et al., 2023).

To address these challenges, numerous experts in the field of smart agriculture have embarked on exploratory research. To overcome the difficulties of identifying clustered tomato fruits and selecting the best picking locations in challenging situations, Bai et al. (2023) developed a two-step localization technique that integrated multi-feature extraction and geometry analysis for target recognition in harvesting. This approach could achieve precise fruit region identification and accurate stem-picking point localization (Bai et al., 2023). To address the challenge of accurately detecting tomato fruits and stems in complex agricultural environments, Miao et al. (2023) proposed an integrated detection algorithm combining traditional image processing with you only look once version 5 (YOLOv5). Through multi-method fusion and error compensation strategies, this research could achieve precise determination of tomato ripeness and accurate stem localization, providing reliable guidance for efficient robotic harvesting (Miao et al., 2023). Gong et al. (2022) suggested a geometric feature reconstruction technique based on multi-source image fusion and an extended mask region-based convolutional neural network (Mask R-CNN) to address the problem of inadequate visual positioning accuracy in fruit-picking robots operating in obscured situations. By integrating multi-source image registration with shape-position recovery algorithms, this approach could achieve high-precision 3D geometric reconstruction and picking point localization for occluded tomatoes (Gong et al., 2022). To address the high labor costs and fruit identification/localization challenges in strawberry picking, Hu et al. (2022) proposed a recognition and localization method integrating instance segmentation with stereo vision. By combining a dual-network architecture of Mask R-CNN and YOLOv3 with the 3D localization technology of the Zeid stereo vision camera, this research could achieve precise identification and 3D spatial localization of ripe strawberries, providing accurate target location information for picking robots (Hu et al., 2022).

To solve the problems of excessive path planning time and low picking efficiency in unstructured orchard environments, Zhang et al. (2024) suggested a heuristic dynamic rapidly-exploring random tree connect (HDRRT) motion planning algorithm for robotic arms obstacle avoidance planning obstacle avoidance planning. By using a dual-structure strategy that combined heuristic dynamic step size strategies and adaptive target gravity, this study could successfully decrease path planning time and path cost while increasing planning success rates (Zhang et al., 2024). Liu (2022) addressed the low efficiency of apple-picking robots in unstructured orchard situations by proposing the hierarchical optimal path planning (HOPP) method. This study significantly reduced the computational time required for three-dimensional picking path planning by combining a two-layer structure with distance-constrained K-means clustering and traveling salesman problem solutions. This approach achieved globally optimal harvesting path planning for multi-objective fruit harvesting (Liu, 2022). A view planner based on an active vision technique was proposed by Yi et al. (2024) to solve the problem of accurately localizing fruit-picking points in heavily obstructed settings. Through a three-step structure, including candidate view generation, spatial coverage score function optimization, and iterative viewpoint adjustment, this research effectively addressed stem occlusion issues, significantly improving the robot's picking success rate and operational efficiency (Yi et al., 2024). Xu et al. (2021) proposed an improved artificial potential field algorithm to address the issues of local minima and insufficient obstacle shape perception in traditional methods for robotic arms 3D path planning. By incorporating a repulsive isopotential surface movement mechanism and a local path optimization structure, this research effectively resolved local minima traps and enabled obstacle shape perception, significantly enhancing path planning success rates and motion smoothness (Xu et al., 2021). In summary, existing research exhibits a typical architecture characterized by "decoupling perception and planning modules" in its technical approach. Its core advantages lie in its perception layer. Techniques such as multi-source information fusion, the integration of traditional and deep learning, and stereo vision effectively enhance the robustness of target recognition and the accuracy of positioning for fruits and vegetables in static environments. At the planning layer, strategies including heuristic random sampling, hierarchical task decomposition, and active perception decision-making significantly optimize path cost and static obstacle avoidance success rates. However, this architecture has fundamental limitations. The perception and planning stages operate in an unidirectional, open-loop manner. They lack real-time visual feedback adjustments based on motion states. The visual module exhibits insufficient generalization capabilities against dynamic occlusions and sudden lighting changes. Moreover, the planning module generally lacks explicit modeling and prediction of dynamic obstacle movement trends. Consequently, the system faces constraints in overall adaptability, real-time responsiveness, and closed-loop stability within highly unstructured, dynamically changing field environments.

YOLOv8 extracts features through a backbone network (BN), fuses multi-scale information via a neck network, and finally performs both bounding box (BOB) regression and classification prediction simultaneously through a detection head (Li et al., 2024). Batch informed trees* (BIT*) combines graph search with random sampling, pruning ineffective regions using heuristic information, and progressively optimizes path costs through iterative batch processing (Kyaw et al., 2022). However, YOLOv8 exhibits insufficient perception of occluded objects and small fruit stems. BIT* lacks a mechanism for reacting to dynamic barriers and has poor processing efficiency in high-dimensional areas (Xu and Li, 2025; Tamizi et al., 2024). XAmong them, the perception module uses YOLOv8 as its framework and incorporates the Swin Transformer as its BN. Its sliding window attention mechanism improves the accuracy of fruit target recognition and localization in complex occlusion environments by enhancing multi-scale feature fusion (MSFF) and global context modeling. The planning module utilizes the BIT* framework, integrating a BiLSTM network to predict dynamic obstacle movement trends. Temporal modeling enhances the robotic arm's foresight and adaptability in path search, enabling efficient and smooth obstacle avoidance in dynamic, unstructured environments. Both modules achieve information integration through hand-eye calibration and coordinate

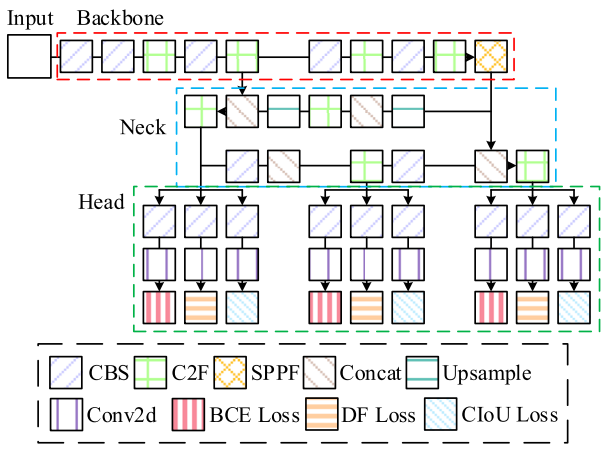

FIGURE 1
Schematic illustration of the YOLOv8 architecture.

transformation, ultimately forming a unified "perception-decision-control" collaborative system. This approach ensures positioning accuracy and planning efficiency while significantly reducing computational and energy consumption costs. Its innovation lies in its ability to achieve synergistic breakthroughs in perception, decision-making, and control. This is accomplished through multi-scale feature enhancement, spatio-temporal context modeling, adaptive sampling strategies, and dynamic cost function optimization.

2 Methodology

This section comprises two parts. The first part introduces the Swin Transformer module based on the YOLOv8 object detection framework to construct a rapid and precise fruit-picking target localization module. It enhances fruit recognition accuracy (RA) in complex environments through MSFF and global context modeling. The second part combines the BiLSTM's temporal prediction capabilities with the BIT* path planning algorithm to develop a RAOA module with dynamic obstacle response capabilities. Finally, the two modules are integrated through hand-eye calibration and coordinate transformation mechanisms to form a complete vision-motion control closed-loop system. This realizes the YOLOv8-B* algorithm architecture from fruit recognition to picking path planning.

2.1 Harvesting target positioning module based on YOLOv8

In automated harvesting systems for fruits and vegetables, robotic arms serve as the core execution units. Their grasping success rate and operational efficiency heavily depend on the precise spatial localization of target fruits. Accurate, real-time identification and localization of fruit positions are fundamental prerequisites for achieving damage-free grasping while avoiding collisions and mispicks. Consequently, this study employs YOLOv8 as the foundation for target localization during harvesting operations. YOLOv8 is selected as the core visual localization framework primarily due to its classic balance in object detection tasks, robust multi-scale feature extraction (MSFE) capabilities, and potential for lightweight deployment. Its efficient cross stage partial network with feature fusion (C2F) architecture and decoupled detector head design provide a stable and scalable baseline. Compared to subsequent versions that focus on specific tasks or architectures, YOLOv8 has broader industrial deployment validation and more experience with lightweight optimization. This makes it better suited for agricultural embedded scenarios with dual constraints on reliability and computational resources (Ma et al., 2024). The architecture of YOLOv8 is illustrated in Figure 1.

In Figure 1, the YOLOv8 network architecture primarily consists of three components: the BN, the neck network, and the detection head. It achieves MSFE and fusion through modules such as convolution + batchnorm + sigmoid, C2F, and spatial pyramid pooling fast (Gao et al., 2024). The detection performance of YOLOv8 relies on optimizing the total loss function. The model learns end-to-end by minimizing the discrepancy between projected values and ground truth annotations while concurrently predicting item BOB coordinates, category labels, and object presence confidence scores during training (Gao et al., 2023). Equation 1 illustrates that the weighted sum of the three terms is the definition of the total loss function.

$$L^{\text{total}} = \lambda_1 L^{\text{cls}} + \lambda_2 L^{\text{box}} + \lambda_3 L^{\text{obj}} \tag{1}$$

In Equation 1, L^{cls} represents the classification loss. L^{box} denotes the BOB regression loss. L^{obj} signifies the object confidence loss. $\{\lambda_1, \lambda_2, \lambda_3\}$ serves as the weighting coefficient for each loss term, balancing the optimization scales across different tasks. Specifically, L^{cls} employs binary cross-entropy (BCE) loss to calculate the discrepancy between predicted and ground-truth categories. L^{box} utilizes a combination of distribution focal (DF) loss and complete intersection over union (CIoU) loss. While CIoU thoroughly takes into account overlap area, center point distance, and aspect ratio to obtain more accurate BOB regression, DF optimizes the focused distribution of BOB position probability. L^{obj} also employs BCE loss to determine whether an object exists within the BOB (Ayyad et al., 2025). Localization results can be directly output as fruit center coordinates and size information for subsequent robotic arms motion planning and grasp pose estimation.

However, the CNN backbone of YOLOv8 has limited capabilities for modeling global contextual information and long-range dependencies. The ST achieves powerful global modeling capabilities while maintaining computational efficiency through its sliding window mechanism (Pal et al., 2023). Therefore, this study incorporates the ST into the BN of YOLOv8 to enable more precise feature extraction and localization of occluded or densely clustered objects in complex environments. Figure 2 depicts the structure of the ST.

In Figure 2, the ST adopts a hierarchical architecture. Based on window-based multi-head self-attention (W-MSA) and shifted window MSA (SW-MSA), it constructs a general-purpose BN capable of efficiently processing visual tasks. Its core lies in the W-MSA computation, where the standard self-attention (SA) calculation is expressed in Equation 2.

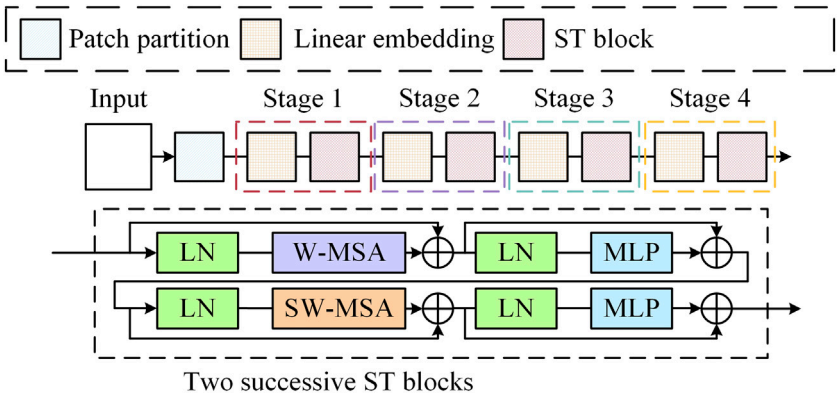

FIGURE 2
Schematic illustration of the ST architecture.

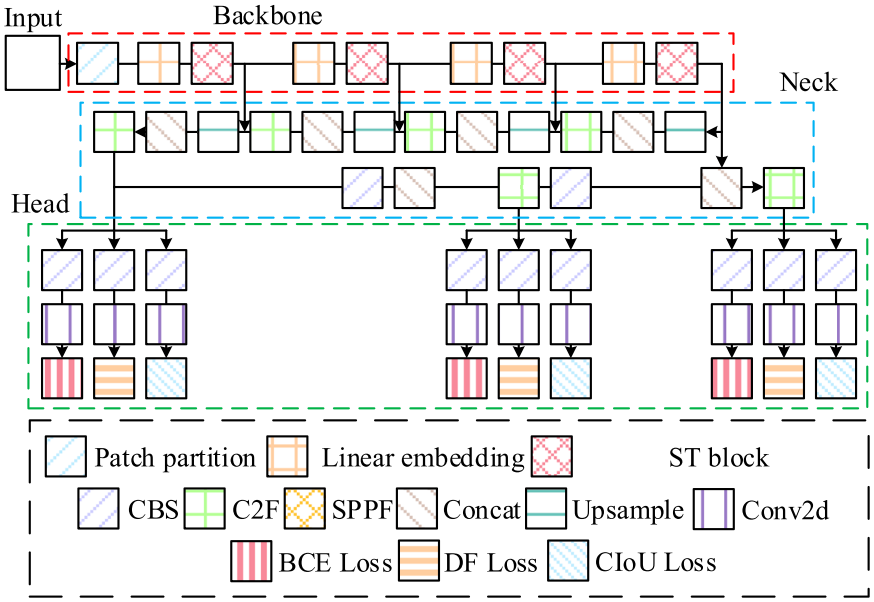

FIGURE 3
Schematic illustration of the harvesting target positioning module architecture.

$$\text{Attention}\,(Q,K,V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}} + B\right)V \tag{2}$$

In Equation 2, $\{Q, K, V\}$ represents the query, key, and value matrices. d_k denotes the dimension of the key vector. $\sqrt{d_k}$ is used to scale the dot product results, preventing softmax gradient saturation. B is the relative position bias, introducing spatial position priors for each attention head to enhance the model's perception of geometric structures. To greatly reduce computational complexity, the ST splits the input image into non-overlapping windows and calculates SA within each window (Wang et al., 2023). To further enable cross-window connections, the alternately applied SW-MSA shifts window partitions, allowing attention computations to extend beyond original window boundaries. Equation 3 can be used to represent two consecutive ST blocks.

$$\begin{cases} \hat{\mathbf{z}}^l = \text{W}-\text{MSA}\left(\text{LN}\left(\mathbf{z}^{l-1}\right)\right) + \mathbf{z}^{l-1} \\ \mathbf{z}^l = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^l\right)\right) + \hat{\mathbf{z}}^l \\ \hat{\mathbf{z}}^{l+1} = \text{SW}-\text{MSA}\left(\text{LN}\left(\mathbf{z}^l\right)\right) + \mathbf{z}^l \\ \mathbf{z}^{l+1} = \text{MLP}\left(\text{LN}\left(\hat{\mathbf{z}}^{l+1}\right)\right) + \hat{\mathbf{z}}^{l+1} \end{cases} \tag{3}$$

In Equation 3, $\{\mathbf{z}^{l-1}, \mathbf{z}^l, \mathbf{z}^{l+1}\}$ represents the output features of layers $l-1$, l, and $l+1$. $\{\hat{\mathbf{z}}^l, \hat{\mathbf{z}}^{l+1}\}$ denotes the residual output after the MSA module. LN indicates the layer normalization (LN) operation. MLP refers to the multilayer perceptron (MLP), which performs nonlinear transformation and feature enhancement. This architecture ensures trainability in deep networks through residual connections and LN, while progressively integrating local and global information at each stage via the alternating W-MSA and SW-MSA mechanism (Tang et al., 2025). Consequently, the study

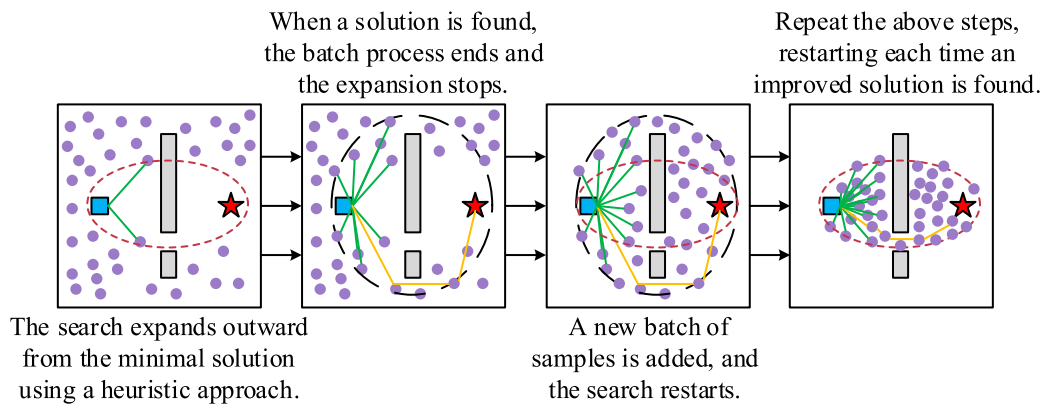

FIGURE 4
Schematic of the BIT* operating flow.

centers on introducing the ST-based YOLOv8 to construct a harvesting target positioning module. Its structure is illustrated in Figure 3.

In Figure 3, this module centers on the YOLOv8 network, replacing the original backbone with a ST for deep feature extraction to enhance representation capabilities for occluded and small target fruits. A feature pyramid network (FPN) and path aggregation network (PANet) structure make up the neck after the backbone, allowing multi-scale feature propagation using both top-down and bottom-up methods. Lastly, a decoupled detection head allows the independent prediction of object category confidence scores and precise geographical coordinates by separating the classification work from the BOB regression task. Through these enhancements, the harvesting target positioning module achieves accurate fruit object recognition and highly reliable localization in complex agricultural environments, providing high-quality visual input for subsequent robotic arms grasping planning.

2.2 Obstacle avoidance module for robotic arms based on BIT* and YOLOv8-B* algorithm construction

The harvesting target positioning module developed in this study achieves high-precision spatial localization of fruit targets. However, its output provides only static coordinate information and lacks dynamic path planning capabilities for robotic arms movements. In unstructured orchard environments, effective obstacle avoidance along the robotic arms' path is crucial for successful harvesting. BIT* significantly enhances RRT's convergence efficiency through batch sampling and heuristic pruning mechanisms. Its incremental graph update structure continuously integrates real-time perception data to adapt to dynamic environments. Unlike gradient-based optimization or data-driven planning methods, BIT does not require differentiable environment models or large-scale labeled trajectories. Through state space sampling and pruning, it achieves robust and efficient dynamic obstacle avoidance in unstructured scenarios. Therefore, this study utilizes BIT* as the foundation for RAOA operations. Figure 4 provides an illustration of its operational procedures (Nenavath and Perumal, 2024).

In Figure 4, the operational flow of BIT* constitutes an iterative batch sampling process. It intelligently expands sampling batches within the state space and searches for random geometric configurations to identify and continuously optimize paths. BIT* explores the solution space by maintaining a tree structure $\mathcal{T} = (V, E)$. Among these, the vertex set V represents explored states, while the edge set E denotes feasible paths between states. Its core lies in generating a sampling batch during each iteration and computing heuristic values to guide the search direction. For any configuration q formed by the joint angles of an arbitrary robotic arms, its heuristic value is jointly determined by the cost c_{current} of the current solution and the estimated cost (EC) $\hat{h}(q)$ to the target. The algorithm first constructs two search trees from the start and target points, respectively, and continuously performs heuristic sorting, as shown in Equation 4 (Huynh et al., 2023).

$$\begin{cases} f(q) = g(q) + \hat{h}(q) \\ v(q) = \min\left(c_{\text{current}}, g(q) + \hat{h}(q)\right) \end{cases} \tag{4}$$

In Equation 4, $g(q)$ represents the actual path cost from the starting point q_{start} to the current state q (e.g., path length (PL)). $\hat{h}(q)$ denotes the heuristic EC from q to the target point q_{goal}, typically using Euclidean distance. $f(q)$ is used to prioritize candidate expansion nodes, favoring exploration in potentially optimal path directions. c_{current} represents the total path cost of currently known feasible solutions. $v(q)$ denotes the upper bound on path cost achievable via node F, used for ranking and pruning (Xu et al., 2022). In each batch processing, if $v(q) > c_{\text{current}}$ holds, it indicates that the node cannot produce a better solution and is pruned. The algorithm only expands vertices that satisfy $f(q) \le c_{\text{current}}$ and $v(q) \le c_{\text{current}}$, thereby effectively pruning search regions unlikely to improve the current solution. Whenever a new solution or a better solution is found, c_{current} is updated, and the search restarts to find a better path on a more finely sampled graph (Johnson et al., 2023).

However, the standard BIT* algorithm is primarily optimized for static environments and struggles to effectively handle dynamic changes such as leaf swaying in orchards. Long-range relationships

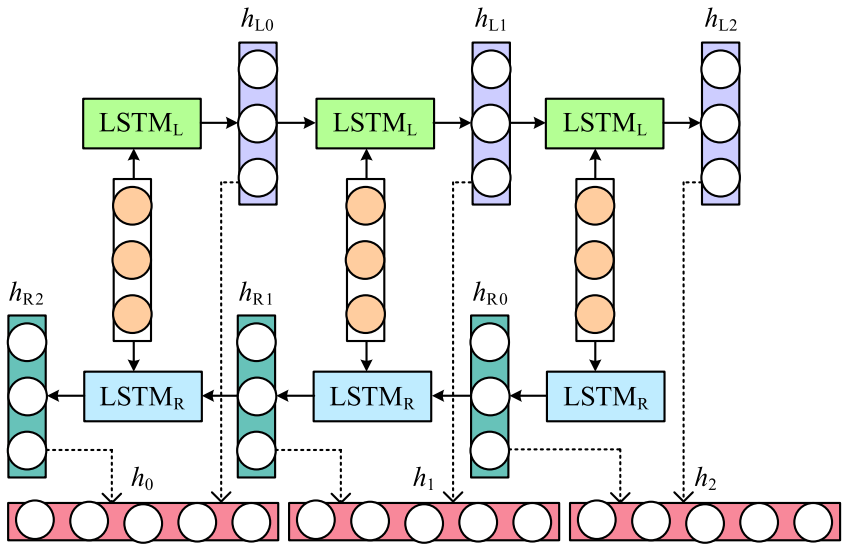

FIGURE 5
Schematic illustration of the BiLSTM architecture.

and forward-backward contextual information in time-series data can be effectively captured by BiLSTM thanks to its special bidirectional gated recurrent structure (Yu et al., 2024). Therefore, this study introduces BiLSTM into BIT*. Its core function is to capture the temporal movement patterns of dynamic obstacles. It learns trends in direction and velocity changes from historical trajectories through a bidirectional gating mechanism, enabling predictions of future positions within short time intervals. These predictions serve as prior knowledge that is fed into the BIT* algorithm. This allows the algorithm to proactively avoid areas where dynamic obstacles are expected to be during the path search. This enhances the planning system's foresight and improves the success rate of dynamic obstacle avoidance. The structure of BiLSTM is shown in Figure 5.

In Figure 5, the BiLSTM consists of two independent LSTM layers, forward and backward, which process the sequence input in the forward and reverse directions, respectively. The hidden state (HS) outputs from both directions are ultimately combined to capture the full contextual information. The core of the BiLSTM is its gating mechanism. Its computational steps involve the forget gate f_t, the input gate i_t, the output gate o_t, and cell state (CS) updates. At time step t, the forward LSTM (denoted as LSTM_L) first determines which information should be forgotten and which new information needs to be stored, as shown in Equation 5 (Kumudham et al., 2024).

$$\begin{cases} f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right) \\ i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right) \\ \hat{C}_t = tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \\ C_t = f_t{}^* C_{t-1} + i_t{}^* \hat{C}_t \end{cases} \tag{5}$$

In Equation 5, x_t represents the current input. $\{b_f, b_i, b_C\}$ denotes the corresponding bias. $\{W_f, W_i, W_i\}$ signifies the corresponding weight. h_{t-1} indicates the HS from the previous time step. f_t determines which information from the previous CS C_{t-1} should be retained or forgotten. i_t, together with the candidate CS $\hat{C}_t$, jointly determines which information needs to be updated into the CS at the current time step. C_t is the current CS, computed jointly by f_t, C_{t-1}, and $\hat{C}_t$. Next, based on the updated CS, LSTM_L computes o_t and the current HS h_t, as shown in Equation 6.

$$\begin{cases} o_t = \sigma\left(W_o \cdot [h_{t-1}, x_t] + b_o\right) \\ h_t = o_t{}^* tanh\left(C_t\right) \end{cases} \tag{6}$$

In Equation 6, W_o and b_o represent the weights and bias of o_t, respectively. For the reverse LSTM (denoted as LSTM_R), it is computed in the same manner but operates in reverse along the time series, thereby generating the reverse HS $h_t^{'}$ (Zhai et al., 2024). Finally, the output of the BiLSTM at time step t is the concatenation of the forward HS h_t and the backward HS $h_t^{'}$, yielding $y_t = [h_t, h_t^{'}]$. This enables the model to fuse bidirectional contextual information across the entire sequence. The BiLSTM takes as input a time-based, sliding-window sequence of dynamic obstacle states, each of which typically contains three-dimensional position coordinates. This sequence is continuously acquired and provided by the system during operation through its real-time perception and tracking module. The network's final output is a predicted sequence of dynamic obstacle positions over several future planning cycles. This sequence is converted into a dynamic cost map that directly guides the generation of collision-free trajectories for the BIT* search. Consequently, this study investigates the BIT* based on the fusion capabilities of the BiLSTM for temporal prediction, constructing a RAOA module. Figure 6 displays its structure.

In Figure 6, this module employs the BIT* algorithm as its core framework. Through its iterative batch sampling and heuristic pruning mechanisms, it achieves efficient and asymptotically optimal path planning for robotic arms in complex, unstructured environments. This module integrates a BiLSTM neural network, leveraging its powerful bidirectional long-range temporal dependency modeling capabilities to accurately predict the movement trends of dynamic obstacles such as swaying branches

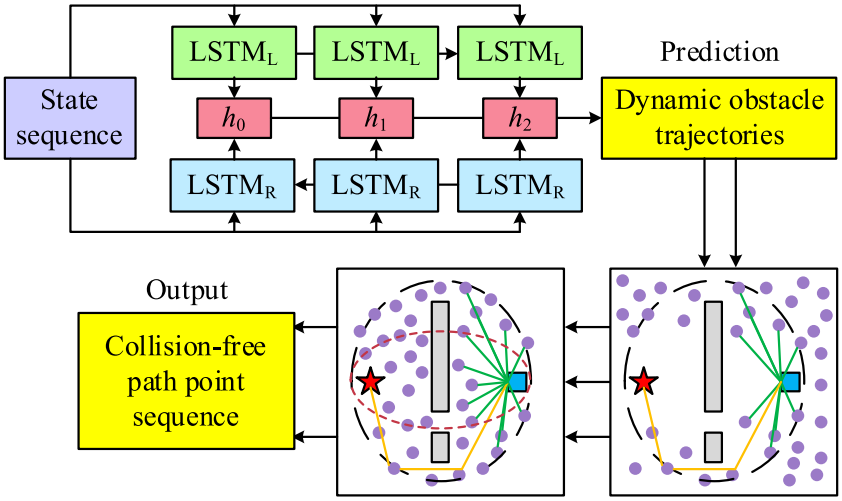

FIGURE 6
Schematic illustration of the RAOA module architecture.

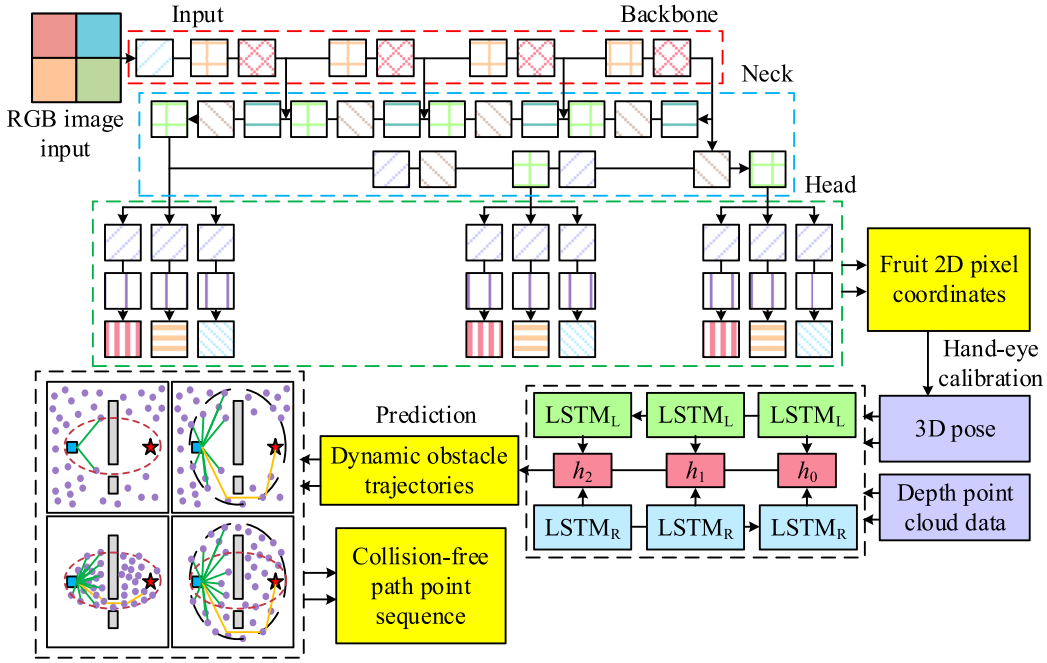

FIGURE 7
Schematic illustration of the YOLOv8-B* structure.

and leaves. This predictive information is incorporated into the BIT* search process from the beginning, which significantly enhances the planning system's forward-looking decision-making capabilities and the robustness of dynamic obstacle avoidance. Ultimately, this ensures the robotic arms generates collision-free trajectories that are safe, smooth, and actively adapt to environmental changes. In summary, this research integrates the harvesting target positioning module with the RAOA module to construct the YOLOv8-B* harvesting target positioning and RAOA algorithm. Its overall structure is illustrated in Figure 7.

In Figure 7, the algorithm first employs the ST BN within the harvesting target positioning module to extract multi-scale global features, enhancing the model's perception of occluded targets and complex backgrounds. The YOLOv8 framework then utilizes its FPN to achieve MSFF. An uncoupled detection head simultaneously performs fruit classification and precise localization, ultimately outputting the fruit's exact pixel coordinates. Subsequently, hand-eye calibration converts the 2D coordinates into a 3D pose within the robot's base coordinate system. This pose, along with depth point cloud data, is input into the RAOA module. Within this module, a

BiLSTM network predicts the motion trajectories of dynamic obstacles. The BIT* algorithm performs real-time, collision-free path planning based on environmental geometry and dynamic prediction results. It ultimately generates an optimal sequence of motion trajectories for the joint space of the robotic arms.

Additionally, in practical deployment, the visual system adopted in this research employs an "eye-on-hand" configuration, where the camera is fixed outside the robot's workspace. This setup stabilizes the camera's field of view during robotic arm movements. This enables continuous observation of the relationships between the robotic arm, the target fruit, and dynamic obstacles. It provides the BIT* planner with stable, global environmental perception input. This setup avoids the severe perspective shifts and occlusion issues inherent in "eye-on-hand" configurations caused by robotic arm motion. It simplifies the complexity of hand-eye calibration and coordinate transformation, thereby enhancing the robustness and real-time performance of the entire vision servo system.

3 Results and analysis

Testing is done in two dimensions: target localization and obstacle avoidance planning, to confirm the efficacy of the suggested YOLOv8-B* algorithm in intricate agricultural settings. The target localization dimension evaluates fruit RA and localization deviation by constructing test sets with varying occlusions and lighting conditions. The obstacle avoidance planning dimension analyzes path planning efficiency by generating dynamic and static obstacles in typical orchard scenarios. The algorithm's efficacy is comprehensively validated through comparative experiments. The testing conducts systematic testing on a mobile robotic platform equipped with a six-degree-of-freedom robotic arm. Using peach trees and their fruits as representative subjects, algorithm validation is performed specifically for their characteristics of dense growth and susceptibility to obstruction by branches and foliage. Subsequent simulations and performance analyses are all based on this specific crop scenario.

To ensure the validity of statistical inference, the study rigorously selects appropriate statistical methods based on data characteristics. Performance metrics for the target localization experiment are calculated using a large-scale independent test set. To account for environmental uncertainty, metrics for the obstacle avoidance planning experiment are obtained through independent, repeated runs across 30 randomly generated dynamic scenarios. For all intergroup comparisons of continuous performance metrics, this study employs independent samples t-tests to assess the significance of mean differences. Benefiting from ample samples and experimental repetitions, the sample mean distributions of performance metrics satisfies the conditions of the central limit theorem, meeting the requirements for parametric testing. All significance results (e.g., $p < 0.05$, $p < 0.01$) are based on this test, indicating that improvements in algorithm performance are statistically significant.

3.1 Target positioning performance testing

In target localization performance testing, the study leverages the PyTorch deep learning framework to implement the YOLOv8-B* architecture. OpenCV is utilized for image preprocessing and result visualization, with the Ultralytics YOLOv8 open-source code repository serving as the foundation for algorithm development. Python 3.8 automates testing frameworks simulated various typical agricultural scenarios, including multi-object occlusion, sudden lighting changes, and foliage interference. This approach supports configurable dynamic environmental parameters and real-time system stress testing. Parameter settings align with those described in the research methodology section. The study employs the PhenoBench dataset as both the test and training sets (stratified randomly split 2:8). This dataset comprises over 100,000 high-resolution aerial images of farmland captured by drones, providing pixel-level annotated crop semantic segmentation masks and annotations for more than 500,000 crop leaf instances. The PhenoBench dataset closely mirrors the visual challenges encountered in close-range harvesting scenarios by encompassing dense crop arrangements, complex foliage occlusions, and variable lighting conditions. Its large-scale, high-quality pixel-level annotations enable models to learn more generalizable feature representations, thereby enhancing robustness in both structured and unstructured orchard environments. Consequently, selecting this dataset for algorithm validation is both reasonable and representative (Weyler et al., 2024).

Additionally, the study compares methods from references (Miao et al., 2023; Gong et al., 2022; Hu et al., 2022) with YOLOv8-B*, specifically YOLOv5 and traditional image processing fusion algorithm (YOLOv5T), multisource image-fused mask R-CNN (MMRC), and YOLOv3 and mask R-CNN integrated dual-network framework (YOLOv3MR). These methods represent state-of-the-art approaches from 2022 to 2024, encompassing technical paradigms such as traditional and deep learning fusion, multi-source information perception, and dual-network collaborative optimization. They provide a comprehensive validation of YOLOv8-B*'s object localization performance. To validate the performance of the algorithm in complex, unstructured field environments, as described in the background section, field images of peach trees exhibiting typical occlusions, uneven lighting, and foliage interference are selected for testing. The target localization performance of different methods is visually compared, with results shown in Figure 8.

In Figure 8, YOLOv5T's feature extraction capability is constrained by the simple fusion of traditional image processing with YOLOv5, resulting in the detection of only 17 fruits (recall rate of 77.3%). This highlights the limitations of local modeling mechanisms in complex environments. MMRC detects 18 fruits (81.8%) by relying on multi-source image registration strategies, but its geometric reconstruction process suffers from cumulative errors. Although YOLOv3MR receives 19 detections (86.4%) through dual network integration with YOLOv3 and Mask R-CNN, it fails to resolve issues of insufficient feature alignment and sensitivity to occlusion. Additionally, YOLOv8-B* significantly enhances spatial perception of partially occluded fruits through ST's W-MSA/SW-MSA, achieving 21 detections (95.5%) to lead the evaluation. The W-MSA/SW-MSA mechanism allows the model to infer and fill in visual details in areas blocked by foliage. This is done by creating connections between non-local windows. This allows the model to use contextual information from unobscured parts of the fruit. This

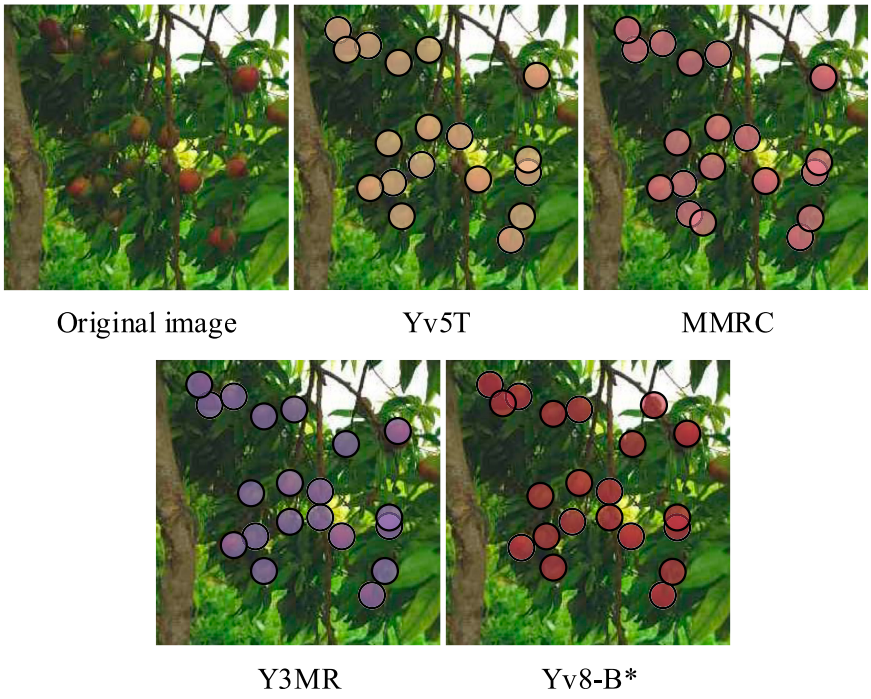

FIGURE 8
Visual validation of the model's object localization performance.

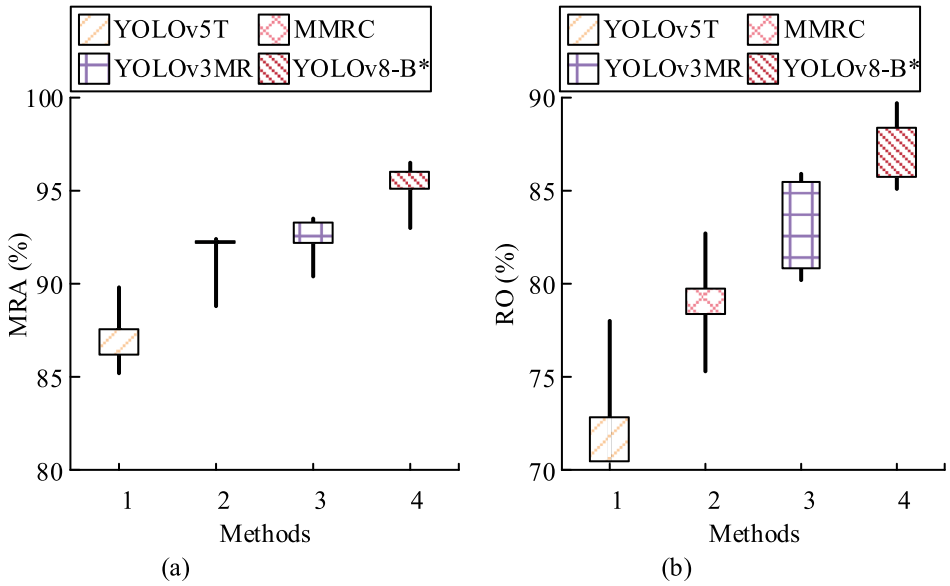

FIGURE 9
Validation of target localization accuracy and robustness. **(a)** MRA difference **(b)** RO difference.

directly validates its effective handling of unstructured challenges, such as "branch occlusion" and "scale variation," as defined in the background. It demonstrates that the introduced global attention mechanism significantly improves robustness of visual perception in complex, real-world environments. Subsequently, to quantitatively assess model accuracy and robustness, the study compares RA and recall under occlusion (RO) across different methods. The former represents the proportion of correctly identified fruits compared to the total number of fruits. The latter indicates the proportion of successfully detected fruits among all obscured fruits under occlusion conditions. The results are shown in Figure 9.

In Figure 9a, YOLOv8-B* achieves a significantly higher RA range of 93.0%–96.5% compared to the baseline model ($p < 0.001$). By incorporating the ST module to enhance MSFF and global

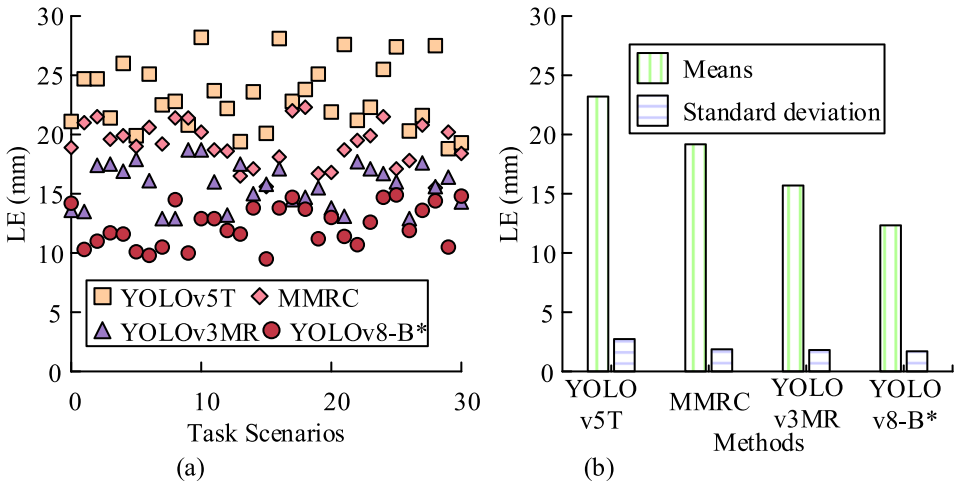

FIGURE 10
Validation of the model's localization accuracy. **(a)** LE difference **(b)** Means and standard deviation.

context modeling, it effectively improves fruit RA detection under complex occlusions. This improvement stems from the Swin Transformer's ability to surpass the local receptive field limitations of traditional CNNs by incorporating discriminative features throughout the entire image. This makes it more robust against inter-class confusion caused by uneven lighting or similar colors. YOLOv5T, relying on traditional image processing and simple YOLOv5 fusion, is limited in feature extraction capability, achieving an RA range of only 85.2%–89.8%. MMRC partially improves perception through its multi-source image fusion strategy, attaining an RA of 88.8%–92.4%. In Figure 9b, YOLOv8-B* also demonstrates a significant lead in RO ranges of 85.1%–89.7% under occlusion scenarios ($p < 0.001$). This advantage stems from the ST's sliding window mechanism, which enhances feature retention and spatial reasoning capabilities for partially occluded objects. Specifically, SW-MSA enables cross-window information exchange through window shifting, allowing the model to "borrow" features from adjacent visible regions to enhance the representation of the occluded fruit body. Although YOLOv3MR achieves relatively high recall rates (80.2%–85.9%) by integrating YOLOv3 and Mask R-CNN, it does not fundamentally resolve the issue of feature loss caused by occlusion. MMRC relies on multi-source registration and geometric reconstruction, yielding RO values of 75.3%–82.7%. The localization error (LE) of the different methods is then compared to evaluate the positioning accuracy of the models. LE is defined as the Euclidean distance between the predicted fruit center and the ground-truth center, as shown in Figure 10.

In Figures 10a,b, YOLOv8-B* exhibits an average LE of 12.33 mm with the lowest standard deviation (1.70 mm), demonstrating significantly superior performance compared to YOLOv5T (23.21 mm, $p < 0.001$), MMRC (19.18 mm, $p < 0.01$), and YOLOv3MR (15.70 mm, $p < 0.05$). The main source of this benefit is the W-MSA and SW-MSA processes of the ST, which improve its capacity to represent global spatial relationships. This mechanism enables BOB regression to anchor more precisely to the visible portion of the fruit and its geometric center, reducing drift errors caused by misleading local features. In typical scenarios, YOLOv8-B* achieves an optimal value of 9.8 mm in Scenario 6, where its sliding window attention effectively captures the geometric features of occluded fruits. The model can more accurately infer the complete contours and center positions of partially obscured fruits through global context, thereby achieving millimeter-level positioning accuracy. YOLOv3MR achieves 13.5 mm in Scenario 1 but overlaps with YOLOv8-B*'s 12.9 mm performance in Scenario 10, revealing limitations in feature alignment during dual-network integration. MMRC's minimum value of 15.5 mm in Scenario 28 remains higher than YOLOv8-B* in most scenarios, indicating that multi-source image registration fails to resolve cumulative error issues. YOLOv5T exhibits a maximum error of 28.2 mm in Scenario 10, highlighting the instability of traditional frameworks under dynamic lighting conditions. To evaluate the model's object localization efficiency and real-time performance, this study compares the processing frame rate (PFR) and inference time (IT) across different methods, as displayed in Table 1.

In Table 1, YOLOv8-B* achieves the optimal performance-speed balance with a frame rate of 32.7 fps and a latency of 32.6 ms. Its IT is significantly lower than MMRC (55.1 ms, $p < 0.001$) and YOLOv3MR (35.0 ms, $p < 0.01$), attributed to YOLOv8's C2F module and decoupled detection head effectively mitigating the computational overhead of ST. Although Swin Transformer introduces global computations, its windowed attention design effectively complements YOLOv8's efficient feature extraction pipeline and keeps computational complexity within acceptable limits. Although YOLOv5T achieves the highest frame rate of 46.1 fps in Scenario 20, this comes at the expense of localization accuracy. MMRC exhibits a worst latency of 59.1 m in Scenario 30, revealing inherent bottlenecks in multi-source fusion. YOLOv3MR achieves the best IT of 31.7 ms in Scenario 25, overlapping with YOLOv8-B* performance, yet its average frame rate of 27.1 fps remains insufficient. YOLOv8-B* simultaneously achieves 34.9 fps and 28.3 ms latency in Scenario 20, validating the synergistic advantages of global modeling and lightweight design.

TABLE 1 Validation of the model's object localization efficiency and real-time performance.

Task scenarios	PFR (fps)				IT (ms)			
	YOLOv5T	MMRC	YOLOv3MR	YOLOv8-B*	YOLOv5T	MMRC	YOLOv3MR	YOLOv8-B*
5	46.3	22.3	30.1	32.0	25.3	62.4	35.3	32.3
10	43.7	22.4	28.1	31.3	20.8	57.6	39.3	34.9
15	45.1	15.4	26.9	31.1	24.3	46.1	38.8	34.3
20	46.1	19.3	25.4	34.9	20.9	52.9	33.8	28.3
25	44.0	20.0	26.9	35.7	19.3	52.4	31.7	34.4
30	40.3	18.5	25.3	31.0	22.1	59.1	30.8	31.6
Means	44.3	19.7	27.1	32.7	22.1	55.1	35.0	32.6
Standard deviation	2.0	2.4	1.6	1.9	2.1	5.3	3.2	2.3

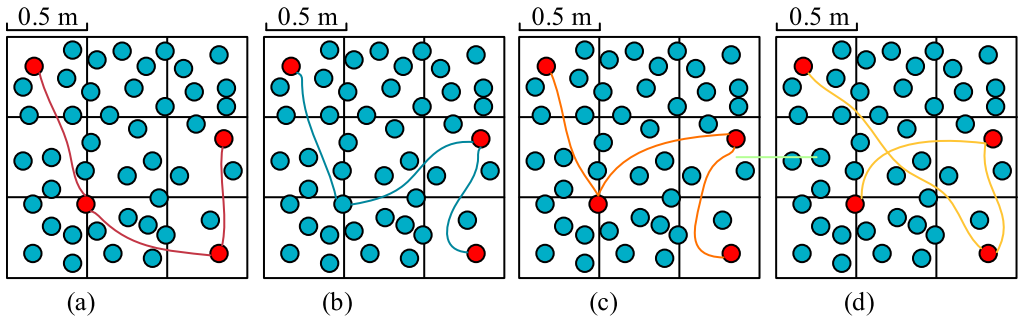

FIGURE 11
Efficiency verification of harvesting path planning. **(a)** Yv8-B* **(b)** HDRRT **(c)** HOPP **(d)** AVVP.

3.2 Obstacle avoidance performance verification of robotic arms

For RAOA performance validation, the study constructs a multi-scenario integrated testing environment within the Gazebo simulation platform, featuring dense orchards, crop row aisles, and mobile obstacles. Robot control and algorithm deployment are implemented via ROS. Continuous multi-source data streams are captured at high precision: point cloud from depth cameras (30Hz), LiDAR scans (40Hz), and robotic arm joint torques (1kHz sampling). These streams encompasses typical stress events marked by dynamic foliage interference, sudden obstacle intrusions, and multi-target harvesting path conflicts. The robot and algorithm implementation architecture aligns with the target positioning performance testing. Furthermore, the study compares YOLOv8-B* with methods from (Zhang et al., 2024; Liu, 2022; Yi et al., 2024): HDRRT, HOPP, and active vision-based view planner (AVVP). These advanced methods from 2022–2024 encompass dynamic sampling path planning, hierarchical optimization decision-making, and active perception planning, comprehensively validating YOLOv8-B*'s RAOA capabilities. The study first selects four ripe fruits as targets within a 1.5 m^3 space. Different methods are employed to control the robotic arms for fruit picking. By comparing the picking paths generated by each method, their planning efficiency is intuitively evaluated, as shown in Figure 11.

In Figure 11a, the YOLOv8-B* algorithm effectively avoids obstacles and generates a globally optimal path through its improved heuristic search structure and dynamic weight adjustment mechanism, achieving a minimum distance of 2.27 m, significantly outperforming the comparison model. The BiLSTM's dynamic obstacle prediction prior enables the BIT* algorithm to proactively avoid areas where obstacles may appear in the future during heuristic pruning. This directs the search toward safer, more direct pathways and prevents path detours caused by temporary obstacle avoidance. In Figure 11b, HDRRT (2.39 m) enhances exploration efficiency through random tree expansion but remains inferior to YOLOv8-B*'s structured search strategy. This fully demonstrates the core influence of algorithmic architecture on path planning performance in complex environments. In Figure 11c, HOPP (2.46 m) relies on a traditional rule base, resulting in numerous sharp angles in the path and generating redundant acceleration/deceleration phases during RA motion. In Figure 11d, AVVP (3.06 m) integrates visual perception but fails to prioritize targets, resulting in the longest planned path. The YOLOv8-B* model has shorter global paths, which directly reduces the overall exposure risk and cumulative collision probability for robotic arms navigating

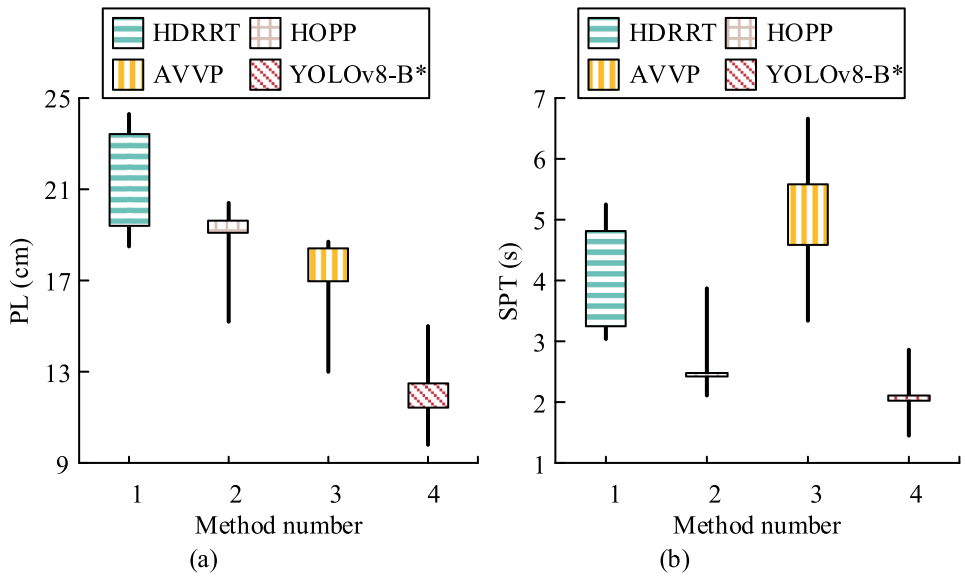

FIGURE 12
Validation of the model's comprehensive efficiency in obstacle avoidance planning. **(a)** PL difference **(b)** SPT difference.

through dense obstacles. This extends the fault-free operation time of robotic arms in unstructured environments with tangled branches, thereby enhancing picking efficiency. Subsequently, the study compares the single obstacle avoidance PL and single planning time (SPT) across different methods to evaluate the overall efficiency of obstacle avoidance planning, as shown in Figure 12.

In Figure 12a, the PL range of YOLOv8-B* (9.8 cm^{-1} to 15.0 cm^{-1}) significantly exceeds that of HDRRT (18.5 cm^{-1} to 24.3 cm^{-1}, $p < 0.001$) and HOPP's 15.2 cm–20.4 cm ($p < 0.01$), and AVVP's 13.0 cm–18.7 cm ($p < 0.05$). Its BIT* algorithm generates compact paths through heuristic pruning and BiLSTM dynamic obstacle prediction. The prediction error of BiLSTM primarily influences the conservatism of pruning: high-confidence predictions enable BIT* to prune future safe regions more aggressively, directly planning shorter paths. Whereas with low confidence, the algorithm retains a wider safety margin, slightly increasing PL to ensure robustness. AVVP achieves an optimal single point of 13.0 cm, but increases to 18.4 cm in Scenario 30, indicating instability in its view iteration mechanism. HDRRT's random sampling results in the highest path redundancy, reaching 23.4 cm in Scenario 0. In environments with dense obstacles, the more compact path of YOLOv8-B* enables the robotic arm's end-effector to navigate narrow spaces with smaller movements and closer adherence to the intended trajectory. This significantly reduces unexpected scrapes or collisions caused by path redundancy. In Figure 12b, YOLOv8-B* also significantly outperforms competitors ($p < 0.001$) with an SPT range of 1.45 s–2.86 s, where its BiLSTM-augmented architecture compresses the search space through spatio-temporal modeling. By preemptively excluding a large number of invalid sampling regions containing future collision risks, BiLSTM's predictions reduce the number of vertices and edges that BIT* needs to evaluate. This substantially lowers the computational overhead per iteration. Although HOPP achieves 2.42 s in Scenario 0, its peak value of 3.87 s overlaps with YOLOv8-B*, revealing the computational burden of hierarchical optimization. HDRRT is the least efficient in random sampling, taking 3.04 s–5.25 s YOLOv8-B*'s extremely short planning time enables the system to perform high-frequency replanning. The robotic arm can adjust its trajectory nearly in real time when encountering sudden dynamic obstacles, such as swaying branches, or target position updates. This ability is a prerequisite for achieving reliable dynamic obstacle avoidance. The study also compares the energy consumption of manipulator (ECM) across different methods to evaluate model efficiency, as shown in Figure 13.

In Figures 13a,b, the average ECM of YOLOv8-B* is 124.58 J, significantly lower than that of HDRRT (228.35 J, $p < 0.001$), HOPP (186.68 J, $p < 0.01$), and AVVP (158.52 J, $p < 0.05$). This advantage stems from the BIT* algorithm generating optimal paths to minimize redundant motion, combined with BiLSTM dynamic prediction to avoid abrupt stops and re-planning. BiLSTM's precise predictions enable the robotic arm to smoothly navigate around dynamic obstacles in advance, avoiding the abrupt braking and re-acceleration processes common in traditional reactive obstacle avoidance. This represents one of the key mechanisms for reducing energy consumption. YOLOv8-B* achieves the lowest energy consumption of 95 J in Scenario 1, where its heuristic pruning and spatio-temporal prediction effectively optimize trajectories. The shorter PL combined with forward-looking speed planning enables the joint motor to operate within its high-efficiency range most of the time. This reduces the additional torque required to overcome inertia and minimizes energy loss. Although AVVP achieves 128 J in Scenario 20, overlapping with YOLOv8-B*'s 142 J performance in Scenario 1, its view iteration mechanism causes additional kinetic energy consumption to rise to 143 J in Scenario 8. HOPP achieves low energy consumption of 153 J in Scenario 23, but does not consider joint torque continuity, resulting in energy consumption as high as

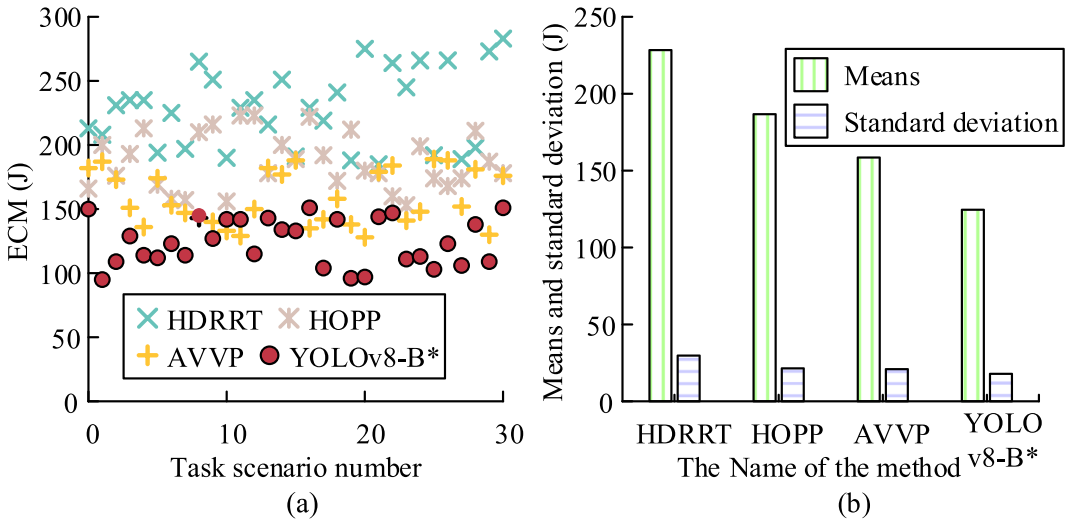

FIGURE 13
Validation of the model's economic efficiency. **(a)** ECM difference **(b)** Means and standard deviation.

TABLE 2 Validation of the model's potential for promotion and application.

Task scenarios	CL (GFlops)				MC (MB)			
	HDRRT	HOPP	AVVP	YOLOv8-B*	HDRRT	HOPP	AVVP	YOLOv8-B*
5	20.3	25.5	72.2	22.7	171	315	212	187
10	23.5	23.9	49.4	21.3	160	224	221	205
15	25.1	32.8	54.5	21.9	167	245	253	207
20	24.4	23.6	64.4	24.5	160	247	186	171
25	23.2	28.2	70.1	24.2	126	232	231	190
30	19.9	22.4	72.6	21.3	167	230	259	160
Means	22.7	26.1	63.9	22.7	159	249	227	187
Standard deviation	2.0	3.5	9.0	1.3	15	31	25	17

223 J in Scenarios 11 and 12. HDRRT's random sampling leads to path redundancy, causing the highest energy consumption of 283 J in cenario 30. YOLOv8-B* has a lower kinetic energy consumption that directly reflects the smoothness and efficiency of the robotic arm's trajectory. This avoids abrupt acceleration and deceleration caused by emergency obstacle avoidance or suboptimal path planning. This stable motion further reduces the risk of contact collisions between the end-effector and fruits or branches due to vibration or inertia. Subsequent studies compares the single-task computational load (CL) and memory consumption (MC) of different methods to evaluate the models' potential for broader application, as displayed in Table 2.

Table 2 shows that YOLOv8-B* achieves optimal resource efficiency with a CL of 22.7 GFlops and MC of 187 MB. Its CL is significantly lower than AVVP's 63.9 GFlops ($p < 0.001$). Its MC is significantly lower than HOPP's 249 MB ($p < 0.01$). This advantage stems from the synergistic optimization of BIT* heuristic search and BiLSTM prediction, which reduces computational iterations, while the ST window attention mechanism minimizes memory usage through parameter sharing. As a lightweight temporal module, BiLSTM replaces the dynamic environment modeling traditionally achieved through extensive sampling and collision detection, fundamentally reducing the computational complexity of the planner. Among these, YOLOv8-B* achieves the lowest CL of 21.3 GFlops in Scenarios 10 and 30. While HDRRT reaches 19.9 GFlops in Scenario 30, overlapping with YOLOv8-B*'s performance, this comes at the cost of reduced path quality. AVVP, benefiting from multi-source perception and view iteration optimization, achieves peak loads of 72.2 GFlops and 72.6 GFlops in Scenarios 5 and 30, respectively. HOPP's hierarchical structure requires pre-storing global path information, leading to MC of 315 MB in Scenario 5, significantly exceeding YOLOv8-B*'s 187 MB performance in the same scenario. The reduced computational and memory demands of YOLOv8-B* ensure stable operation of the algorithm on onboard computing units. This frees ample resources for processing high-frequency visual feedback and continuous obstacle avoidance planning. This safeguards the real-time performance and

TABLE 3 Verification of absorption/replacement for basic modules.

Settings	YOLOv8	Swin transformer	BIT*	BiLSTM	RA (%)	PL (cm)	CL (GFlops)
Full (Yv8-B*)	✓	✓	✓	✓	96.5	12.4	22.7
A1 (w/o Swin Transformer)	✓	×	✓	✓	90.1	13.1	18.4
A2 (w/o BiLSTM)	✓	✓	✓	×	95.8	18.7	21.8
A3 (w/o SwinT and BiLSTM)	✓	×	✓	×	89.6	19.2	17.2
YOLOv5	—	✓	✓	✓	93	13.8	24.5
YOLOv10	—	✓	✓	✓	95.3	12.9	25.7
ECA	✓	—	✓	✓	92.7	14.5	19.3
SE attention	✓	—	✓	✓	93.4	14.2	19.5
MobileViT	✓	—	✓	✓	94.6	13.6	20.3
ConvFormer	✓	—	✓	✓	95	13.3	23
RRT*	✓	✓	—	✓	96	15.9	19.7
RRTX	✓	✓	—	✓	96.1	16.4	26.4
LSTM	✓	✓	✓	—	95.9	13.8	22
TCN	✓	✓	✓	—	95.8	14.1	23.4

reliability of the entire perception-planning loop in complex scenarios, forming the foundational system for achieving sustained safe obstacle avoidance. The deep collaboration between modules significantly reduces the system's overall resource consumption compared to the simple sum of individual modules, demonstrating the superiority of the architectural design.

To validate the necessity of the base module design, the study performs ablation tests on the base module. Visual backbones are replaced with YOLOv5, YOLOv10, efficient channel attention (ECA), squeeze-and-excitation (SE) attention, mobile vision Transformer (MobileViT), and convolution-enhanced Transformer (ConvFormer). Planners are replaced with rapidly-exploring random tree star (RRT*) and RRT with eXact anytime optimization (RRTX). LSTM and temporal convolutional network (TCN) are used to replace the temporal predictor (BiLSTM). The results are shown in Table 3.

As shown in Table 3, the complete model (Yv8-B*) achieves optimal overall performance in terms of RA (96.5%), PL (12.4 cm), and CL (22.7 GFlops). Ablation experiments reveal that removing Swin Transformer (A1) significantly degrades RA ($p < 0.001$) and increases PL. This demonstrates its critical role in enhancing RA and generating compact paths through global modeling. Removing BiLSTM (A2) substantially increases PL to 18.7 cm ($p < 0.001$), validating dynamic prediction's core contribution to path planning efficiency. In module replacements, MobileViT and ConvFormer both yields lower RA (94.6%, 95.0%) than the full model with longer paths. ECA and SE demonstrates weaker accuracy and path performance. Replacing BIT with RRT and RRTX increases PL to 15.9 cm and 16.4 cm respectively ($p < 0.01$), with RRTX achieving higher CL. Substituting BiLSTM with LSTM and TCN also results in longer PLs. Experiments quantitatively confirm that the selected Swin Transformer and BiLSTM modules achieve the optimal balance among RA, path planning efficiency, and CL.

4 Discussion and conclusion

To address the challenges of inaccurate fruit localization and inefficient dynamic obstacle avoidance in complex agricultural environments, this study proposed the YOLOv8-B* fusion algorithm based on an enhanced YOLOv8 and BIT*. By incorporating the ST module to enhance MSFF and global context modeling, and integrating a BiLSTM network to endow the BIT* algorithm with dynamic obstacle prediction capabilities, an integrated perception-decision-control harvesting robot system was constructed. Experiments demonstrated that YOLOv8-B* achieved RA of 93.0%–96.5%, RO of 85.1%–89.7%, and a mean LE of 12.33 mm in the target localization dimension. Compared to the optimal reference model, it improved accuracy by 3.5% and reduced LE by 21.5%. Moreover, in the obstacle avoidance planning dimension, it achieved a PL of 9.8 cm–5.0 cm and a planning time of 1.45 s–2.86 s, reducing PL by 17.8% and improving planning efficiency by 38.2% compared to the optimal comparison model. In actual deployment, the ECM is reduced to 124.58 J, with single-task CL and MC at 22.7 GFlops and 187 MB respectively. Compared to mainstream methods, resource consumption is reduced by an average of 42.3% and 24.9%, validating the algorithm's comprehensive advantages in accuracy, efficiency, energy consumption, and resource economy.

The architectural innovation of YOLOv8-B* lies in its dual-module coordination mechanism: The visual perception module based on ST overcomes the local perception limitations of traditional CNNs through a sliding window attention mechanism, significantly enhancing the representation capability of occluded object features. The BIT*-enhanced planning module addresses the response lag issue for dynamic obstacles by combining spatio-temporal context prediction with heuristic search. The two components form a closed-loop system through hand-eye calibration, enabling seamless transition from fruit recognition to

path planning. This research has achieved a significant reduction in the CL and memory footprint of single-task operations, compared to the typical computing power and memory capacity of mainstream embedded AI computing platforms like Jetson Orin NX. This indicates that the Yv8-B* algorithm architecture possesses the potential for direct porting to such platforms and achieving real-time operation. However, the research has several limitations. First, the ST module has high computational demands, so it needs to be optimized and validated further for deployment on embedded devices. BiLSTM's dynamic prediction relies on historical data quality, potentially leading to error accumulation under extreme occlusion scenarios. Future work will address these challenges through the design of a lightweight hybrid attention mechanism that balances computational efficiency and model performance. Additionally, the development of model lightweighting and operator optimization deployment strategies tailored for edge computing platforms like Jetson will ensure the stable, real-time operation of the algorithm in actual onboard robot systems. Additionally, a multi-sensor fusion dynamic obstacle trajectory compensation algorithm will be developed to enhance system robustness in adverse environments.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

Author contributions

YX: Writing – original draft, Formal Analysis, Methodology, Investigation, Writing – review and editing, Data curation, Conceptualization.

Funding

The author(s) declared that financial support was received for this work and/or its publication. The research is supported by: Project Level: Key Provincial Teaching Research Project of Higher Education Institutions in Anhui Province; Research on Strategies and Paths to Improve Teachers' Informatization Ability under Educational Digital Transformation; (2022jyxm946).

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that generative AI was not used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Ayyad, S. M., Sallam, N. M., Gamel, S. A., and Ali, Z. H. (2025). Particle swarm optimization with YOLOv8 for improved detection performance of tomato plants. *J. Big Data* 12 (1), 152–153. doi:10.1186/s40537-025-01206-6

Bai, Y., Mao, S., Zhou, J., and Zhang, B. (2023). Clustered tomato detection and picking point location using machine learning-aided image analysis for automatic robotic harvesting. *Precis. Agric.* 24 (2), 727–743. doi:10.1007/s11119-022-09972-6

Droukas, L., Doulgeri, Z., Tsakiridis, N. L., Triantafyllou, D., Kleitsiotis, I., Mariolis, I., et al. (2023). A survey of robotic harvesting systems and enabling technologies. *J. Intell. Robot. Syst.* 107 (2), 21–22. doi:10.1007/s10846-022-01793-z

Gao, X., and Zhang, Y. (2023). Detection of fruit using YOLOv8-based single stage detectors. *Int. J. Adv. Comput. Sci. Appl.* 14 (12), 83–84. doi:10.14569/IJACSA.2023.0141208

Gao, Z. C., Huang, J. X., Chen, J. H., Shao, T. Y., Ni, H., and Cai, H. H. (2024). Deep transfer learning-based computer vision for real-time harvest period classification and impurity detection of porphyra haitnensis. *Aquac. Int.* 32 (4), 5171–5198. doi:10.1007/s10499-024-01422-6

Gong, L., Wang, W., Wang, T., and Liu, C. (2022). Robotic harvesting of the occluded fruits with a precise shape and position reconstruction approach. *J. Field Robot.* 39 (1), 69–84. doi:10.1002/rob.22041

Hu, H., Kaizu, Y., Zhang, H. D., Xu, Y. W., Imou, K. J., Li, M., et al. (2022). Recognition and localization of strawberries from 3D binocular cameras for a strawberry picking robot using coupled YOLO/mask R-CNN. *Int. J. Agric. Biol. Eng.* 15 (6), 175–179. doi:10.25165/j.ijabe.20221506.7306

Huynh, L. Q., Tran, L. V., Phan, P. N., Yu, Z., and Dao, S. V. (2023). Intermediary RRT*-PSO: a multi-directional hybrid fast convergence sampling-based path planning algorithm. *Comput. Mater. Contin.* 76 (2), 2281–2300. doi:10.32604/cmc.2023.034872

Johnson, J. J., Qureshi, A. H., and Yip, M. C. (2023). Learning sampling dictionaries for efficient and generalizable robot motion planning with transformers. *IEEE Robot. Autom. Lett.* 8 (12), 7946–7953. doi:10.1109/LRA.2023.3322087

Kumudham, R., Shakir, M., and Abishek B, E. (2024). Enhancing brix value prediction in strawberries using machine learning: a fusion of physiochemical and color-based features for improved sweetness assessment. *Malays. J. Comput. Sci.* 37 (2), 107–123. doi:10.22452/mjcs

Kyaw, P. T., Le, A. V., Veerajagadheswar, P., Elara, M. R., Thu, T. T., Nhan, N. H. K., et al. (2022). Energy-efficient path planning of reconfigurable robots in complex environments. *IEEE Trans. Robot.* 38 (4), 2481–2494. doi:10.1109/TRO.2022.3147408

Li, H., Huang, J., Gu, Z., He, D., Huang, J., and Wang, C. (2024). Positioning of mango picking point using an improved YOLOv8 architecture with object detection and instance segmentation. *Biosyst. Eng.* 247 (1), 202–220. doi:10.1016/j.biosystemseng.2024.09.015

Liu, D. W. (2022). Hierarchial optimal path planning (HOPP) for robotic apple harvesting. *Int. J. Health Sci. Res.* 4 (3), 6–12. doi:10.36838/v4i3.2

Liu, J., and Liu, Z. (2024). The vision-based target recognition, localization, and control for harvesting robots: a review. *Int. J. Precis. Eng. Manuf.* 25 (2), 409–428. doi:10.1007/s12541-023-00911-7

Ma, B., Hua, Z., Wen, Y., Deng, H., Zhao, Y., Pu, L., et al. (2024). Using an improved lightweight YOLOv8 model for real-time detection of multi-stage apple fruit in complex orchard environments. *Artif. Intell. Agric.* 11 (1), 70–82. doi:10.1016/j.aiia.2024.02.001

Miao, Z., Yu, X., Li, N., Zhang, Z., He, C., Li, Z., et al. (2023). Efficient tomato harvesting robot based on image processing and deep learning. *Precis. Agric.* 24 (1), 254–287. doi:10.1007/s11119-022-09944-w

Nenavath, D. N., and Perumal, B. (2024). On-tree mango fruit count using live video-split image dataset to predict better yield at pre-harvesting stage. *Int. J. Elect. Comput. Eng. Syst.* 15 (9), 771–782. doi:10.32985/ijeces.15.9.5

Pal, S., Roy, A., Shivakumara, P., and Pal, U. (2023). Adapting a swin transformer for license plate number and text detection in drone images. *Artif. Intell. Appl.* 1 (3), 145–154. doi:10.47852/bonviewAIA3202549

Panduranga, K. M., and Ranganathasharma, R. H. (2024). Sustainability insights on learning-based approaches in precision agriculture in internet-of-things. *Int. J. Elect. Comput. Eng.* 14 (3), 3495–3511. doi:10.11591/IJECE.V14I3.PP3495-3511

Tamizi, M. G., Honari, H., Nozdryn-Plotnicki, A., and Najjaran, H. (2024). End-to-end deep learning-based framework for path planning and collision checking: bin-picking application. *Robotica* 42 (4), 1094–1112. doi:10.1017/S0263574724000109

Tang, J., Yu, Z., and Shao, C. (2025). TransSSA: invariant cue perceptual feature focused learning for dynamic fruit target detection. *Comput. Mater. Contin.* 83 (2), 2829–2850. doi:10.32604/cmc.2025.063287

Wang, C., Yang, G., Huang, Y., Liu, Y., and Zhang, Y. (2023). A transformer-based mask R-CNN for tomato detection and segmentation. *J. Intell. Fuzzy Syst.* 44 (5), 8585–8595. doi:10.3233/JIFS-222954

Weyler, J., Magistri, F., Marks, E., Chong, Y. L., Sodano, M., Roggiolani, G., et al. (2024). Phenobench: a large dataset and benchmarks for semantic image interpretation in the agricultural domain. *IEEE Trans. Pattern Anal. Mach. Intell.* 46 (12), 9583–9594. doi:10.1109/TPAMI.2024.3419548

Xu, J., and Li, W. (2025). Lightweight improvement algorithm for target detection of Pu'er tea harvesting robotic arm based on YOLOv8. *Int. J. Inf. Commun. Technol.* 26 (8), 1–18. doi:10.1504/IJICT.2025.145720

Xu, T., Zhou, H., Tan, S., Li, Z., Ju, X., and Peng, Y. (2021). Mechanical arm obstacle avoidance path planning based on improved artificial potential field method. *Ind. Robot.* 49 (2), 271–279. doi:10.1108/IR-06-2021-0120

Xu, J., He, Y., Tian, H., and Wei, Z. (2022). A random path sampling-based method for motion planning in many dimensions. *IEEE Trans. Instrum. Meas.* 73 (1), 1–8. doi:10.1109/TIM.2022.3212036

Yi, T., Zhang, D., Luo, L., and Luo, J. (2024). View planning for grape harvesting based on active vision strategy under occlusion. *IEEE Robot. Autom. Lett.* 9 (3), 2535–2542. doi:10.1109/LRA.2024.3357397

Yu, Y., An, X., Lin, J., Li, S., and Chen, Y. (2024). A vision system based on CNN-LSTM for robotic citrus sorting. *Inf. Process. Agric.* 11 (1), 14–25. doi:10.1016/j.inpa.2022.06.002

Zeeshan, S., and Aized, T. (2023). Performance analysis of path planning algorithms for fruit harvesting robot. *J. Biosyst. Eng.* 48 (2), 178–197. doi:10.1007/s42853-023-00184-y

Zhai, W., Xu, Z., Liu, J., Xiong, X., Pan, J., Chung, S. O., et al. (2024). Feature deformation network with multi-range feature enhancement for agricultural machinery operation mode identification. *Int. J. Agric. Biol. Eng.* 17 (4), 265–275. doi:10.25165/j.ijabe.20241704.8831

Zhang, B., Yin, C., Fu, Y., Xia, Y., and Fu, W. (2024). Harvest motion planning for mango picking robot based on improved RRT-connect. *Biosyst. Eng.* 248 (1), 177–189. doi:10.1016/j.biosystemseng.2024.10.008

Zhou, H., Wang, X., Au, W., Kang, H., and Chen, C. (2022). Intelligent robots for fruit harvesting: recent developments and future challenges. *Precis. Agric.* 23 (5), 1856–1907. doi:10.1007/s11119-022-09913-3