FASE FRONTIERS OF AGRICULTURAL SCIENCE AND ENGINEERING
Engineering Transactions of CAE
RESEARCH ARTICLE
2026 / Volume 13 / Issue 2 / 25653

Sweet potato grasping point recognition in complex backgrounds: a method based on salient object detection

Ranbing YANG[1,2], Ang ZHAO[1,2], Danyang LV[2,3], Yongfei PAN[2,3], Hongfei ZHU[3], Xinyu GUO[1], Jian ZHANG (✉)[1,2], Jianqi HOU[1]

1 College of Mechanical and Electrical Engineering, Hainan University, Haikou 570228, China.
2 Key Laboratory of Tropical Intelligent Agricultural Equipment, Ministry of Agriculture and Rural Affairs, Danzhou 570228, China.
3 College of Information and Communication Engineering, Hainan University, Haikou 570228, China.

KEYWORDS

Agricultural robots, edge detection, salient object detection, sweet potato

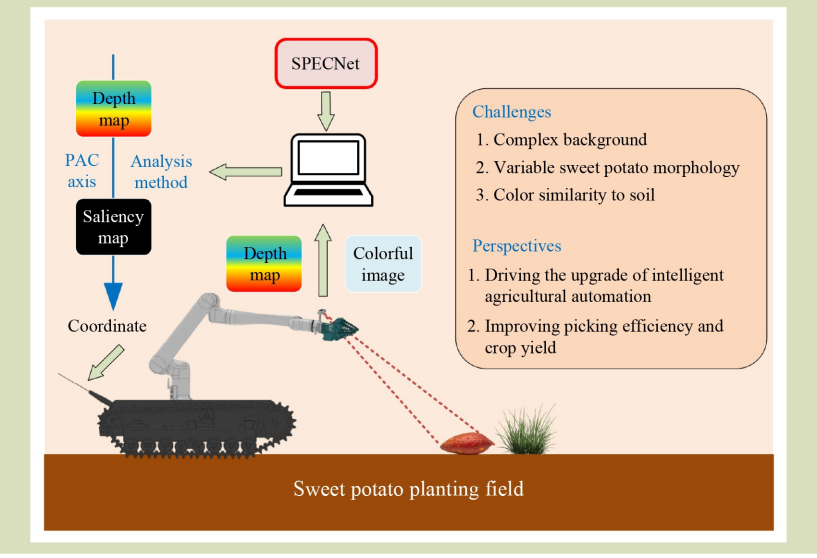
HIGHLIGHTS

- A novel SPECNet architecture integrating Dynamic Convolution (Dynamic Conv), Haar wavelet downsampling (HWD), and Edge-Enhanced SE attention (ESE) is proposed, significantly improving sweet potato contour extraction under complex backgrounds.
- In experiments, SPECNet achieved MAE = 0.102, *F*-measure = 0.970, *E*-measure = 0.984, and *S*-measure = 0.968 on the test set, outperforming six mainstream comparison models.
- Applied to sweet potato harvesting robot platform. SPECNet processes at 35 fps in real time and achieves a 78% grasp success rate, demonstrating stability and practicality in field conditions.
- Ablation studies confirm that the Dynamic Conv, HWD, and ESE modules each contribute substantial performance gains, enhancing feature representation, edge preservation, and detail refinement.

Received March 25, 2025;
Accepted July 4, 2025.

Correspondence: zhangjian_qau@163.com

GRAPHICAL ABSTRACT



ABSTRACT

To address challenges in crop grasping tasks for agricultural robots, specifically, poor crop background segmentation and limited adaptability in grasp point localization, this paper proposes a saliency guided segmentation approach. This method improves both object recognition and grasp point detection, thereby optimizing robot grasping performance and increasing success rates, even under complex environmental conditions. The proposed network uses a boundary aware detection strategy built on an encoder decoder architecture with an improvement module. First, standard convolutions are replaced by dynamic convolution to improve feature representation. Second, a Haar wavelet downsampling module is introduced to improve multi scale feature

extraction. Finally, the standard squeeze and excitation attention block is improved with edge enhancement, which is embedded at each decoding stage to emphasize boundary information. In benchmark tests, the proposed model achieved a mean absolute error of 10.9%, with *F*-, *E*- and *S*-measures of 97.0%, 98.4%, and 96.8%, respectively. When deployed on an agricultural robot platform, it achieved a 78.0% grasping success rate, processing images at 35 frames per second. These results demonstrate that the proposed network reliably identifies and localizes optimal grasp points under real world conditions.

1 Introduction

Sweet potato is an important agricultural crop being one of the main food crops worldwide. Its rich nutritional value makes it widely applicable in food processing, feed production and the pharmaceutical industry. The cultivation of sweet potato in Hainan Province of China has a history of over 300 years and has long been an important production area in the southern potato region of China[1]. The terrain of Hainan Province is predominantly mountainous in the central region, gradually transitioning to hills, terraces and plains, forming a ringlike geomorphological structure. This topographical feature presents challenges for agricultural mechanization, particularly in the planting and harvesting of crops such as sweet potatoes. Large machinery struggles to operate in these areas, so sweet potatoes are primarily hand harvested. Therefore, research on sweet potato picking robots is of significant importance for sweet potato harvesting in Hainan Province.

Existing methods in agricultural robotics often struggle to accurately segment crops from complex backgrounds during grasp point localization, leading to difficulties separating crops from the background and limited adaptability to varying environments[2–5]. Researchers have made significant progress in addressing these issues. For example, Wang et al.[6] used the K-means color clustering algorithm to extract apple target regions and performed contour smoothing, achieving an average localization error of 4.54°. Wang et al.[7] used an improved Canny operator for edge detection in foggy conditions, attaining an average precision of 0.615. Gao et al.[8] used deep learning for point cloud registration, proposing an edge-aware point cloud registration algorithm that improved the recall rate of key point registration in low-overlap point cloud scenarios (3DLoMatch) by about 5%. Despite these advancements, challenges remain, such as the difficulty of the K-means clustering algorithm to provide precise segmentation when target and background colors are similar (Fig. 1(b)), the potential for the Canny operator to confuse object and background edges (Fig. 1(c)) and the high costs associated with point cloud registration algorithms. Figure 1 provides a comparison of these methods.

In recent years, the rise of deep learning technologies has demonstrated excellent adaptability and robustness in addressing the above-mentioned challenges, and has been widely applied to the recognition of contours and pest detection in crops[9] such as tomatoes[10], rice[11], oranges[12], and maize[13], and pineapple[14,15]. Yu et al.[16] proposed the CASENet deep category aware semantic edge detection network, which significantly improved the performance of the original model through a depth unsupervised nested architecture. Li and Yu[17] used deep convolutional neural networks to extract multiscale features, demonstrating significant improvements in visual saliency models on the DUT-OMRON public data set. As a major branch of deep learning, salient object detection is an important direction in computer vision, aiming to separate the most salient objects from complex backgrounds, and through extensive training, it can accurately extract the contours of specified objects (Fig. 1(d)). Recently, fully convolutional networks[18] have been used for saliency object detection, significantly improving recognition performance on public data sets. Due to its characteristics, saliency object detection has broad application prospects for crop contour recognition in agricultural robots. He et al.[19] applied this method to improve the effectiveness and robustness of machine vision for recognizing mature mulberries in natural harvesting environments. Yuan et al.[20] applied this method to visual detection of tissue culture plantlet

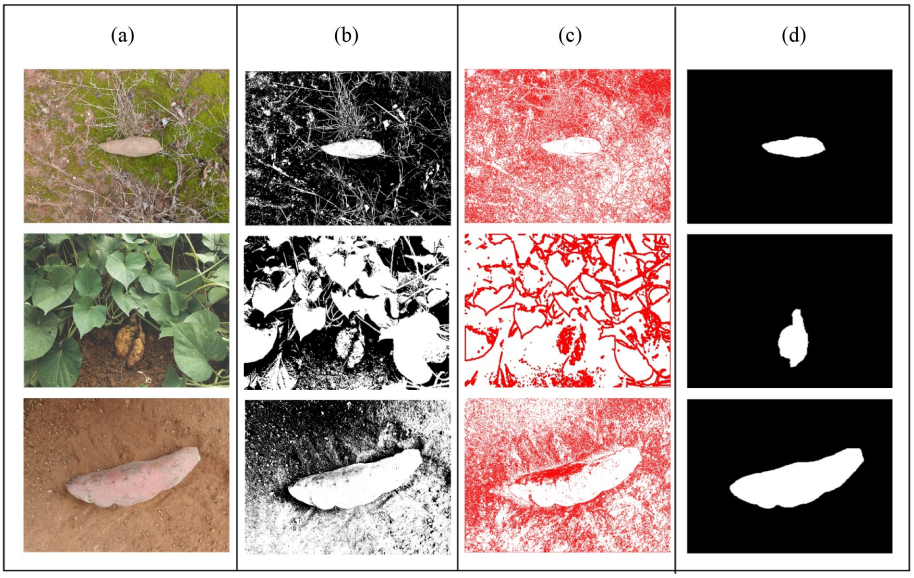


Fig. 1 Contour recognition based on three commonly-used algorithms: (a) unprocessed original image; (b) K-means.; (c) Canny; and (d) salient object detection.

picking points in the automated rapid breeding of orchids, significantly enhancing adaptability and efficiency. Although significant progress has been made in contour extraction, research on sweet potato contour detection in complex environments is still limited. Therefore, addressing issues such as background interference and the difficulty in determining the grasping point in sweet potato picking in complex environments, this work investigates sweet potato contour recognition and grasping point extraction under such conditions. A saliency detection method is proposed for accurately locating the grasping point, with an improved model to enhance the ability to extract sweet potato contours in complex environments, helping sweet potato picking robots complete the task. This approach aims to meet the needs of sweet potato picking tasks in complex field environments.

2 Materials and methods

To build a comprehensive data set, a total of 5423 sweet potato images were captured using an Intel RealSense D435i depth camera. The images were collected in October 2024 at the Sweet Potato Experimental Field (red soil) of the Intelligent Agricultural Machinery Equipment Research and Development Base, Hainan University, and in May 2025 at the Haitou Sweet Potato Provincial Modern Agriculture Industrial Park (sandy soil) in Haitou Town, Danzhou City, Hainan Province. Images were collected daily during different time periods. The collected images were organized, and 3500 images (1750 from each site) were randomly selected. These were partitioned into a training set (2500 images), a validation set (500 images) and a test set (500 images), ensuring a 1:1 ratio of sandy soil to red soil images in each subset. Data augmentation techniques (including flipping and blurring) were subsequently applied, expanding the data set to 14,000 images. Ultimately, the data set was split into 10,000 training images, 2000 validation images, and 2000 test images. During the data preparation phase, the open-source annotation tool Labelme was used to manually label ground truth masks for all images, ensuring accurate and reproducible annotations.

After organizing the data set, characteristics of harvested sweet potato images were determined (Fig. 2).

(1) Surface contamination: since sweet potatoes grow underground, their surfaces are often covered with soil or moist stains during harvesting, which affects the overall color and clarity of the images.

(2) Background clutter: during harvesting, the background may

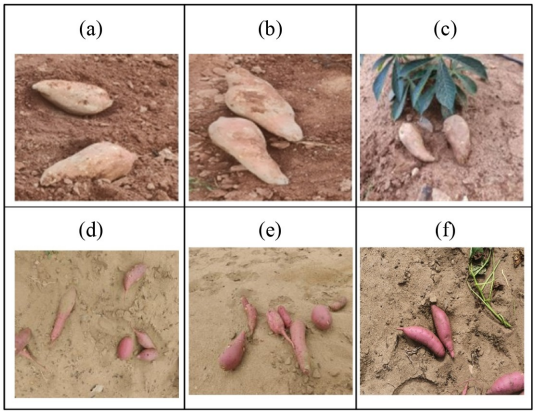


Fig. 2 Pictures of sweet potatoes with different characteristics: (a, d) soil cover; (b, e) edge overlap; and (c, f) environmental disturbance in two contexts, red soil (a, b, c) and sandy soil (d, e, f).

contain other non-target objects such as stones, leaves, soil or crop residues, and there may even be cases where the edges of sweet potatoes overlap.

3 Visual inspection algorithms

3.1 Sweet potato contour identification network structure

Due to the challenges in sweet potato picking, such as complex harvesting environments, cluttered backgrounds and soil color similarity, there is a need for enhanced feature extraction and edge recognition capabilities. BASNet[21] combines the segmentation capability of an encoder-decoder architecture with a predict-refine module. By integrating a triple hybrid loss function comprising BCE, SSIM and IoU, it effectively highlights salient regions and accurately captures intricate boundaries. Also, BASNet-based variants (e.g., U^2-Net[22] and Mobile BASNet[23]) are mature and available on code-sharing platforms, such as GitHub, providing abundant pretrained weights and implementation examples. This availability significantly reduces development complexity and ensures the stability and reproducibility of both training and deployment when customizing the network within this framework. Therefore, we used BASNet architecture as the foundation of our proposed model.

Dynamic convolution[24] generates variable convolutional kernels from input features, enabling the network to adaptively adjust its filters based on image content and thus capture diverse appearance characteristics. In the co-saliency detection task, the CADC model, built on dynamic convolution, efficiently localizes fine-grained targets by aggregating shared features across an image group and then applying dynamic convolution to each image. By applying this approach to sweet potato contour detection, the ability of the network to capture diverse shape features can be significantly improved. Therefore, in the proposed model, the convolutional layers in the transition phase of the predict module and in the refine module are replaced with dynamic convolution to enhance representational capacity, improve generalization and adapt to contour extraction tasks in complex backgrounds.

In multimodal or high-resolution image segmentation experiments, the Haar wavelet downsampling (HWD) module[25] yields a 1% to 2% increase in mIoU compared to standard downsampling methods, while significantly reducing uncertainty in feature representations. The HWD module preserves fine protrusions and curved boundaries in sweet potato contours. When combined with attention mechanisms and dynamic convolution, it effectively enhances overall segmentation quality. Therefore, the HWD module was incorporated into the downsampling stage of the refine module to strengthen its feature extraction capabilities

For sweet potato contour detection, an additional edge-prediction branch is introduced on intermediate feature maps to more accurately capture irregular, curved edges. The predicted boundary map was then fused with the original feature map via weighted summation, suppressing responses in non-contour regions while emphasizing true boundary structures. This feature-interaction-based edge-attention strategy has been validated in optical remote-sensing image saliency detection. Consequently, we modified the standard squeeze -and -excitation (SE) attention mechanism[26] by addition of edge-enhancement (ESE), integrating it into every decoding stage of the predict module. This enables the network to more effectively capture edge information.

Based on these improvements, a Sweet Potato Edge-Enhanced Contour Network (SPECNet) model was designed for sweet potato contour extraction under complex background conditions (Fig. 3).

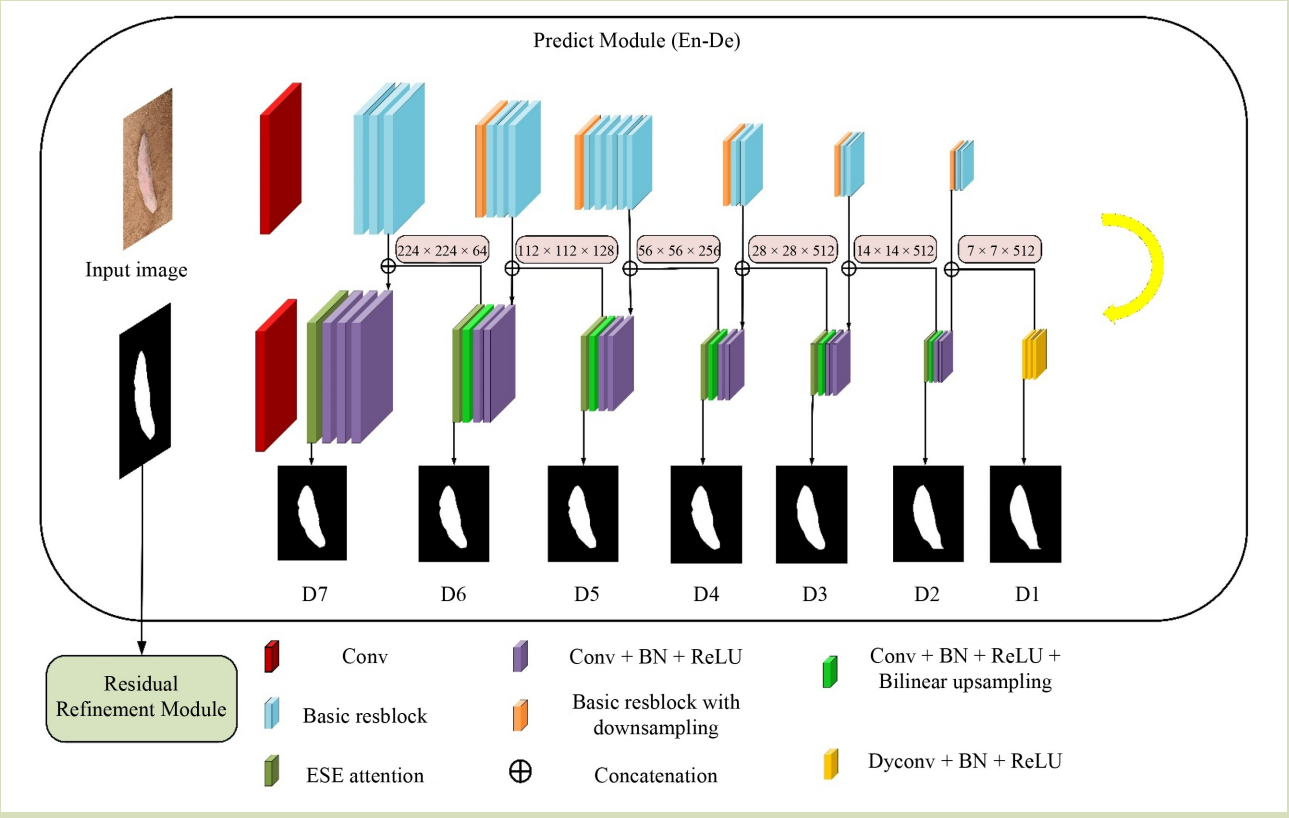


Fig. 3 SPECNet model. D1–D7 represent the output step diagrams of the decoder after each encoding-decoding process, with the final output being transferred to the residual refinement module for secondary feature extraction.

3.2 Predict module

The prediction module (Fig. 3) uses an encoder-decoder network as its backbone to simultaneously capture high-level global context and low-level details. The encoder consists of an input convolutional layer followed by six stages, each comprising basic residual blocks. The first four stages are directly adapted from ResNet-34[27].The input layer uses 64 convolutional filters of size 3 × 3 with stride 1 and no pooling operations, ensuring that feature maps before the second stage maintain the same resolution as the input image. This modification allows the network to acquire higher-resolution features in the early layers while reducing the overall receptive field. To maintain the receptive field of ResNet-34, we add two additional stages after the fourth stage. Each new stage consists of three basic residual blocks with 512 filters, followed by a 2 × 2 non-overlapping max pooling layer. Between the encoder and decoder, there is a bridging stage composed of three dynamic convolutions, batch normalization and rectified linear unit (ReLU) activations. The decoder mirrors the structure of the encoder, with each stage comprising three convolutional layers, batch normalization, ReLU activation and an Edge-enhanced squeeze-and-excitation (SE) attention mechanism (ESE) attention mechanism. It receives concatenated features from both the previous stage and the corresponding encoder stage. The outputs of the bridging stage and decoder stages undergo 3 × 3 convolutions, bilinear upsampling and S-shaped functions to generate seven saliency maps. The final map, which has the highest precision, serves as the network output and is forwarded to the refinement module.

3.3 Dynamic convolution

Due to the diverse shapes and varying growth postures of sweet potatoes, standard fixed convolutional kernels deliver limited generalization capability. Dynamic convolution dynamically generates convolutional kernels based on input data. Unlike

static convolutions, dynamic convolution adaptively adjusts its kernels for each input, enhancing the expressive power and computational efficiency of the model. This adaptability allows the model to better accommodate the diverse forms of sweet potatoes against various backgrounds, improving detection accuracy. Dynamic convolution aggregates information from multiple kernels, capturing richer feature representations that help highlight sweet potato contours in complex backgrounds (Fig. 4).

In the dynamic convolution module, the traditional perceptron is defined as:

$$y = g(W^{\mathrm{T}}x + b) \tag{1}$$

where, W used the weight matrix, x is the input vector, b is the bias term and g is the activation function.

The calculation formulas for other parameters are as follows:

$$y = g\left(\widetilde{W}^{\mathrm{T}}x + \widetilde{b}\right) \tag{2}$$

$$\widetilde{W} = \sum_{k=1}^{K} \pi_k(x)\,\widetilde{W}_k \tag{3}$$

$$\widetilde{b} = \sum_{k=1}^{K} \pi_k(x)\,\widetilde{b}_k \tag{4}$$

$$s.t.\ 0 \leqslant \pi_k(x) \leqslant 1, \sum_{k=1}^{K} \pi_k(x) = 1 \tag{5}$$

In dynamic convolution, the attention weights, denoted as π_k in Eqs. (2)–(5), are not fixed but vary with the input. This dynamic adjustment allows the convolutional kernels to adapt to different input data, enhancing the feature representation capabilities of the model compared to static convolutions.

3.4 Haar wavelet downsampling module

Accurate contour recognition relies on preserving edge details. However, traditional downsampling methods, such as average and max pooling, often lose high-frequency information like edges and may amplify background noise and soil interference. The HWD module effectively reduces the spatial resolution of feature maps while preserving edge and texture details, thereby minimizing information loss. This approach significantly enhances the performance of semantic segmentation models by addressing the limitations of traditional downsampling (Fig. 5).

3.5 Edge-enhanced squeeze-and-excitation attention module

Sweet potato images often contain interferences such as soil and withered leaves, and their contours and skin textures typically correspond to feature responses in specific channels. However, in deep neural networks with multiple layers, the SE attention module enhances the representational capacity of the network by adaptively adjusting the importance of the features of each channel. The SE module first compresses the features of each channel through global average pooling, then generates channel weights via two fully connected layers, thereby emphasizing important features and suppressing less significant ones. However, for saliency detection tasks, relying solely on channel attention may not fully capture edge information, especially in complex backgrounds.

To address this issue, we propose an improved ESE attention module to better capture and recognize sweet potato contours (Fig. 6).

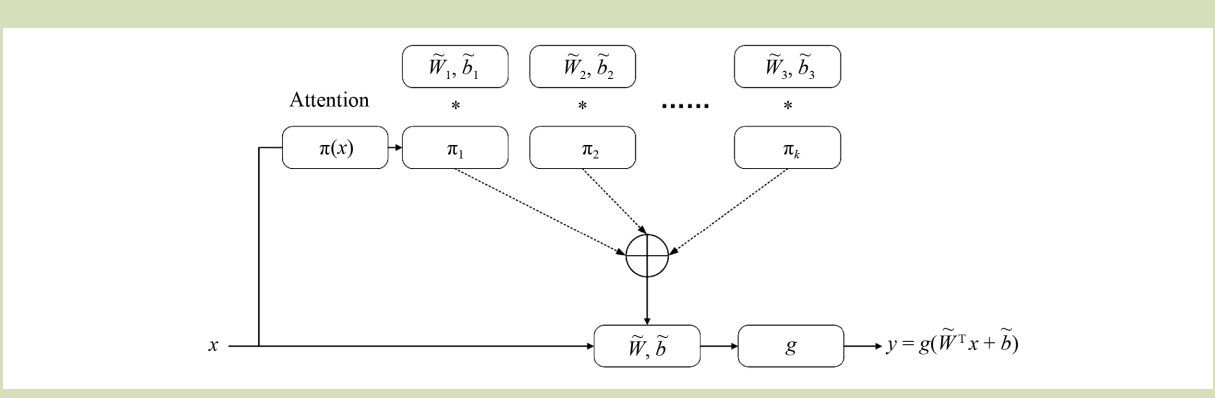


Fig. 4 Dynamic convolution. π(*x*), π_1, π_2 are dynamically generated weighting coefficients; * is convolution operation; W is convolution kernel; b is bias; g is activation function.

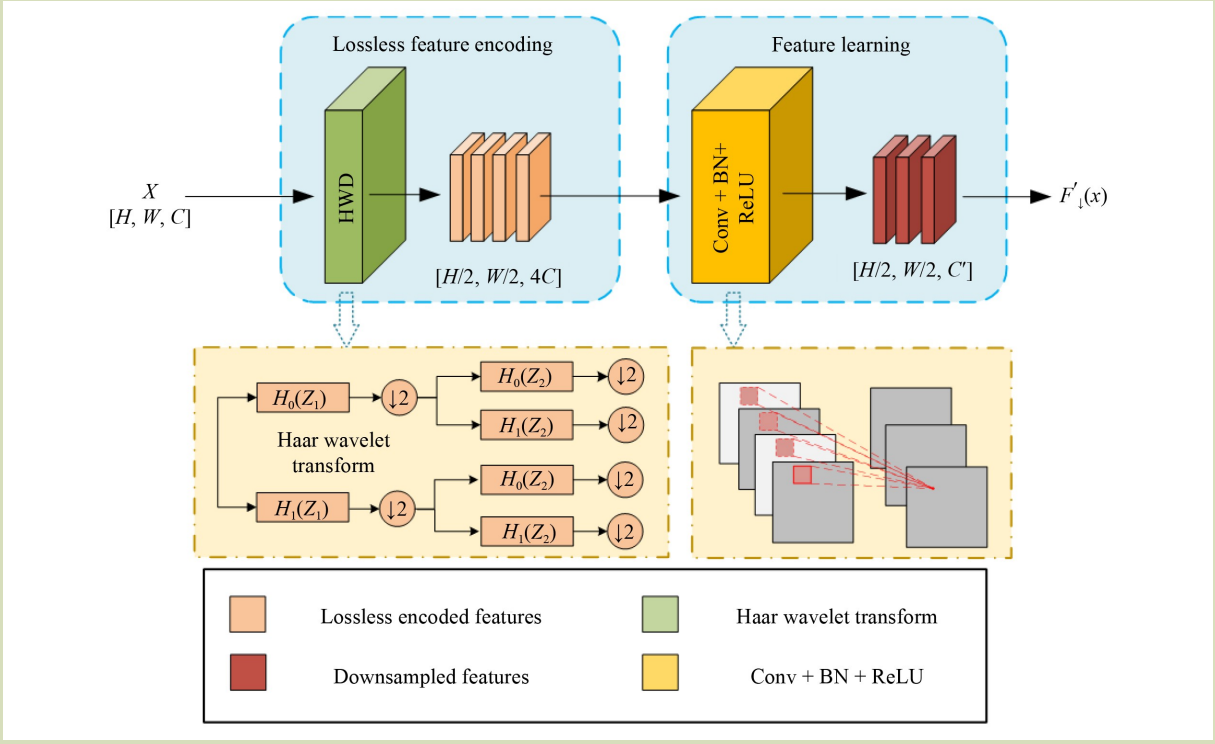


Fig. 5 Haar wavelet downsampling module.

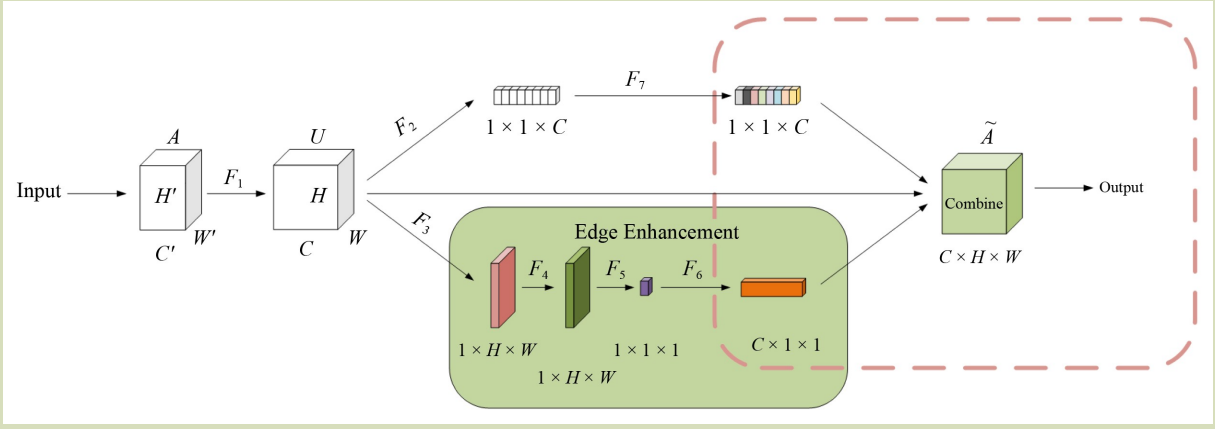


Fig. 6 ESE attention module. F_i is the output of a branch.

The ESE attention module comprises two components: the original SE mechanism and an edge guided module. The original SE mechanism compresses the input feature map into a global feature representation through adaptive average pooling, then generates channel weights via fully connected layers and ReLU activations, and finally adjusts the weights using a sigmoid function. The edge-guided module extracts and refines edge features through 3 × 3 convolutions and forms a global edge representation via adaptive average pooling. During forward propagation, the edge features are fused with

the SE weights to generate comprehensive attention weights, highlighting edge information and enhancing performance in tasks such as edge detection and semantic segmentation.

3.6 Refinement module

The refinement module aims to further refine the coarse saliency map by learning the residuals between the saliency map and the ground truth. The original refinement module, based on an encoder-decoder architecture, consists of an input layer, encoder, bridging layer, decoder and output layer. Each stage uses 3 × 3 convolutions, batch normalization, and ReLU activations, ultimately generating the saliency map. However, this structure has limited expressive power and suffers from feature information loss, restricting its capacity for feature representation.

The improved model introduces dynamic convolution to enhance representational ability and generalization, adapting to contour extraction tasks in complex backgrounds. Additionally, Haar wavelet downsampling (HWD) module achieves more efficient downsampling, retaining more critical feature information. Also, optimized skip connections and convolution operations strengthen feature fusion capabilities, enhancing the detail and accuracy of the output. These enhancements lead to superior performance in saliency detection and edge refinement tasks via the refinement module (Fig. 7).

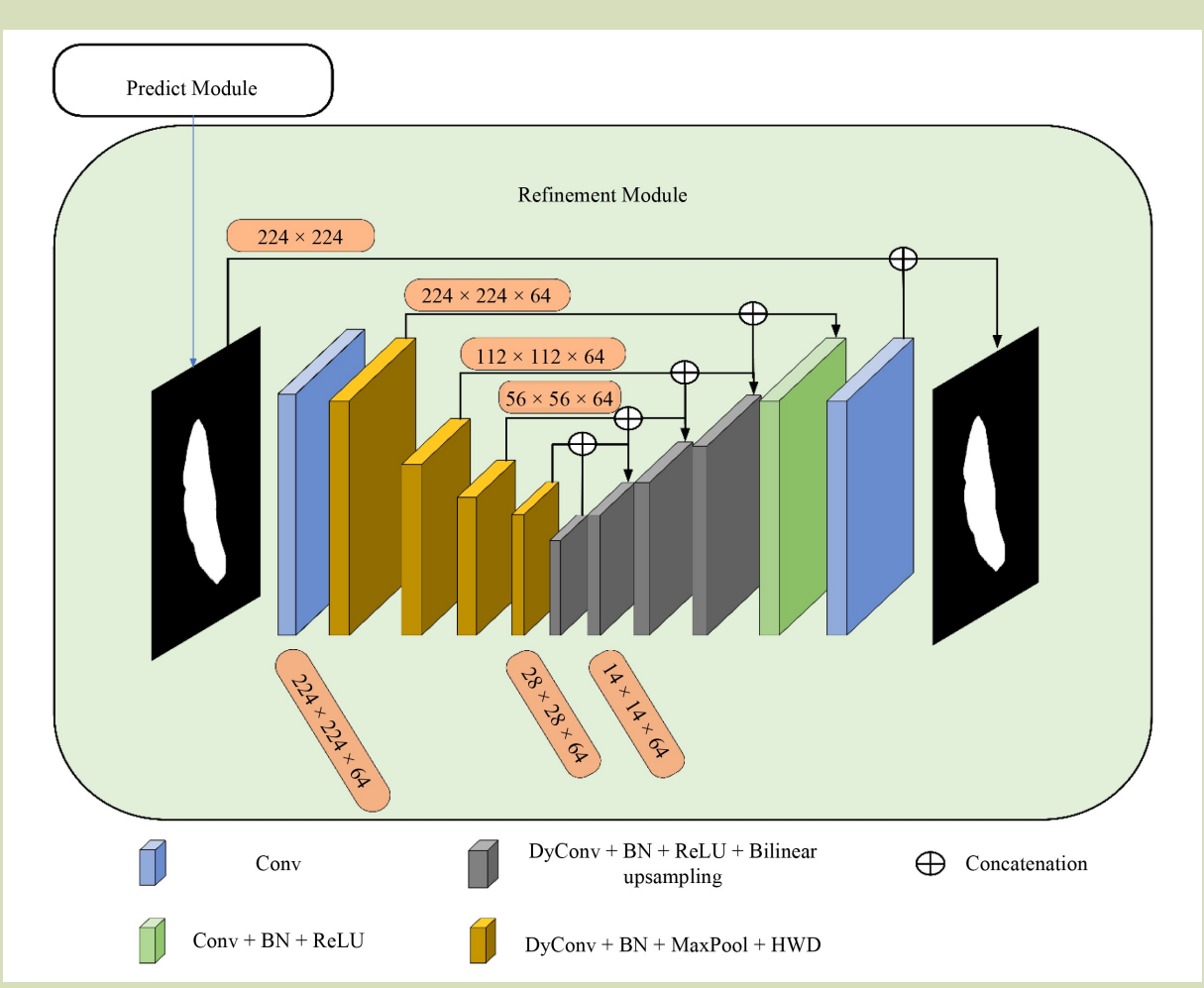


Fig. 7 Residual refinement module.

3.7 Method for calibrating sweet potato coordinates

In natural environments, agricultural robots typically use image processing techniques to extract the contours of sweet potatoes and morphological methods to calculate grasping points. In complex backgrounds, the contours and colors of sweet potatoes subtly differ from their surroundings. Salient object detection technologies can effectively distinguish sweet potatoes from background regions and identify their contours. By recognizing these contours, the most suitable grasping points can be accurately located. When combined with the three-dimensional coordinates of the grasping points measured by depth cameras, precise grasping positions are provided for robotic arms.

3.8 Potato anchor extraction

In agricultural product grading, classification is typically based on quantitative or qualitative assessments of shape and size. Crops often face rejection due to deformities, leading to food waste and economic losses for both growers and consumers[28]. For sweet potatoes, those with lower grades, such as deformed or asymmetrical ones, are often rejected during harvest. Postharvest, nearly 65% of sweet potatoes left in the field are edible and undamaged but do not meet consumer expectations for size and shape[29].

High grade sweet potatoes, with an ideal curvature value of 1.05 resembling a spindle shape[30], are most preferred. The spindle shape features tapered ends and a thicker middle, with radial symmetry. Therefore, for planar images extracted by depth cameras, using the centroid as the grasping point is scientifically valid.

Based on the ground truth map obtained from saliency object detection, image processing is first performed to obtain the desired contour. Then, the centroid position of the contour is calculated using geometric moments, serving as the center point of the target. Sweet potatoes have a certain degree of symmetry, and in terms of grasping orientation, the optimal grasping angle is typically parallel to the main axis of the object or slightly inclined. This approach maximizes contact area, increases friction and reduces the risk of slipping during grasping.

To calculate the centroid of a sweet potato, first, let $C(x,y)$ represent the centroid position in the image, where x and y are the horizontal and vertical coordinates, respectively. The contour consists of multiple pixel points. Then, let the coordinates of each point on the contour be (x_i,y_i) with a total of N points.The centroid coordinates (C_x and C_y) are calculated as:

$$c_x = \frac{\sum_{i=1}^{N} x_i}{N} \tag{6}$$

$$c_y = \frac{\sum_{i=1}^{N} y_i}{N} \tag{7}$$

In the calculation of the centroid, x_i and y_i are the coordinates of the contour points, and N is the total number of points in the contour. Principal component analysis (PCA)[31] is used to determine the main axis of the sweet potato contour. PCA provides directional information about the data distribution, which can be used to identify the longest axis of the object. For a given data set X, PCA seeks a transformation matrix P that converts the data X into a new lower-dimensional representation Z:

$$Z = X \times P \tag{8}$$

where, the columns of P are the principal components of the data.

The steps to adjust the complete grasping attitude of sweet potato are shown in Fig. 8.

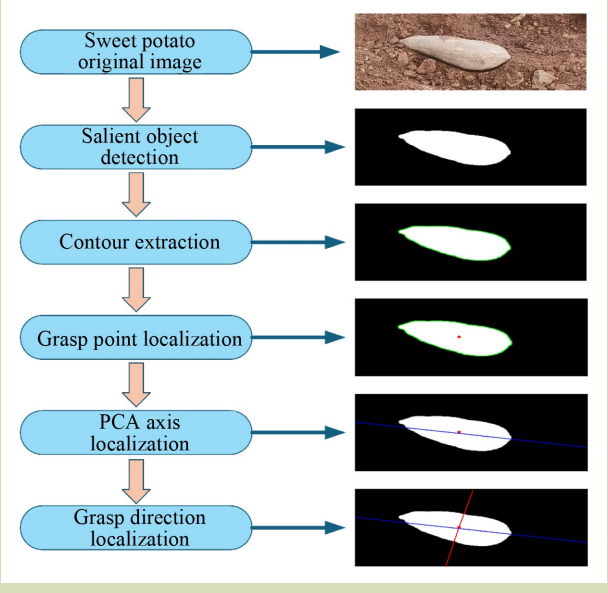


Fig. 8 Flow chart of sweet potato grasping posture recognition.

3.9 Depth camera captures the 3D coordinates of the point

In the experiment, the depth camera was used to determine the positioning and grasping direction of sweet potatoes (Fig. 9). The shape of the sweet potato can be approximated as a spindle. Point E represents the centroid of the tuber, and line segment AB is the projection of the cross-section of the sweet potato onto the saliency map. Point D corresponds to the spatial point of the projection center of the outer surface of the tuber near the camera. The camera coordinate system is denoted as O_C, with axes X_C, Y_C and Z_C, while the image pixel coordinate system is represented as O_{xy}. According to projection principles, points A, B, C and D on the sweet potato's cross-section project onto the saliency map as points a, b, c and d, respectively. The main axis line PP' is determined through principal component analysis (PCA), and line segment cd represents the grasping direction of the robotic arm obtained from the saliency map analysis. By utilizing the depth camera, the depth value of point D on the outer surface projection center of the tuber can be acquired, enabling precise control of the grasping position of the robotic arm.

4 Results and discussion

4.1 Evaluation indicators

In the task of sweet potato contour extraction, the objective is to ensure the robustness and accuracy of the algorithm under various scenarios and conditions. Therefore, evaluation metrics such as F-, E- and S-measures[32–34], and mean absolute error (MAE)[35] are selected.

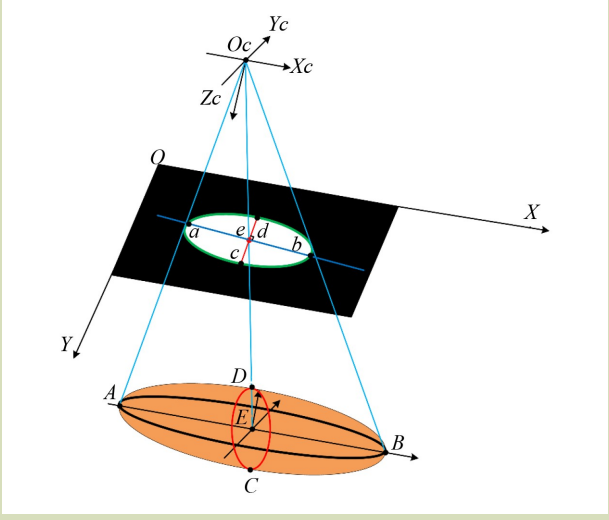


Fig. 9 Identification diagram of sweet potato grasping point.

The F-measure combines precision and recall to assess the overall performance of the model in detection tasks. It effectively reflects whether the detected contours are complete and accurate, making it suitable for evaluating the comprehensive performance of the model. The F-measure is calculated as:

$$F_\beta = \frac{(1+\beta^2)\cdot P_r \cdot R}{\beta^2 \cdot P_r + R} \tag{9}$$

where, P_r is precision, R is recall, and β is a constant factor. When $\beta = 1$, it corresponds to F_1, where precision and recall are weighted equally. When $\beta > 1$, recall has a greater impact, and when $0 < \beta < 1$, precision has a greater impact. Therefore, to better emphasize the precision of the model, β is set to 0.3.

The E-measure is a more comprehensive evaluation metric that combines pixel-level and region-level similarities in an image, providing a better reflection of global perceptual quality. In sweet potato contour extraction, the E-measure reflects the ability of the model to perceive overall shape and structure, making it suitable for evaluating the robustness of the model when handling complex backgrounds and shape variations:

$$E_\phi = \frac{1}{W \cdot H}\sum_{i=1}^{W}\sum_{j=1}^{H}\phi(P(i,j),G(i,j)) \tag{10}$$

where, W and H are the width and height of the image, $P(i,j)$ is the pixel value of the saliency detection result at position (i,j) and $G(i,j)$ is the pixel value of the ground truth map at position (i,j).

$\phi(P,G)$ is the enhancement consistency function, defined as:

$$\phi(P,G) = \frac{2 \cdot P \cdot G}{P^2 + G^2 + \epsilon} \tag{11}$$

where, ϵ is a small positive number used to prevent division by zero.

The S-measure is used to measure the structural similarity of saliency detection results. This measure is particularly suitable for evaluating the ability of the model to accurately extract contours while maintaining the overall structure, which is especially important for targets like sweet potatoes that have clear boundaries and shapes:

$$S = \alpha \cdot S_o + (1-\alpha) \cdot S_r \quad (12)$$

where, S_o is object similarity, S_r is region similarity and α is the weight factor, typically set to 0.5 to balance the contributions of object and region similarity.

Mean absolute error (MAE) is used to calculate the average pixel difference between the predicted result and the ground truth. It reflects the precision of the model at the pixel level and helps evaluate its ability to handle details, such as the accuracy of sweet potato edges and detail restoration:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (13)$$

where, n is the total number of samples, y_i is the ground truth of the i-th sample, $\hat{y}_i$ is the predicted value of the i-th sample and $|\, y_i - \hat{y}_i \,|$ is the absolute error.

4.2 Configuration of test environment and network parameter settings

We implemented a network using the publicly available framework: PyTorch 0.3.8[36]. A server equipped with an Intel Xeon Silver 4316 CPU @ 2.30GHz and an NVIDIA GeForce RTX 4090 GPU was used for both training and testing. During training, each image was first resized to 256 × 256 and then randomly cropped to 224 × 224. Some encoder parameters were initialized with a ResNet-34 model. The training loss converged after 528k iterations (200 epochs) with a batch size of 20, and the entire training process took about 30 h. During testing, the input image was resized to 256 × 256 and fed into the network to obtain its saliency map, which was then resized back to the original size of the input image. In response to expert feedback and investigate the impact of higher-resolution inputs, we additionally trained a model variant. This variant uses the same network architecture and training strategy, but with input images resized to 512 × 512 pixels.

4.3 Ablation experiment

In this subsection, we validate and analyze the modifications described above using the evaluation metrics, *F*-, *E*- and *S*-measures, and MAE. To demonstrate the effectiveness of the modified dynamic convolution, HWD downsampling module and edge-enhanced attention mechanism, we compare the altered model against the original attention mechanism and other popular, analogous modules integrated into BASNet (Table 1)

As shown in Table 1, the improved edge-enhanced attention mechanism, HWD downsampling module and dynamic convolution are more suitable for the sweet potato saliency detection task under complex backgrounds compared to other popular modules of the same type.

For issues such as insufficient depth and breadth of feature extraction and inadequate edge information capture, this study has made improvements to the U-Net[44]-based network BASNet and conducted related experiments based on experimental results are shown in Table 2.

According to the module ablation experiment (Table 2), in the saliency detection model, the performance of the model gradually improves with the stepwise addition of different modules (Dynamic convolution, HWD and ESE). Although the initial introduction of Dynamic convolution slightly increased the error, the overall model still maintained a good feature extraction ability due to its reduction in computational complexity and the number of parameters. With the addition of the HWD module, the multiscale feature expression capability was enhanced, making the model more accurate in perceiving global features and boundary information. Finally, by introducing the edge-enhanced attention mechanism, the model placed more emphasis on the edge regions of the target object, significantly improving edge detection capabilities and overall structural similarity. In summary, after the improvements to the BASNet network, with optimized computational load, parameter count and model size, the MAE decreased by 15%, while the *F*-, *E*- and *S*-measures all showed improvements.

4.4 Comparative analysis of different classification models

We compared the improved saliency detection model with six common models: ACCoNet[45], U^2-net, BASNet, BiconNet[46], YOLOv11m-Seg[47] and deeplabv3+[48]. To ensure a fair comparison, we used the same number of training epochs on identical hardware. We evaluated the segmentation quality using precision-recall (PR) curves and *F*-measure for each model. The proposed model outperformed these baselines on key metrics (Fig. 10; Table 3), demonstrating superior saliency detection performance under complex backgrounds.

The PR curve is a commonly used method to evaluate the performance of saliency probability maps. By thresholding the

Table 1 Comparative experiments on attention mechanisms

Type comparison	Configuration	MAE (mean absolute error)	Max F_β (F-measure)	Max E_ϕ (Enhanced-alignment measure)	Max S_m (Structure-measure)
Attention mechanism	BASNet	0.120	0.960	0.977	0.952
	BASNet + GE[37] GE: gather-excite attention	0.130	0.961	0.978	0.950
	BASNet + SCSE[38] SCSE: concurrent spatial and channel squeeze-and-excitation	0.122	0.964	0.979	0.953
	BASNet + SPA[39] SPA: stand-alone self-attention	0.117	0.968	0.979	0.955
	BASNet + SE	0.124	0.953	0.969	0.943
	BASNet + ESE	0.112	0.967	0.979	0.956
Downsampling strategy	B + AA[40] AA: anti-alias downsampling	0.112	0.964	0.973	0.954
	B + CA[41] CA: content-adaptive downsampling	0.120	0.965	0.975	0.956
	B + HWD	0.112	0.967	0.978	0.951
Convolution method	B + DCLS[42]	0.116	0.966	0.978	0.954
	B + partial convolution[43]	0.117	0.968	0.981	0.960
	B + dynamic convolution	0.115	0.968	0.980	0.960

Table 2 Module Ablation Experiment

Configuration	MAE (mean absolute error)	Max F_β (F-measure)	Max E_ϕ (Enhanced-alignment measure)	Max S_m (Structure-measure)
U-Net	0.567	0.822	0.916	0.843
BASNet	0.120	0.960	0.977	0.952
BASNet + dynamic convolution	0.115	0.968	0.980	0.960
BASNet + dynamic convolution + HWD	0.110	0.970	0.979	0.956
BASNet + dynamic convolution + HWD + ESE	0.102	0.970	0.984	0.968

saliency maps and comparing them with the ground truth masks, precision and recall can be calculated. In the data set, different binarization thresholds generate varying precision and recall values. As the threshold changes from 0 to 1, a series of precision-recall pairs are obtained, and these points are plotted to form the PR curve. Our proposed model is closer to the upper-right corner of the PR curve (Fig. 10(a)), indicating that, compared to other models, it has higher accuracy (precision) and a lower miss rate (recall) when predicting positive classes.

The F-measure threshold curve shows the balance between precision and recall at different thresholds. Specifically, this curve is typically used to measure the overall performance of the model by calculating the trend of the F-value as the threshold varies. Our proposed model maintains a more balanced performance at different decision thresholds (Fig. 10(b)), making it suitable for applications that require high precision and recall.

Additionally, Table 3 summarizes the results of all data sets and

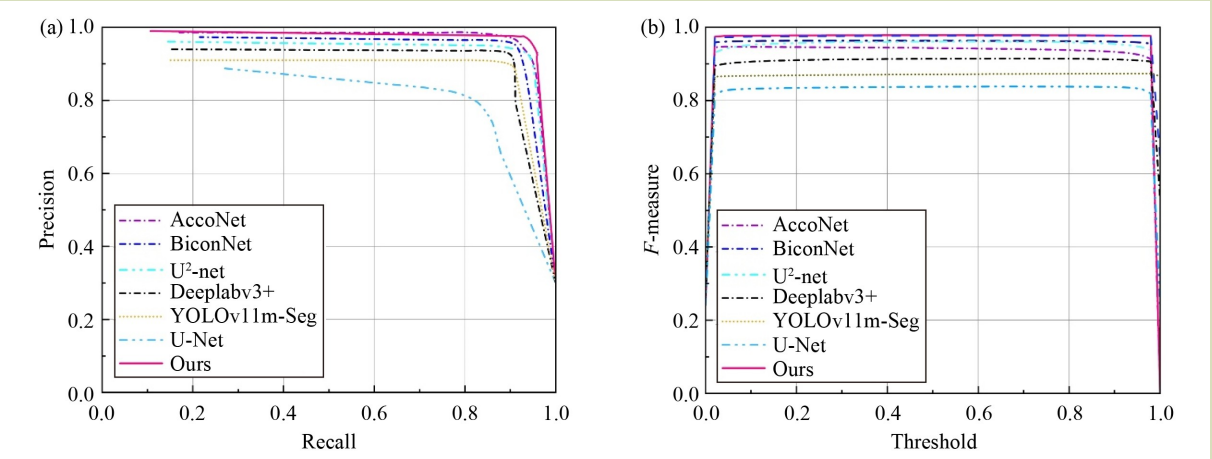


Fig. 10 Precision-recall curve (a) and *F*-measure threshold curve (b) for the proposed model and six comparator models.

Table 3 Model performance comparison

Model	MAE (mean absolute error)	Max F_β (*F*-measure)	Max E_ϕ (Enhanced-alignment measure)	Max S_m (Structure-measure)
ACCoNet	0.116	0.970	0.962	0.933
U2-net	0.169	0.958	0.967	0.949
BASNet	0.120	0.960	0.977	0.952
BiconNet	0.114	0.964	0.974	0.958
YOLOv11m-Seg	0.143	0.874	0.879	0.823
Deeplabv3+	0.110	0.914	0.899	0.902
Proposed model (224 × 224)	0.102	0.970	0.984	0.968
Proposed model (512 × 512)	0.082	0.982	0.990	0.981

provides a qualitative comparison with six other methods. We can see that the accuracy of our proposed model in identifying sweet potato contours in complex backgrounds is significantly higher than that of other models.

In response to various issues in sweet potato contour recognition under complex backgrounds, we visually compared our proposed model with other models using GT maps as the benchmark (Fig. 11). We can see that our method is able to accurately segment salient objects in a variety of challenging scenarios, including images with multiple cluttered sweet potatoes, semi-buried sweet potatoes, overlapping sweet potatoes and complex background interference. The saliency probability maps generated by our method are more uniform compared to other methods. Also, our results are clearer, more comprehensive and precise, enabling more effective identification and segmentation of the target regions, thus improving the overall detection performance.

4.5 Platform deployment and validation

To further validate the effectiveness of the improved model in practical applications, it was deployed on a laptop equipped with an Intel Core i5-13500HX CPU and an NVIDIA GeForce RTX 4060 Laptop GPU for remotely controlling the sweet potato picking robot. The sweet potato picking robot consisted of an image acquisition module, gripping device, robotic arm and control module. The system used a depth camera to capture color and depth images of the sweet potato (Fig. 12).

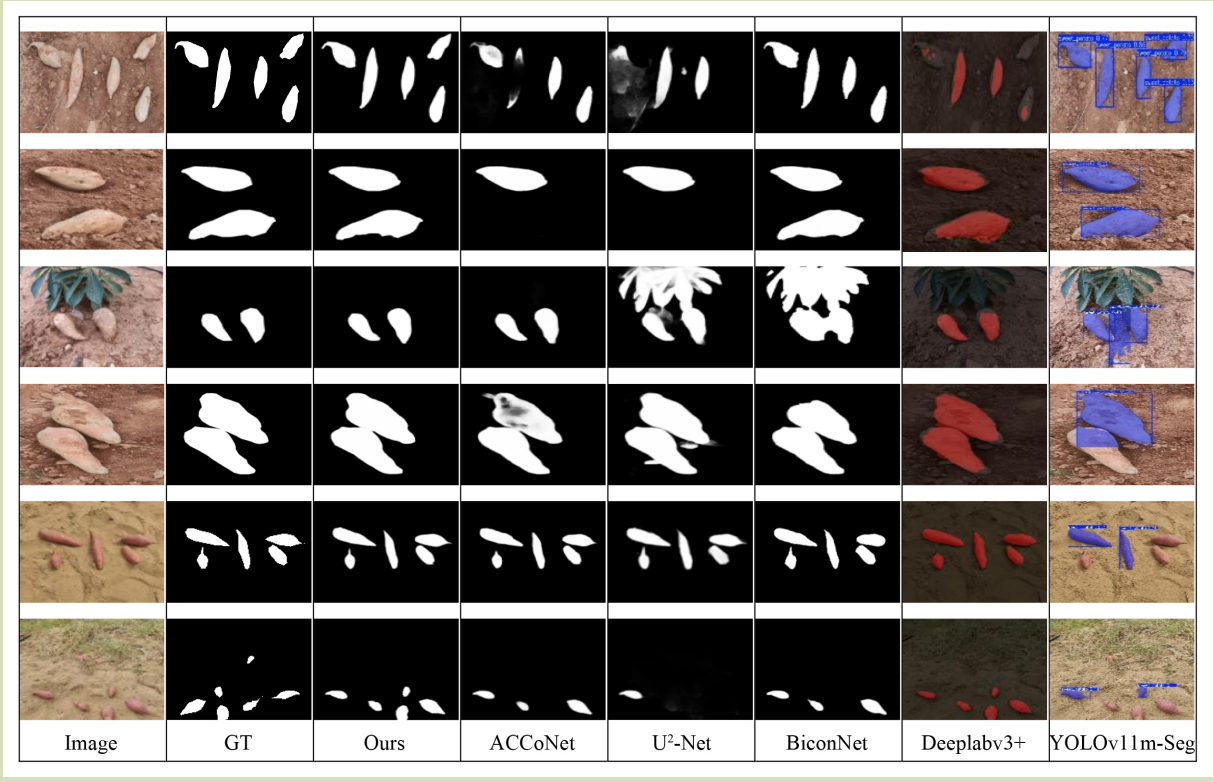


Fig. 11 Qualitative comparison of results for the proposed model and six other models.

The color image is transmitted to a server running the saliency detection algorithm, which extracts the sweet potato contour.

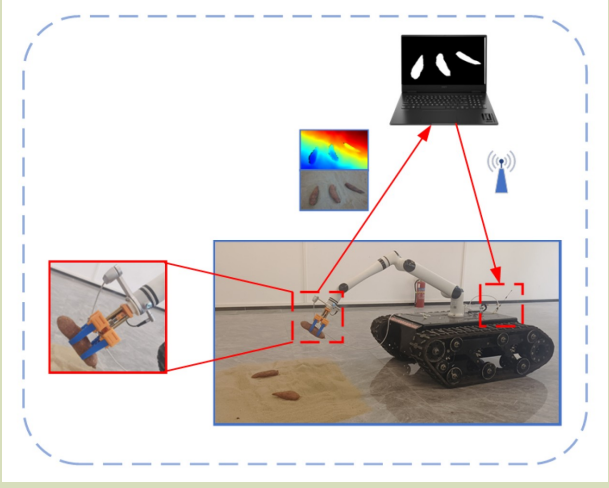


Fig. 12 Structure diagram of sweet potato picking robot.

Based on the contour, the grasping point and direction are determined and their corresponding 3D coordinates on the depth map are transmitted to the robotic arm, which then controls the end-effector to complete the sweet potato grasping task.

For this study, 100 sweet potatoes were selected for recognition and grasping experiments. First, during the sweet potato grasping trials, the shapes of the specimens were organized and classified, as summarized in Table 4, with their characteristic morphologies illustrated in Fig. 13. The root length L is defined as the maximum axial length of each sweet potato and the maximum root diameter W is defined as the largest cross-sectional diameter measured perpendicular to the longest axis.

Due to site constraints, the experiments were first conducted in an indoor simulated sandy-soil environment (Fig. 12), and subsequently the grasping trials were performed in a field soil environment in the Batou (locality) experimental field in Sanya, Hainan Province (Fig. 14). Through experimentation, it

Table 4 Classification of sweet potato shapes

Type	Shape characteristic	Quantitative description	Proportion
A	Slender	$L \geqslant 18$ cm and $L/W \geqslant 1.8$	80%
B	Sturdy	$W \geqslant 5$ cm and $1.2 < L/W < 1.8$	10%
C	Compact	$L \leqslant 12$ cm and $L/W \leqslant 1.2$	10%

Fig. 13 Representation of the sweet potato type used in the sweet potato grasping experiment. Types A–C are summarized in Table 4.

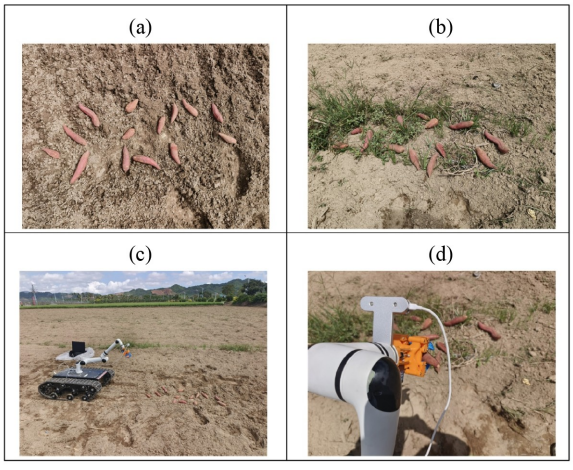


Fig. 14 Experimental contexts: (a, c) field and (b, d) grassland.

has been observed that the average recognition speed reaches 35 frames per second.

The samples were divided into 10 independent groups, and a randomized complete block design was used to randomly distribute the sweet potatoes across the experimental field, thereby minimizing systematic error due to inter -group variation. Grasping trials for each group were conducted at 7:00, 12:00 and 18:30 to ensure a comprehensive assessment of system performance under varying illumination conditions. A grasp was deemed successful if the manipulator lifted the target sweet potato from its original position and accurately placed it within a predetermined target area. To guarantee experimental reproducibility, each group was tested three times at each time point, resulting in a total of 10 groups × 3 times of day × 3 replicates being 90 individual units. The aggregated results of these experiments are presented in Table 5.

From the perspectives of experimental environment, experimental timing and sweet-potato morphological characteristics, the latter emerges as the primary factor affecting grasp success rate. Of the three shape categories, the thick and short types (B and C in Fig. 13) are more prone to slippage or grasp failure during end-effector contact. Field observations indicate that this issue stems mainly from geometric mismatch between the end structure of the manipulator and the sweet-potato morphology, resulting in insufficient grip stability.

Additionally, environmental factors in the field also influence grasp performance. In particular, ground interferences such as weeds and withered leaves can occlude visible regions of the sweet potato, thereby impairing the ability of the vision algorithm to localize the target accurately. Such occlusions can lead to misestimation of the centroid position, causing off-target or missed grasps.

With respect to experimental timing, assessments conducted at noon (when illumination is the strongest) had a slightly higher grasp success rate compared to morning and late-afternoon

Table 5 Grasping success rates (%) under varying environments and times: overall and two types as illustrated in Fig. 13

Experimental environment	Time of day	Success rate (%)			
		Overall	Type A	Type B	Type C
Sandy-soil (indoor)	–	82	96	10	40
Field-soil	Morning	81	93	20	40
	Noon	82	93	20	50
	Evening	79	93	10	30
Grassland	Morning	74	85	20	40
	Noon	76	87	10	50
	Evening	72	83	20	30
Average	–	78	90	16	40

Note: Success rate, the success rate of this type of capture. Types A–C are summarized in Table 4

sessions. This suggests that stronger lighting conditions can improve image-recognition quality. However, the overall difference across the three times of day was minor, indicating that the improved system maintains a degree of robustness against variations in natural light.

In summary, the proposed enhanced grasping model demonstrates reliable stability and robustness when faced with varying sweet-potato shapes and complex environmental backgrounds. These results validate its feasibility and potential for practical deployment in real-world agricultural fields.

5 Conclusions

This study addresses the challenges faced by agricultural manipulators when grasping sweet potatoes, including difficulties in locating grasp points, poor segmentation between the crop and background, and insufficient environmental adaptability. An improved method for sweet potato grasp point localization based on a U-net-based saliency detection network (BASNet) is proposed. The proposed method has three key advantages.

(1) The method replaces standard convolution modules with dynamic convolutions to enhance representation and generalization capabilities, enabling accurate extraction of the key features of sweet potato contours in complex backgrounds.

(2) An HWD module is embedded during the downsampling process of the network. This module retains more semantic information and local details while reducing the resolution of feature maps, thereby improving the segmentation between sweet potatoes and the background.

(3) An ESE attention mechanism is introduced, which builds upon the traditional SE attention mechanism by emphasizing edge features. By effectively extracting and reinforcing target boundary information, the accuracy of sweet potato contour detection is improved, making grasp point localization more precise.

Based on these improvements, a sweet potato grasp point localization method based on the SPECNet network is presented.

In ablation experiments, the proposed module improvements reduced the mean absolute error of the model by 15%, with similar gains in other indicators. With all modifications combined, the proposed model achieved an overall recognition mean absolute error of 10.9%, an F-measure of 97.0%, an E-measure of 98.4%, and an S-measure of 96.8%. Compared with existing mainstream recognition models, the proposed model provides superior recognition performance. Also, in sweet potato picking experiments, a grasp success rate of 78% was achieved, meeting practical application requirements.

In summary, the SPECNet model proposed in this study demonstrates high accuracy and reliability in saliency detection and grasp point localization tasks, showing promising potential for practical applications.

Acknowledgements

This research was sponsored by National Natural Science Foundation of China (32272003), National Modern Agricultural Industry Technology System Project, China (CARS-09-P32), and Key Research and Development Project of Hainan Province, China (ZDYF2023XDNY039).

Compliance with ethics guidelines

Ranbing Yang, Ang Zhao, Danyang Lv, Yongfei Pan, Hongfei Zhu, Xinyu Guo, Jian Zhang, and Jianqi Hou declare that they have no conflicts of interest or financial conflicts to disclose. This article does not contain any studies with human or animal subjects performed by any of the authors.

REFERENCES

1. Huang J. High-yield Cultivation Techniques of Sweet Potato. Haikou: *Hainan Publishing House and Sanhuan Publishing House*, 2010 (in Chinese)
2. Xie B W, Jin M H, Duan J L, Li Z W, Wang W S, Qu M Y, Yang Z. Design of adaptive grippers for fruit-picking robots considering contact behavior. *Agriculture*, 2024, **14**(7): 1082
3. Yu J C, Weng K J, Liang G Y, Xie G H. A Vision-based Robotic Grasping System Using Deep Learning for 3D Object Recognition and Pose Estimation. In: Proceedings of the 2013 IEEE International Conference on Robotics and Biomimetics (ROBIO), Shenzhen, China. *IEEE*, 2013, 1175–1180
4. Indira P, Arafat I S, Karthikeyan R, Selvarajan S, Balachandran P K. Fabrication and investigation of agricultural monitoring system with IoT & AI. *SN Applied Sciences*, 2023, **5**(12): 322
5. Alaaudeen K M, Selvarajan S, Manoharan H, Jhaveri R H. Intelligent robotics harvesting system process for fruits grasping prediction. *Scientific Reports*, 2024, **14**(1): 2820
6. Wang D D, Xu Y, Song H B, He D J. Localization method of picking point of apple target based on smoothing contour symmetry axis algorithm. *Transactions of the Chinese Society of Agricultural Engineering*, 2015, **31**(5): 167–174 (in Chinese)
7. Wang H M, Ge Q B, Wu Q T, Zheng R J, Zhu J L. Edge detection method for UAV in fog based on improved Canny operator. *Control Engineering*, 2025, **32**(6): 1030–1038 (in Chinese)
8. Gao J J, Wang R A, Wang Z X, Dong Q J, Gao H H, Xin S Q. Edge-aware point cloud registration. *Journal of Computer-Aided Design & Computer Graphics*, 2024, **36**(7): 1122–1130 (in Chinese)
9. Patel V K, Abhishek K, Selvarajan S. Optimized recurrent neural network-based early diagnosis of crop pest and diseases in agriculture. *Discover Computing*, 2024, **27**(1): 43
10. Fuentes A, Yoon S, Kim S C, Park D S. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*, 2017, **17**(9): 2022
11. Wan Y, Yang H Y, Wang Y L, Luo J J, Mei M. Recognition of rice disease based on image segmentation and convolution neural network. *Acta Agriculturae Boreali-occidentalis Sinica*, 2022, **31**(2): 246–256 (in Chinese)
12. Lv J D, Zhao D A, Ji W, Ding S H. Recognition of apple fruit in natural environment. *Optik*, 2016, **127**(3): 1354–1362
13. Yu X, Wang Z, Jing H T, Jin X L, Nie C W, Bai Y, Wang Z. Maize tassel segmentation based on deep learning method and RGB image. *Journal of Zhejiang University (Agric. & Life Sci.)*, 2021, **47**(4): 451–463 (in Chinese)
14. Wu F Y, Zhu R, Meng F, Qiu J J, Yang X P, Li J H, Zou X J. An enhanced cycle generative adversarial network approach for nighttime pineapple detection of automated harvesting robots. *Agronomy*, 2024, **14**(12): 3002
15. Meng F, Li J H, Zhang Y Q, Qi S J, Tang Y C. Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks. *Computers and Electronics in Agriculture*, 2023, **214**: 108298
16. Yu Z D, Feng C, Liu M Y, Ramalingam S. CASENet: Deep Category-aware Semantic Edge Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA. *IEEE*, 2017, 1761–1770
17. Li G B, Yu Y Z. Visual saliency detection based on multiscale deep CNN features. *IEEE Transactions on Image Processing*, 2016, **25**(11): 5012–5024
18. Wang L Z, Wang L J, Lu H C, Zhang P P, Ruan X. Salient object detection with recurrent fully convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019, **41**(7): 1734–1746
19. He F L, Guo Y C, Gao C, Chen J. Image segmentation of ripe mulberries based on visual saliency and pulse coupled neural

network. *Transactions of the Chinese Society of Agricultural Engineering*, 2017, **33**(6): 148–155 (in Chinese)

20. Yuan C, Zhang X, Wang J H, Zhao M X, Xu D W. Pinch point extraction method for *Phalaenopsis* tissue-cultured seedlings based on salient features. *Transactions of the Chinese Society of Agricultural Engineering*, 2023, **39**(13): 151–159 (in Chinese)
21. Qin X B, Zhang Z C, Huang C Y, Gao C, Dehghan M, Jagersand M. BASNet: Boundary-aware Salient Object Detection. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA. *IEEE/CVF*, 2019, 7471–7481
22. Qin X B, Zhang Z C, Huang C Y, Dehghan M, Zaiane O R, Jagersand M. U^2-Net: going deeper with nested U-structure for salient object detection. *Pattern Recognition*, 2020, **106**: 107404
23. Qin X B, Fan D P, Huang C Y, Diagne C, Zhang Z C, Cabeza S A A, Suárez A, Jagersand M, Shao L. Boundary-aware segmentation network for mobile and web applications. *arXiv preprint*, 2021, arXiv:2101.04704
24. Zhang N, Han J W, Liu N, Shao L. Summarize and Search: Learning Consensus-aware Dynamic Convolution for Co-saliency Detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada. *IEEE*, 2021, 4147–4156
25. Xu G P, Liao W T, Zhang X, Li C, He X W, Wu X L. Haar wavelet downsampling: a simple but effective downsampling module for semantic segmentation. *Pattern Recognition*, 2023, **143**: 109819
26. Hu J, Shen L, Sun G. Squeeze-and-Excitation Networks. In: Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. *IEEE*, 2018, 7132–7141
27. He K M, Zhang X Y, Ren S Q, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA. *IEEE*, 2016, 770–778
28. Boyette M D, Tsirnikas A L. Evaluating the Shape and Size Characteristics of Sweetpotatoes Using Digital Image Analysis. In: Proceedings of the 2017 ASABE Annual International Meeting, Spokane, WA, USA. *American Society of Agricultural and Biological Engineers*, 2017, 1700038
29. Johnson L K, Dunning R D, Bloom J D, Gunter C C, Boyette M D, Creamer N G. Estimating on-farm food loss at the field level: a methodology and applied case study on a North *Carolina* farm. *Resources, Conservation and Recycling*, 2018, **137**: 243–250
30. Liu H J, Hunt S, Yencho G C, Pecota K V, Mierop R, Williams C M, Jones D S. Predicting sweetpotato traits using machine learning: impact of environmental and agronomic factors on shape and size. *Computers and Electronics in Agriculture*, 2024, **225**: 109215
31. Yang J, Zhang D, Frangi A F, Yang J Y. Two-dimensional PCA: a new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2004, **26**(1): 131–137
32. Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned Salient Region Detection. In: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA. *IEEE*, 2009, 1597–1604
33. Fan D P, Gong C, Cao Y, Ren B, Cheng M M, Borji A. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI - 18), Stockholm, Sweden. *AAAI Press*, 2018, 698–704
34. Fan D P, Cheng M M, Liu Y, Li T, Borji A. Structure-measure: a new way to evaluate foreground maps. *arXiv preprint*, 2017, *arXiv*:1708.00786
35. He K M, Chen X L, Xie S N, Li Y H, Dollár P, Girshick R. Masked Autoencoders Are Scalable Vision Learners. In: Proceedings of the IEEE/CVF 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA. *IEEE*, 2022, 15979–15988
36. Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A. Automatic Differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems Workshop on Autodiff (NIPS-W). Long Beach, California, USA: *NIPS*, 2017, 1–4
37. Hu J, Shen L, Albanie S, Sun G, Vedaldi A. Gather-Excite: Exploiting Feature Context in Convolutional Neural Networks. In: 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada. *Curran Associates*, 2018
38. Roy A G, Navab N, Wachinger C. Concurrent Spatial and Channel Squeeze & Excitation in Fully Convolutional Networks. In: Proceedings of the 21st International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2018), Granada, Spain. Cham: *Springer*, 2018, 421–429
39. Prajit R, Niki P, Ashish V, Irwan B, Anselm L, Jonathon S. Stand-alone Self-attention in Vision Models. In: Proceedings of the 33rd International Conference on Neural Information Processing Systems. *Curran Associate*, 2019, **7**: 68–80
40. Richard Z. Making Convolutional Networks Shift-invariant Again. In: Proceedings of the 36th International Conference on Machine Learning (ICML 2019). Long Beach, California, USA: *PMLR*, 2019, **97**: 7324–7334
41. Hesse R, Schaub-Meyer S, Roth S. Content-adaptive Downsampling in Convolutional Neural Networks. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Vancouver, BC, Canada. *IEEE*, 2023, 4544–4553
42. Ismail K H, Thomas P, Timothée M. Dilated Convolution with

Learnable Spacings. In: Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Honolulu, United States. *OpenReview*, 2023, 1–7

43. Liu G L, Reda Fitsum A, Shih Kevin J, Wang T C, Tao A, Catanzaro B. Image Inpainting for Irregular Holes Using Partial Convolutions. In: Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich, Germany. *Springer*, 2018, 85–100
44. Olaf R, Philipp F, Thomas B. U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Munich, Germany. *Springer*, 2015, 234–241
45. Li G Y, Liu Z, Zeng D, Lin W S, Ling H B. Adjacent context coordination network for salient object detection in optical remote sensing images. *IEEE Transactions on Cybernetics*, 2023, **53**(1): 526–538
46. Yang Z Y, Soltanian-Zadeh S, Farsiu S. BiconNet: an edge-preserved connectivity-based approach for salient object detection. *Pattern Recognition*, 2022, **121**: 108231
47. Sapkota R, Karkee M. Comparing YOLOv11 and YOLOv8 for instance segmentation of occluded and non-occluded immature green fruits in complex orchard environment. *arXiv preprint*, 2024, arXiv:2410.19869.
48. Chen L C, Zhu Y K, Papandreou G, Schroff F, Adam H. Encoder-decoder with Atrous Separable Convolution for Semantic Image Segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV 2018), Munich. Cham: *Springer*, 2018, 833–851