Detection of the farmland plow areas using RGB-D images with an improved YOLOv5 model

Jiangtao Ji[1,2,3], Zhihao Han[1], Kaixuan Zhao[1,2,3], Qianwen Li[2,3,4*], Shucan Du[1]

(1. *College of Agricultural Equipment Engineering, Henan University of Science and Technology, Luoyang 471003, Henan, China*;
2. *Science & Technology Innovation Center for Completed Set Equipment, Longmen Laboratory, Luoyang 471023, Henan, China*;
3. *Collaborative Innovation Center of Machinery Equipment Advanced Manufacturing of Henan Province, Luoyang 471003, Henan, China*;
4. *School of Art and Design, Henan University of Science and Technology, Luoyang 471003, Henan, China*)

Abstract: Recognition of the boundaries of farmland plow areas has an important guiding role in the operation of intelligent agricultural equipment. To precisely recognize these boundaries, a detection method for unmanned tractor plow areas based on RGB-Depth (RGB-D) cameras was proposed, and the feasibility of the detection method was analyzed. This method applied advanced computer vision technology to the field of agricultural automation. Adopting and improving the YOLOv5-seg object segmentation algorithm, first, the Convolutional Block Attention Module (CBAM) was integrated into Concentrated-Comprehensive Convolution Block (C3) to form C3CBAM, thereby enhancing the ability of the network to extract features from plow areas. The GhostConv module was also utilized to reduce parameter and computational complexity. Second, using the depth image information provided by the RGB-D camera combined with the results recognized by the YOLOv5-seg model, the mask image was processed to extract contour boundaries, align the contours with the depth map, and obtain the boundary distance information of the plowed area. Last, based on farmland information, the calculated average boundary distance was corrected, further improving the accuracy of the distance measurements. The experiment results showed that the YOLOv5-seg object segmentation algorithm achieved a recognition accuracy of 99% for plowed areas and that the ranging accuracy improved with decreasing detection distance. The ranging error at 5.5 m was approximately 0.056 m, and the average detection time per frame is 29 ms, which can meet the real-time operational requirements. The results of this study can provide precise guarantees for the autonomous operation of unmanned plowing units.

Keywords: plow areas, RGB-D camera, YOLO, object segmentation, contour boundary, average distance
DOI: 10.25165/j.ijabe.20241703.8810

Citation: Ji J T, Han Z H, Zhao K X, Li Q W, Du S C. Detection of the farmland plow areas using RGB-D images with an improved YOLOv5 model. Int J Agric & Biol Eng, 2024; 17(3): 156–165.

1 Introduction

With the rapid development of intelligence and information technology, as well as smart agriculture, the unmanned driving and autonomous operation of agricultural machinery have become popular research fields. Intelligent agricultural equipment will constitute the main direction of agricultural mechanization development[1-3].

An intelligent farmland operation machine should be able to correctly identify not only the marking line[4] between the unploughed area and the plowed area but also, simultaneously, the boundary of the farmland to ensure the integrity of the farmland operation process. The traditional automatic navigation system needs to be manually punched to mark the boundary of the ground beforehand, which additionally increases the workload; however, machine vision can obtain images of the farmland environment in real-time, obtain the required target features, and realize distance detection via target detection, which has the advantages of high speed, high accuracy, noncontact nature, automation, and multifunctionality[5], which is important for the automation and intelligence of agricultural machinery. The farmland boundary of a farm field is bounded by ridges and road edges, which are important markers for distinguishing plowed and unplowed areas. Accurately identifying and calculating relative distances are crucial for the subsequent operation of farm machinery.

Current related researches mainly focused on methods based on traditional vision[6-8] and distance sensors[9-11]. Traditional visual inspection uses cameras to obtain images of farmland boundaries and then to obtain boundary lines. Wang et al.[12] employed machine vision techniques to divide farmland boundary images into 8 subregions, each of which was solved for grayscale jump feature points and linearly fitted with a robust regression method to obtain the main extensions of irregular land heads. Cai et al.[13] used a support vector machine algorithm to segment paddy field ridge images based on superpixel segmentation and then extracted the ridge boundaries using the Hough transform. Zhu et al.[14] performed image processing with the HIS color space model using an improved region splitting aggregation algorithm and the Moore boundary tracking algorithm to extract field road boundaries. Ollis et al.[15] developed an automatic harvester based on a visual system that can recognize the boundaries of the field, adapt to local changes in lighting and crops, and eliminate the interference of shadows. Astrand et al.[16] proposed a robust recognition method for plant

Received date: 2024-01-14 **Accepted date:** 2024-04-29
Biographies: **Jiangtao Ji**, PhD, Professor, research interest: agricultural production mechanization, Email: jjt0907@163.com; **Zhihao Han**, Master candidate, research interest: modern agricultural equipment theory and technology, Email: zh.han@stu.haust.edu.cn; **Kaixuan Zhao**, PhD, Associate Professor, research interest: machine vision. Email: kx.zhao@haust.edu.cn; **Shucan Du**, Master candidate, research interest: modern agricultural equipment theory and technology, Email: sc.du@stu.haust.edu.cn.
***Corresponding author: Qianwen Li**, PhD, research interest: automatic control. Henan University of Science and Technology, Luoyang 471003, Henan, China. Tel: +86-18623790613, Email: qianwenli166@163.com.

rows based on the Hough transform, which determines whether a crop row has ended and reached the boundary of the farmland by detecting anomalies in the offset and heading angle of the crop row. The above-mentioned traditional 2D image processing for farmland boundary recognition is limited by perspective issues, and missing depth information, which limit its accuracy and robustness.

The utilization of 3D image processing will enhance the accuracy of information processing. Wei et al.[17] acquired 3D data via a binocular camera, extracted crop height information via an improved clustering method and color image segmentation, and subsequently extracted harvest boundary points. Hong et al.[18] used binocular cameras to perform adaptive threshold point cloud extraction and interference cancellation on 3D point clouds constructed from disparity maps, achieving the recognition of field ridge boundaries. However, the construction of 3D point clouds typically requires a large amount of memory and processing time, resulting in low efficiency.

Radar ranging is another method used to detect the distance between farmland boundaries and agricultural machinery. Chen Binbin et al.[19] proposed the laser and inertial measurement unit algorithm, which accurately and in real-time detects grain boundaries by considering the correlation between sampling points and laser harnesses. Using laser nondestructive detection technology, Wei et al.[20] developed an online recognition system for harvesting boundaries of combine harvesters. Zhang et al.[21] proposed the fusion detection of cameras and millimeter wave radar, which obtains accurate information such as the shape, distance, and height of the field ridges ahead via visual detection and vertical placement of the radar. When using distance sensors for detection, it is necessary to establish a distinct protrusion feature on the farmland boundary and use sensors such as LiDAR to perceive the boundary. The above-mentioned traditional image-based methods for recognizing farmland boundaries generally suffer from long processing time, which usually takes hundreds of milliseconds. Sensor-based boundary detection methods, on the other hand, face many drawbacks such as high cost, increased complexity, and limited versatility. Against this backdrop, the application of deep learning technologies has brought breakthrough improvements to agricultural field boundary recognition. Intelligent machine vision, primarily utilizing deep learning and convolutional neural networks, has become the mainstream research direction.

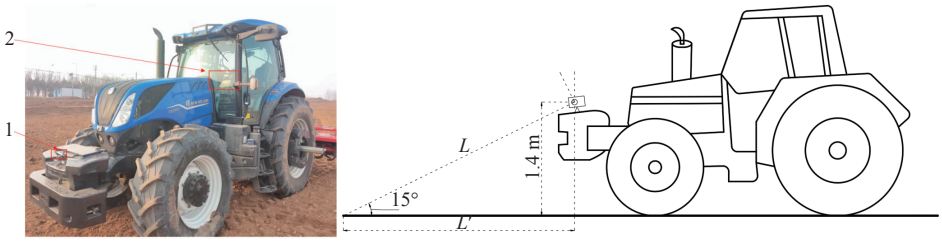
Persello et al.[22] used a fully convolutional network with globalization and grouping algorithms to effectively detect and delineate farmland boundaries by learning complex spatial-contextual features. Qiao et al.[23] developed a deep learning method for recognizing farmland boundary images, constructing a six-category dataset trained on the MobileNetV2 network, which achieved a Top-1 accuracy of 98.5% and an F1-score of 97.0% on validation and test sets, respectively.

In summary, to meet the real-time detection and ranging requirements of the farmland boundary of intelligent agricultural machines during plowing operations, this study proposed a plow area detection method based on RGB-D images with an improved YOLO model. First, color and depth map information of the front area of agricultural machinery was obtained through a depth camera, and the improved YOLO model was used to perform forward processing on the color map to obtain plow area information. Second, combined with depth images, the distance between the working area boundary and the nonworking area boundary. Last, the reliability of depth recognition farmland boundary detection is verified via real-field machine experiments.

2 Materials and methods

2.1 Data acquisition and calibration

When agricultural machinery is operating in the field, obstacles such as trees, ditches, and road edges may appear at the end of the field. The data collected in this study include data from areas with different situations, as previously mentioned. The Intel RealSense D455 depth camera produced by Intel company (USA) was installed on the Newfoundland 2204 tractor at an angle of 15° to the ground. The camera has an installation height of 1.4 m, an image resolution of 640×480 pixels, and a capture frame rate of 60 fps, and 3000 images were captured in an unplowed farmland environment. Figure 1 shows the installation diagram of the camera, and Figure 2 shows the collected image samples of the farmland boundary in several scenarios.



a. Actual installation location diagram b. Installation location schematic diagram

1. Camera 2. Computer terminal

Figure 1 Diagram of data collection platform

According to the work requirements mentioned earlier, the AnyLabeling tool is used for manual annotation after scaling the image. This tool combines the intelligent annotation function of Segment Anything and YOLO model, which can efficiently and accurately annotate the image and assign mask labels for subsequent segmentation tasks.

Before training, in order to ensure obtaining a more accurate dataset and improve the robustness of the network model, the dataset is subjected to Mosaic image rotation, mirror symmetry, translation transformation, and mean processing to expand the amount of data[24], and the training set, validation set, and test set were divided in an 8:1:1 ratio.

2.2 Modeling of farmland edge segmentation

Segmentation is the process of identifying the desired target pixel by pixel in the original image. Instance segmentation requires both bounding box detection and localization of the target, as well as pixel-level foreground and background segmentation of the target within the bounding box. The YOLOv5 model is the most stable version, fully integrated with support for instance segmentation. Therefore, this article adopts instance segmentation based on the

YOLOv5 model. Compared with other segmentation models, YOLOv5 changes the resulting mask of instance segmentation from the entire image to only the mask within the detection result box to improve memory utilization efficiency and computational speed. Limiting the segmentation focus to the area of interest can save considerable storage space and accelerate the segmentation process. This method is important for detecting and segmenting plowed areas and provides strong support for improving practical application scenarios.

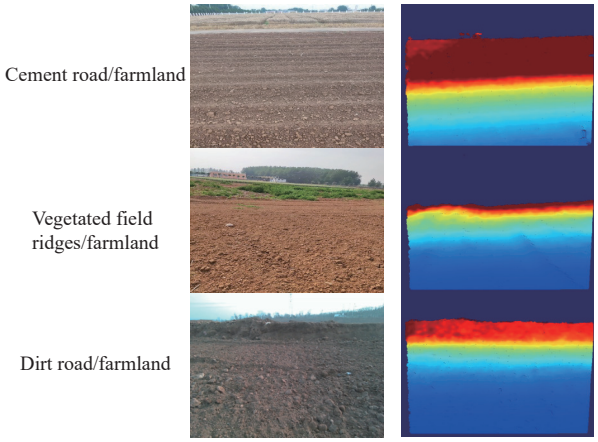


Figure 2 Image samples of different terrain scenes

2.2.1 YOlOv5 model improvement

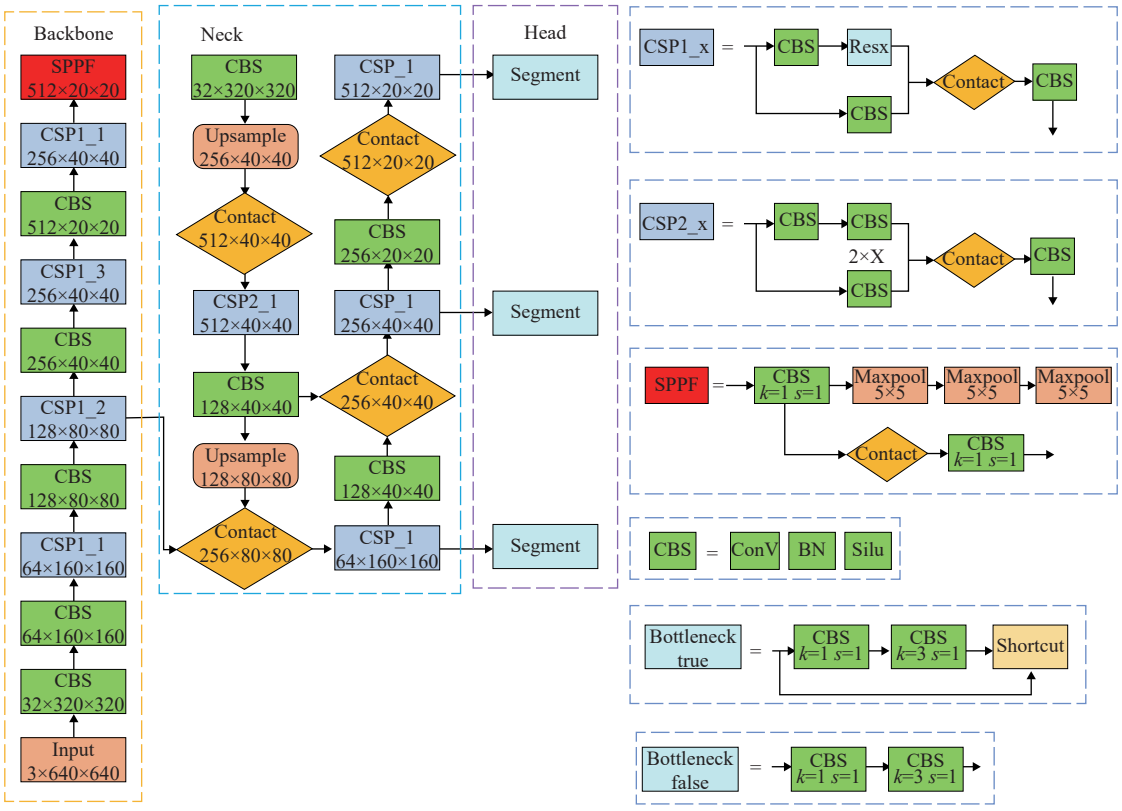
Presently, YOLOv5 (hereinafter referred to as v5) has been updated to a stable version 7.0, integrating segmentation support. To consider the lightweight and real-time segmentation requirements of the model, in this article, we use YOLOv5s-seg (hereafter referred to as 5s-seg) to construct the model. The segmentation improvement is implemented on the basis of the original model, and these differences are reflected mainly in the following five aspects, as listed in Table 1.

Table 1 Differences between YOLOv5s and YOLOv5s-seg

Category	Code entry	Data loading and processing	Network changes	Loss	Evaluation indicators
5s	Root directory	jpg+txt	yaml+head (detect)	Object Detection+Giou	ap
5s-seg	New segment	jpg+txt	yaml+head (detect+segment)	Cross-Entropy+Dice	ap

Note: ap, average precision.

The 5s-seg model structure is shown in Figure 3. The network is divided into three parts: feature extraction, fusion, and prediction output. The CSPDarket53 architecture is used to improve the Darknet53 network, and performance is enhanced through CSP connections[25]. In feature extraction, the CSP1_X structure divides the input into two branches: one undergoes convolution after passing through multiple residual structures, and the other is directly convolved before connecting to the first branch. CSP2_X uses 2×X CBS instead of residual structure, primarily used in the neck networks. The SPPF module replaces the original SPP structure with three 5×5 max pooling layers to reduce complexity and improve speed. The neck structure adopts Feature Pyramid Network (FPN)+Path Aggregation Network (PAN), which extracts and fuses features through up and down sampling to enhance feature representation. The head layer uses segmentation instead of the original detection method and is constructed through yaml files. The segmentation network inherits the detection class and includes segmentation and detection functions.



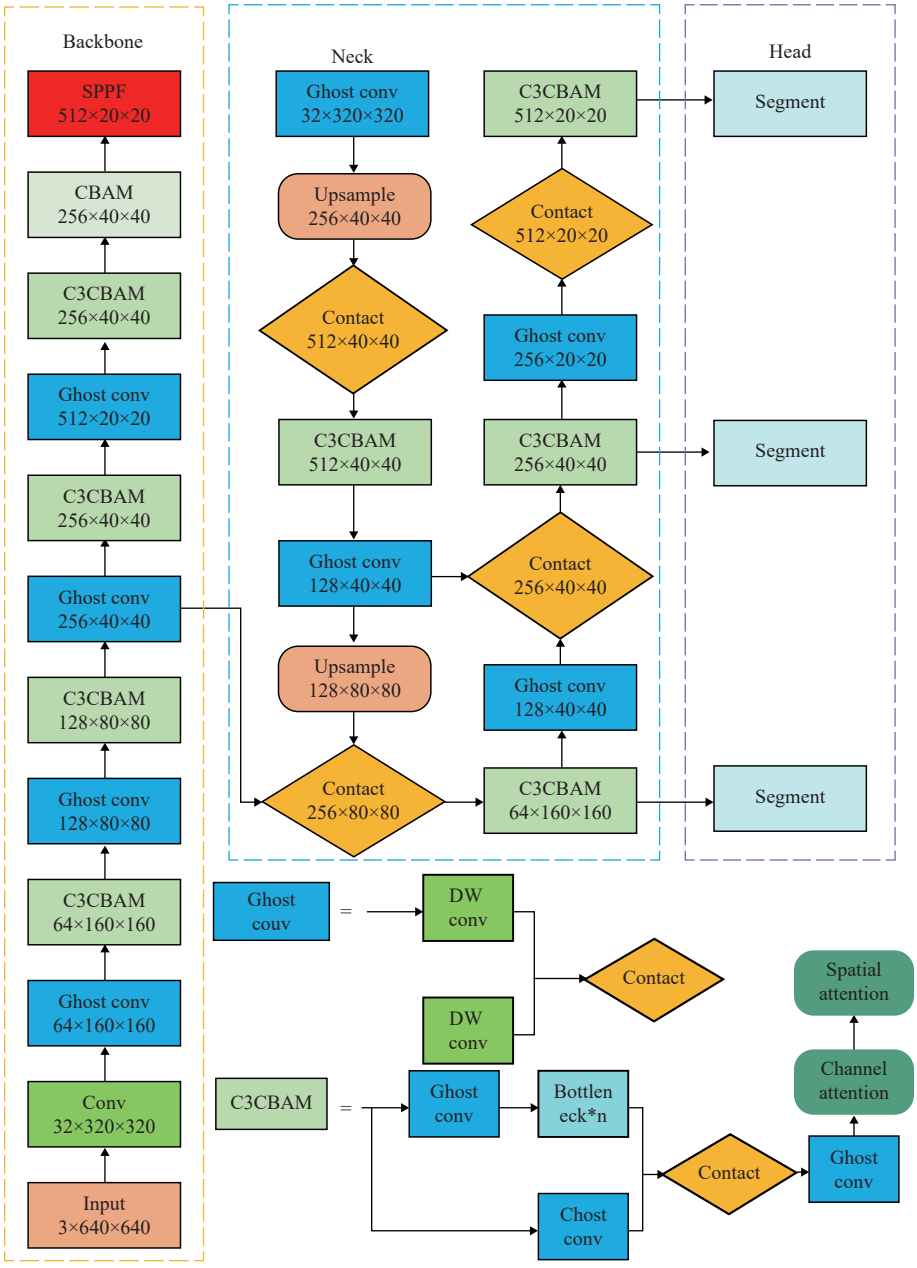
Note: SPPF, Spatial Pyramid Pooling-Fast; CSP, Cross Stage Partial-connections; CBS, Conv+BatchNorm (BN)+SiLu; ResX consists of a CBL and *X* residual components; CBL, Conv+BN+Leaky; *k* is the size of the convolution kernel; *s* is the stride.

Figure 3 YOLOv5s-seg network structure

The improved model structure is shown in Figure 4. By using ghost convolution instead of regular convolution and replacing the CSP module of the network with the Convolutional Block Attention Module (CBAM), the network is lightweight. One CBAM is added after three C3CBAMs in the backbone network to enhance the extraction of input features and to improve the accuracy of network target segmentation.

2.2.2 Improving the C3CBAM

Based on the above network structure and the comprehensive research of this article and to meet the needs of agricultural machinery to detect plow areas during field operations, it is necessary to accurately determine the spatial position information of the land promptly. This study used the CBAM, which combines channel attention (CA) and spatial attention (SA), to adaptively adjust the channel weights of the feature map. In instance segmentation, the model needs to consider target details and global contextual information. The CBAM, as a lightweight and universal module, can seamlessly integrate into any CNN architecture with slightly increased computational complexity and can efficiently capture important features and information about the target, enabling the model to adapt to changes in the shape and size of the target.



Note: CBAM, Convolutional Block Attention Module; C3, Concentrated-Comprehensive Convolution Block; DW conv: Depthwise Conv.

Figure 4 Improving the YOLOv5s-seg network structure

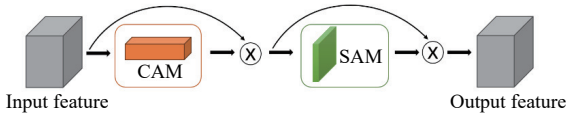
Figure 5 shows the structure of the CBAM, which works given feature map $\boldsymbol{F}$ of height H, width W, and dimension C. First, the feature map goes through the channel attention mechanism to obtain a one-dimensional CA feature map, followed by multiplication of the convolution result by the original map. Second, the output result is employed as an input to go through the spatial attention mechanism to obtain a two-dimensional SA feature map. Last, the output result is multiplied by the original map to obtain the refined feature map. The entire calculation process is represented as follows:

$$\boldsymbol{F} \in R^{C\times H\times W} \tag{1}$$

$$F' = M_C(F) \otimes F \tag{2}$$

$$F'' = M_S(F') \otimes F' \tag{3}$$

where, ⊗ denotes element-by-element multiplication, R is the real numbers, M_c is the channel attention module, M_S is the spatial attention module, F' is the feature map output by the spatial attention module, and F'' is the final output feature map.



Note: ⊗, element-by-element multiplication; CAM, channel attention module; SAM, spatial attention module.

Figure 5 CBAM structure diagram

The initial channel attention mechanism involved global average pooling and max pooling operations on the width and height of the input feature maps, followed by element-wise weighting of the output from the multi-layer perceptron (MLP), and activation through a sigmoid function to obtain the final channel attention feature maps. However, this method tends to overlook the interaction of information within channels when performing global average pooling and max pooling. Therefore, this study eliminated the max pooling operation and used only average pooling to aggregate the spatial information of the feature maps, thus more effectively capturing global features while avoiding the neglect of important information. Additionally, this study drew on the idea of the Efficient Channel Attention (ECA) model, replacing the MLP with one-dimensional convolutions of specific kernel size to enhance the interaction between different channels, and simplifying the model output through a sigmoid function. This improvement reduces computational complexity, and enhances model performance and efficiency, making the network more streamlined and easy to implement, particularly suitable for visual tasks such as image classification and object detection. The improved channel attention module CAM-E is shown in Figure 6, and the computational process is expressed as follows:

$$M_{EC}(F) = \sigma(conv1d_{k=3}(AvgPool(F))) \tag{4}$$

where, M_{EC} is the CAM-E module, Avgpool is the average pooling, σ is the sigmoid activation function, conv1d is 1 for convolution, k denotes the convolution kernel size, and the optimal size of k is determined to be 3 after cross-validation.

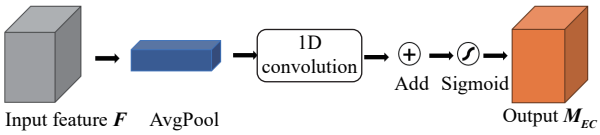


Figure 6 CAM-E structure diagram

The convolution operation of the final CBAM is simplified as follows:

$$\mathrm{CBAM}(F) = F \otimes M_{EC}(F) \otimes M_S(F') \tag{5}$$

2.2.3 Improving the backbone network

The 5s network structure adopts traditional convolutional operations, and traditional feature extraction methods generate numerous parameters and computations, as well as rich and redundant feature maps. Therefore, this study adopted GhostConv instead of traditional convolutional layers.

As shown in Figure 7, the original 3×3 convolutional operation is divided into two smaller convolutional phases, where the first phase with the larger convolutional kernel (k×k) is referred to as the "main convolution", which utilizes a small number of convolutional kernels for feature extraction. The second stage of the convolution kernel (1×1) is referred to as "phantom convolution", where the phantom convolution kernel performs a cheaper linear variation of the feature maps from the previous section; it is concatenated to generate the final feature map. This decomposition operation allows the network to more efficiently learn features and significantly reduces computational and parameter overhead.

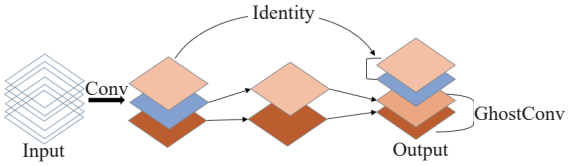


Figure 7 Schematic of conventional and GhostConv

For example, when the input feature map data are height (h); width (w); and channle (c) and the output is n, h'; and w' feature maps, the computation using conventional convolution P_1 is

$$P_1 = nh'w'ckk \tag{6}$$

where, c is the number of input image channels.

The number of parameters using the Chost convolutional network P_2 is

$$P_2 = \frac{n}{s}h'w'ckk + (s-1)\frac{n}{s}h'w'dd \tag{7}$$

where, n=m·s, m is the number of feature maps generated in the first stage, s is the ghost feature map generated in the second stage, and d is the size of the convolution kernel for linear operation, $s \ll c$.

Therefore, the ratio of the number of parameters of the two is

$$\frac{P_1}{P_2} = \frac{nh'w'ckk}{\frac{n}{s}h'w'ckk + (s-1)\frac{n}{s}h'w'dd} \approx \frac{sc}{s+c-1} \approx s \tag{8}$$

According to the above equation, when k and d are equal in size, the number of parameters occupied by the Ghost convolution is 1/s that of the conventional convolution.

2.3 Model training

2.3.1 Experimental environment and training parameter settings

To start training for the improved model, this article uses the deep learning framework PyTorch for model training. The hardware utilized for deep learning was a 12th Gen Intel(R) Core(TM) i7-12650H 2.30 GHz computer, an NVIDIA GeForce RTX3050 central processor and graphics card were employed, the graphics memory was 4 GB, and cuda11.6 was selected to improve the network training speed.

The initial learning rate of this experiment is set to 0.01, the learning rate is adjusted using the cosine annealing restart[26] learning rate mechanism, and the training is performed using warmup_3, after which the learning rate can be achieved by the warmup mechanism, and the network parameters are optimized using the momentum stochastic gradient descent (SGD) algorithm. The input image is set to 640×640, the batch_size is set to 8, and the total number of training rounds is 300.

2.3.2 Model evaluation indicators

To evaluate the performance of the model, this article uses precision, recall, mean average precision, and model size as evaluation metrics. P reflects the accuracy of the model in classifying samples, R represents the ability of the model to obtain

positive samples, mAP is the average accuracy AP of all categories, and AP is calculated by integrating the PR curve with accuracy P as the vertical axis and recall R as the horizontal axis[27], which comprehensively reflects the overall performance of the model in detecting all categories. The specific calculation formula is presented as follows:

$$P = \frac{T_p}{T_p + F_p} \tag{9}$$

$$R = \frac{T_P}{T_P + F_N} \tag{10}$$

where, N denotes the number of categories in the dataset, T_p denotes the number of positive samples correctly predicted as positive samples by the model, F_P denotes the number of negative samples incorrectly predicted as positive samples by the model, and F_N denotes the number of positive samples incorrectly predicted as negative samples by the model.

$$\mathrm{AP} = \int_0^1 \mathrm{P} \cdot R \mathrm{dR} \tag{11}$$

$$\mathrm{mAP} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{AP}_i \tag{12}$$

In object detection tasks, the determination of positive and negative samples is based on the intersection union (IoU), which is the ratio of the area of overlap between the predicted segmentation and the actual target segmentation to the area of union. When the IoU is greater than the threshold, the sample is considered to be positive; when the IoU is less than the threshold, the sample is considered to be negative. When the IoU is set to 0.5, the average accuracy of the YOLOv5 model is represented as $\mathrm{AP}_{0.5}$ (AP at an IoU of 0.5), while the average accuracy of all categories is described as $\mathrm{mAP}_{0.5}$ (mAP at an IoU of 0.5).

This article starts with pixel accuracy (PA) and uses the mean intersection over union (mIoU) as the standard indicator for segmentation models. The PA represents the ratio of the number of pixels correctly classified to the total number of pixels, and the mIoU represents the average of the ratios of the intersection to the union between the predicted and true segmentations for all classes or instances. The expression is

$$\mathrm{PA} = \frac{\mathrm{TP} + \mathrm{TN}}{T} \tag{13}$$

$$\mathrm{mIOU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{14}$$

where, the total number of pixels is T=TP+TN+FP+FN, TP is the number of pixels correctly predicted by the model (true positives); TN is the number of background pixels correctly predicted by the model (true negatives), and in the case of a single category, the background pixels are the pixels other than those of the target category; FP is the number of pixels for which the model incorrectly predicted the background pixels to be in the target category (false-positives); and FN is the number of pixels for which the target category pixels are incorrectly predicted to be in the background pixels (false-negatives).

2.4 Contour processing of plow areas

For the detection and distance measurement of agricultural machinery operation area boundaries in agricultural environments, this study proposed a method combined with depth cameras based on the above training model. The boundary of the segmentation area is divided, and the distance is calculated by combining depth maps. Figure 8 shows the technical flowchart of this study.

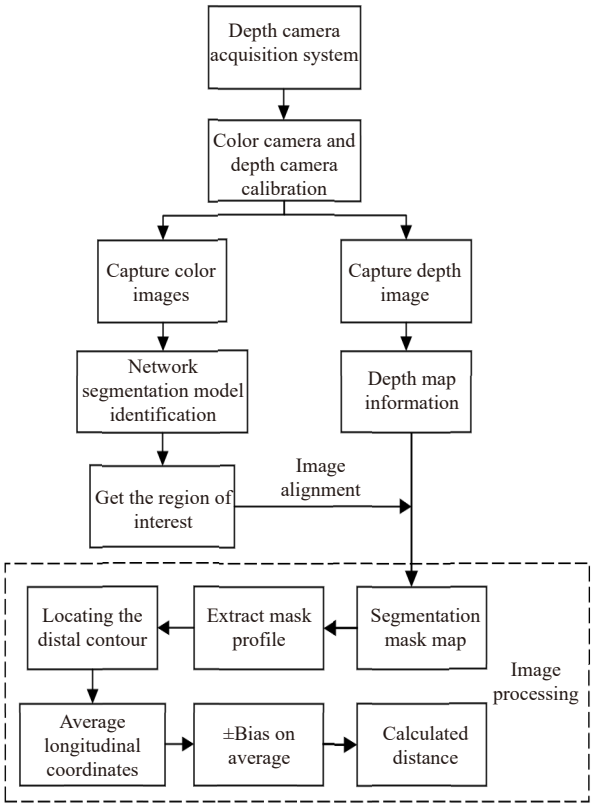


Figure 8 Technical flowchart

2.4.1 Image alignment

Figure 8 shows that the prerequisite for contour processing is image alignment, which requires aligning the RGB image captured by the camera with the color image of the depth image. When the intrinsic and extrinsic matrices of the camera are known, alignment between RGB images and depth images can be achieved through matrix transformation[28]. The conversion relationship between the pixel coordinate system and the world coordinate system is expressed as follows:

$$Z_C \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{T} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} = \boldsymbol{KM} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{15}$$

where, $\boldsymbol{K}$ is the internal reference matrix of the camera; $\boldsymbol{M}$ is the external reference matrix; $\boldsymbol{R}$ is a 3×3 unit orthogonal matrix (referred to as the rotation matrix); $\boldsymbol{T}$ is a 3×1 translation matrix; $f_x = f/\mathrm{d}x$, $f_y = f/\mathrm{d}y$; f is the focal length of the camera; dx and dy denote the actual lengths of the unit pixels in the row direction and column direction, respectively, mm; c_x and c_y are the width and height, respectively, of the image in half, pixels.

The coordinate transformation formula for aligning the color map with the depth image is

$$T_{d2c} = T_{w2c} T_{w2d}^{-1} = \begin{bmatrix} R_{w2c} R_{w2d}^{-1} & t_{w2c} - R_{w2c} R_{w2d}^{-1} t_{w2d} \\ 0 & 1 \end{bmatrix} \tag{16}$$

where, T_{w2c} and T_{w2d} are the external reference matrices for the conversion of the world coordinate system to the color coordinate system and the depth camera coordinate system, respectively; T_{d2c} is the external reference matrix for the conversion of the depth camera coordinate system to the color coordinate system.

Through the above process, the corresponding pixel points in the depth image and color image have the same spatial coordinates,

which achieves the purpose of image alignment.

2.4.2 Image contour processing

After aligning the images, the region masks recognized by the deep learning model are processed, and to increase the accuracy of the subsequent segmentation, a new concept was introduced: the monoclinic (homography) transform, which is used to describe the positional mapping relationship of an object between the world coordinate system and the pixel coordinate system. The uni-responsive transform is applied to the generated mask map, and the uni-responsive matrix is defined by the above equation:

$$H = \lambda \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} [r_1 \ r_2 \ t] = \lambda K [r_1 \ r_2 \ t] \quad (17)$$

where, λ denotes the scale factor.

According to the perspective matrix transformation of the homography matrix, the pixel coordinates of the four corner points of the mask were obtained, and the image from the camera perspective was transformed into an image based on the vehicle width state.

Image masking was performed on the obtained new mask, which is a pixel-level operation that selectively preserves or excludes pixels in the original image using pixel information in the mask to obtain specific regions.

As shown in Figure 9, Contour1 was extracted from the mask map, contour1 was analyzed, the lateral coverage of contour1 was calculated as (w_s, w_e), bias was set to 15 pixels according to the contour characteristics, and the image of the (w_s+bias, w_e+bias) region was extracted from maks1 and constructed as Mask2. The target contour corresponding to the smaller value was calculated for Measures 1 and 2, which are denoted as contour-min. The contour-min was analyzed to calculate the longitudinal coverage of the contour, as well as its transverse coverage, and the masked contour external matrix mask Mask2 was constructed. By averaging the distal contour and adding the bias value, possible errors are effectively offset, and a single contour is affected by characteristics such as noise and shape variations. This operation stabilizes the position of the distal contour, improves its accuracy, and better separates the tilled area from the background by constructing a rectangular box to mark the boundary of the tilled area.

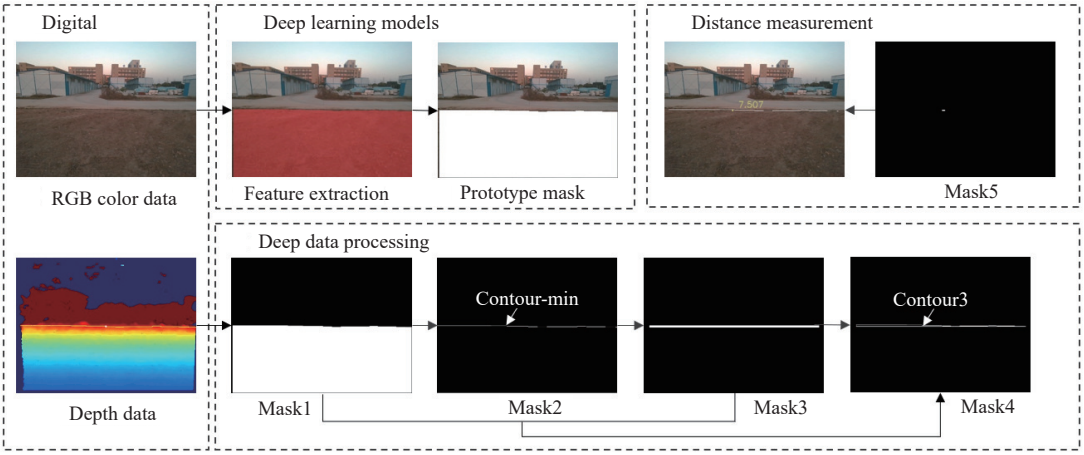


Figure 9 Flowchart of the contour processing of data

After the rectangular box was determined, it intersected with the original Mask1 to determine the final boundary region, and then contour extraction was performed by the Canny operator[29] to obtain contour3, which excludes non-edge pixels and retains only the lines of the candidate edges by suppressing the non-extremely large values of the processed image.

Using the contour center distance measurement method, the contour obtained by the Canny operator is processed to calculate the geometric center of the contour, as shown in Figure 10. First, according to the coordinates of polygon vertices $P_i(x_i, y_i)$, the vector Vec_i pointing to the vertices at the origin is computed $Vec_i = P_i - origin$ and the origin is the origin coordinate. Second, the vector Vec_i is computed and written as SUM. The vector of the geometric center of the contour is SUM/n, which is written as Vec_c. Last, the coordinate position of the geometric center is expressed as Vec_c+origin, which is written as P_c, and the expression of the calculation process is presented as follows:

$$P_c = \frac{1}{n} \sum_{i=1}^{n} (P_i - \text{origin}) + \text{origin} \quad (18)$$

After determining the position of the center point, to reduce errors, with the center point as the origin, 5 pixels each from the top, bottom, left, and right were selected to form a 121-pixel square. The distance from the average value to this area is calculated.

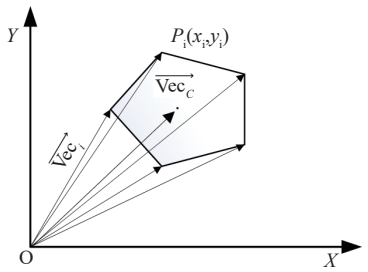


Figure 10 Contour center coordinates

3 Results and analysis

3.1 Deep learning model

3.1.1 Ablation test results

To analyze the effect of each model combination on the performance of the improved model, this study designs ablation experiments using uniform hyperparameters and training 300 epochs; the results of the experiments are listed in Table 2. In the model listed in Table 2, GC denotes replacing the traditional convolution with a phantom convolution structure, BC denotes replacing the backbone layer C3 structure with C3CBAM, AC denotes replacing all the C3 structures with C3CBAM, and GM denotes adding all the modules to obtain the final model, which is the improved model proposed in this study.

Table 2 Farmland boundary recognition ablation test

Model	*P*	$mAP_{0.50-0.95}$	*R*	Weights/MB	Speed/fps
5s-seg	0.939	0.753	0.878	14.8	45
GC-5s-seg	0.961	0.889	0.97	12.4	90
BC-5s-seg	0.937	0.643	0.97	13.7	52
AC-5s-seg	0.952	0.848	0.954	12.7	62
GM-5s-seg	0.991	0.923	0.976	9.82	98

As listed in Table 2, by applying the phantom structure to the backbone and neck parts of the YOLOv5s-seg and adjusting the network width, the GC-5s-seg model was developed. This modification reduced the model's parameter count without losing detection accuracy, improved precision by 3 percentage points, and increased the mean average precision by 13 percentage points while doubling the detection speed. To further enhance the model's precision, the CBAM was applied to optimize the baseline model and was integrated at specific locations, resulting in 3 different models: BC-5s-seg, AC-5s-seg, and GM-5s-seg. Among these, the GM-5s-seg model achieved a precision (*P*) of 0.991, which is an improvement of 6 and 4 percentage points over the first two models respectively. It has a parameter size of 9.82 MB and an average detection speed of 0.011 s per image, which is twice as fast as the 5-seg model. The integrated inference results of GM-5s-seg on $mAP_{0.50-0.95}$ showed a performance increase of 17 percentage points over the original model.

This article uses the $mAP_{0.50-0.95}$ as the measurement standard for the farmland boundary detection model. As shown in Figure 11, the improved model has achieved a good fitting effect.

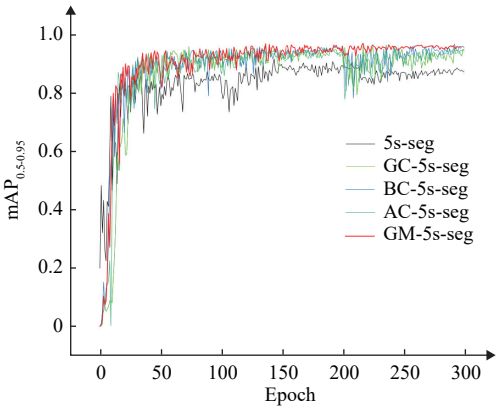


Figure 11 Comparison of $mAP_{0.50-0.95}$ obtained in ablation experiments

3.1.2 Comparison between different attention mechanisms

To further verify the better performance of the model employed in this article, based on the 5s-seg model, different attention mechanism modules were combined to select the mainstream modules ECA and CA to replace the position of the CBAM in the network. In addition, the GhostNet network structure was introduced to replace BottleneckCSP with a Ghost bottleneck for comparison. The experimental results are listed in Table 3.

Table 3 Comparison of different attention models

Model	*P*	$mAP_{0.5-0.95}$	*R*	Weights/MB	Speed/fps
5s-seg	0.939	0.753	0.879	14.8	45
GM-5s-seg	0.991	0.923	0.976	9.82	98
ECA-5s-seg	0.96	0.894	0.967	10.38	76
CA-5s-seg	0.937	0.643	0.96	13.6	62
GN-5s-seg	0.972	0.682	0.946	12.7	76

As listed in Table 3 and shown in Figure 12, compared with the original YOLOv5-seg model, the model with the CA and GN modules yields an increase in the precision rate *P* and recall rate *R*, but the $mAP_{0.50-0.95}$ significantly decreases. The evaluation metrics with the ECA module have all improved, showing significant enhancements in mean Average Precision compared to CA and GN. Compared to the other three attention modules, models incorporating the GM module have achieved excellent results in both accuracy and recall, while also having the smallest model parameters and computational requirements, demonstrating superior overall performance. These experiments showed that adding the GM module is superior to the other attention modules on the self-constructed dataset of this study.

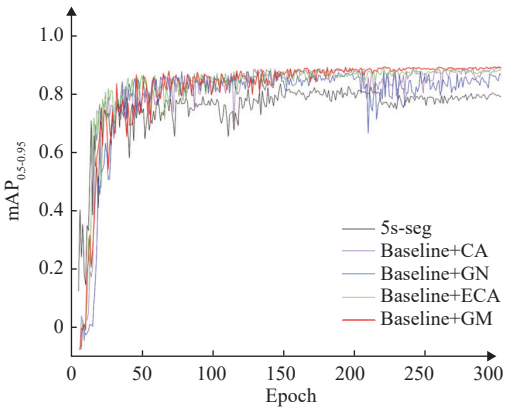


Figure 12 Comparison of the $mAP_{0.50-0.95}$ obtained in fusion experiments with different attention

3.1.3 Comparison of different algorithms

To verify the superior segmentation performance of the improved model in this article, the common instance segmentation algorithms Mask R-CNN[30] and YOLACT[31] are selected to conduct comparative experiments with the same experimental environment and configuration parameters; the results are listed in Table 4. An examination of the results in Table 4 clearly shows that the model proposed in this paper significantly improves the PA, average intersection and merger ratio (mIoU), and average accuracy. Additionally, the model has a substantial advantage over the other models in terms of weight and detection speed, which indicates its effectiveness.

Table 4 Comparison of different algorithms

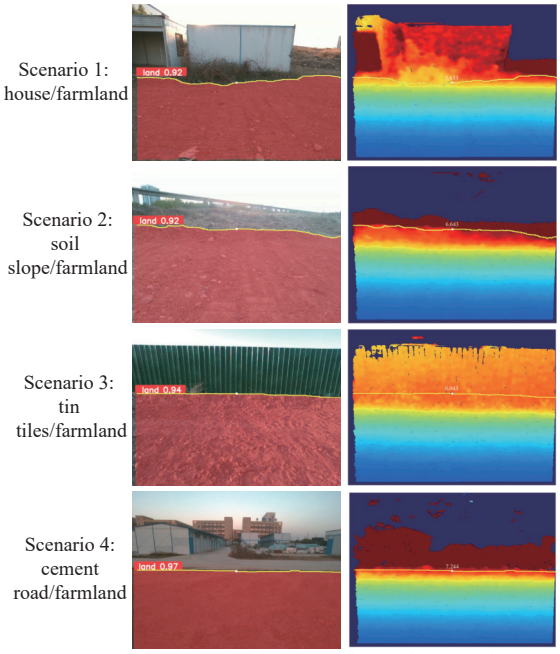
Model	PA	mIoU	$mAP_{0.5-0.95}$	Weights/MB	Speed/fps
GM-5s-seg	0.970	0.952	0.923	9.82	98
Mask R-CNN	0.910	0.801	0.875	249.86	5
YOLACT	0.922	0.833	0.911	194.44	19

3.2 Measurement of boundary distance in plow areas

To ensure the generality of the boundary object segmentation algorithm in this article, detection and distance measurement experiments are conducted for different ground head boundaries. The following are recognition, detection, and depth information distance measurement maps for several types of land boundary conditions.

As shown in Figure 13, it can be observed that in several different scenarios, the improved farmland boundary recognition and ranging model can meet practical needs. Specifically, in Figure 13, the model performs well in recognition and can accurately and effectively partition the plow area from the background, even if the scene boundary is a soil slope, the improved

model of this study can still accurately identify it, which further verifies the robustness and applicability of the model. For the depth map, the depth information was successfully removed from the background by setting an appropriate threshold, thus avoiding errors in subsequent distance measurements.



a. Color recognition map b. Depth recognition map

Note: The yellow line in the figure represents the boundary contour line, and the white point represents the center point of the contour.

Figure 13 Detection and ranging samples in different terrain scenarios

In order to verify the representativeness of the contour average boundary distance, two boundary measurement distance indicators are set: the contour average boundary distance and the center boundary distance, and the center boundary distance is the distance at the intersection of the image midline, and the boundary. Combined with the installation height of the camera in Figure 1 and the measured depth distance value, the distance value in the horizontal direction was calculated by the collinear theorem. The results are listed in Table 5.

Table 5 Farmland boundary distance detection results

Image	Center distance/m	Average contour distance/m	True distance/m	Error/m
Scenario 1	5.463	5.556	5.5	0.056
Scenario 2	5.882	5.957	6.0	0.043
	6.602	6.606	6.5	0.106
Scenario 3	6.068	6.043	6.0	0.043
Scenario 4	6.640	6.542	6.5	0.042
	7.773	7.260	7.5	0.260

The driving speed of the agricultural machine is 0.8-1.5 m/s during the operation, considering the control time of the machine and the resolution of the sensor, the machine is able to recognize the boundary of the ridge and try to measure the distance at a distance (8.0 m), with an allowable distance error of 0.400 m; the machine is able to accurately recognize the boundary of the farmland to assist the turn control at a distance (5.0 m), with an allowable distance error of 0.100 m. It can be seen in Table 5, that with the increase of test distance, the error will gradually increase, and the error is 0.056 m at the near (5.5 m) and 0.260 m at 7.5 m. Table 5 shows that with the increase of the test distance, the error will gradually increase, and the error is 0.056 m at the near place (5.5 m), and 0.260 m at 7.5 m. The average detection time for each image is 29 ms, which verifies the effectiveness of the algorithm.

4 Conclusions

1) This study proposed a farmland plow area recognition algorithm based on object segmentation. By fully utilizing the powerful feature learning ability of deep learning, this algorithm can capture the color, texture, and shape features of plowed areas, significantly improving recognition accuracy and overcoming the limitations of traditional manual feature extraction.

2) The model parameters and computational cost were reduced by improving the network model structure by replacing the traditional convolution with ghost convolution, replacing the original CSP structure with the C3CBAM, and adding the CBAM to the detection header to enhance the feature extraction and detection capability of the model.

3) A target-ranging method based on RGB-D data fusion was proposed. By improving the object detection network, camera calibration, and image registration, the depth image was successfully matched with the RGB image coordinates to obtain the depth information of the object to be measured, and this information was subsequently converted to the relative distance of the target. The experiment has proven the effectiveness of this method, with ranging errors controlled at the centimeter level. Compared to single vision-based ranging methods, the ranging accuracy has been significantly improved.

4) The distance detection accuracy of this algorithm for the boundary of the farmland plow area was closely related to the detection distance. The experimental results reveal that the distance detection accuracy is 0.260 m at a distance of 7.5 m from the boundary of the farmland, that the average distance detection error is approximately 0.056 m at a distance of 5.5 m from the ridge, and that the detection time of each image is 29 ms, which satisfies the requirements of distance detection of the boundary during the autonomous operation of the farm machine with a traveling speed less than 1.5 m/s.

In conclusion, the proposed target-ranging method based on RGB-D data fusion fully integrates depth and image information. By optimizing the network structure, camera calibration, and image alignment, high-accuracy target ranging is achieved. The method is expected to be widely employed in practical applications and to provide a high-precision solution for ranging tasks.

Acknowledgements

This work was financially supported by the National Key Research and Development Program (NKRDP) projects (Grant No. 2023YFD2001100), Major Science and Technology Programs in Henan Province (Grant No. 221100110800), and Major Science and Technology Special Project of Henan Province (Longmen Laboratory First-class Project) (Grant No. 231100220200).

[References]

[1] Luo X W, Liao J, Zang Y, Qu Y G, Wang P. Developing from mechanized to smart agricultural production in China. Strategic Studi of CAE, 2022; 24(1): 46–54.

[2] Liu C L, Lin H Z, Li Y M, Gong L, Miao Z H. Analysis on status and development trend of intelligent control technology for agricultural equipment. Transactions of the CSAM, 2020; 51(1): 1–18.

[3] Zhang M, Ji Y H, Li S C, Cao R Y, Xu H Z, Zhang Z Q. Research progress

of agricultural machinery navigation technology. Transactions of the CSAM, 2020; 51(4): 1–18. (in Chinese)

[4] Zhou J, Ji C Y, Liu C L. Visual navigation system of agricultural wheeled-mobile robot. Transactions of the CSAM, 2005; 36(3): 90–94. (in Chinese)

[5] Chen B Q, Wu Z H, Li H Y, Wang J. Research of machine vision technology in agricultural application: Today and the future. Science & Technology Review, 2018; 36(11): 54–65. (in Chinese)

[6] Xie Y S, Chen K, Li W T, Zhang Y, Mo J Q. An improved adaptive threshold RANSAC method for medium tillage crop rows detection. In: 2021 6th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an: IEEE, 2021; pp.1282–1286.

[7] Ahmadi A, Nardi L, Chebrolu N, Stachniss C. Visual servoing-based navigation for monitoring row-crop fields. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), Paris: IEEE, 2020; 4920–4926.

[8] Zheng H, Wang Q, Ji J M. Navigation line extraction based on image processing for weeding robot. In: IECON 2022 - 48th Annual Conference of the IEEE Industrial Electronics Society, Brussels: IEEE, 2022; pp.1–5.

[9] Xue J L, Zhang S S. Navigation of an agricultural robot based on laser radar. Transactions of the CSAM, 2014; 45(9): 55–60. (in Chinese)

[10] Yun C, Kim H-J, Jeon C-W, Kim J-H. Stereovision-based guidance line detection method for auto-guidance system on furrow irrigated fields. IFAC-PapersOnline, 2018; 51(17): 157–161.

[11] Zhao T, Noboru N, Yang L L, Kazunobu I, Chen J. Development of uncut crop edge detection system based on laser rangefinder for combine harvesters. Int J Agric & Biol Eng, 2016; 9(2): 21–28.

[12] Wang Q, Liu H, Yang P S, Meng Z J. Detection method of headland boundary line based on machine vision. Transactions of the CSAM, 2020; 51(5): 18–27. (in Chinese)

[13] Cai D Q, Li Y M, Tan C J, Liu C L. Detection method of boundary of paddy fields using support vector machine. Transactions of the CSAM, 2019; 50(6): 22–27, 109. (in Chinese)

[14] Zhu S X, Xie Z H, Xu H, Pan J D, Ren S G. An approach to field path detection based on computer vision. Jiangsu Journal of Agriculture, 2013; 29(6): 1291–1296.

[15] Ollis M, Stentz A. Vision-based perception for an automated harvester. In: Proceedings of the 1997 IEEE/RSJ International Conference on Intelligent Robot and Systems: Innovative Robotics for Real-world Applications. IROS'97, Grenoble: IEEE, 1997; 3: 1838–1844.

[16] Åstrand B, Baerveldt A J. A vision based row-following system for agricultural field machinery. Mechatronics, 2005; 15(2): 251–269.

[17] Wei X H, Zhang M, Liu Q S, Li L. Extraction of crop height and cut-edge information based on binocular vision. Transactions of the CSAM, 2022; 53(3): 225–233. (in Chinese)

[18] Hong Z J, Li Y M, Lin H Z, Gong L, Liu C L. Field boundary distance detection method in early stage of planting based on binocular vision. Transactions of the CSAM, 2022; 53(5): 27–33, 56. (in Chinese)

[19] Chen B B, Cao Q X. Grain height and boundary detection based on based on laser and IMU. Mechatronics, 2020; 26(Z1): 46–51.

[20] Wei L G, Zhang X C, Wang F Z, Che Y, Sun X W, Wang Z W. Design and experiment of harvest boundary online recognition system for rice and wheat combine harvester based on laser detection. Transactions of the CSAE, 2017; 33(Z1): 30–35. (in Chinese)

[21] Zhang Y, Pan S Q, Xie Y S, Chen K, Mo J Q. Detection of ridge in front of vehicle based on fusion of camera and millimeter wave radar. Transactions of the CSAE, 2021; 37(15): 169–178. (in Chinese)

[22] Persello C, Tolpekin V A, Bergado J R, de By R A. Delineation of agricultural fields in smallholder farms from satellite images using fully convolutional networks and combinatorial grouping. Remote Sensing of Environment, 2019; 231: 111253.

[23] Qiao Y J, Liu H, Meng Z J, Chen J P, Ma L Y. Method for the automatic recognition of cropland headland images based on deep learning. Int J Agric & Biol Eng, 2023; 16(2): 216–224.

[24] Li X Y, Cai C, Zhang R F, Ju L, He J R. Deep cascaded convolutional models for cattle pose estimation. Computers and Electronics in Agriculture, 2019; 164: 104885.

[25] Wang C Y, Liao H Y M, Wu Y H, Chen P Y, Hsieh J W, Yeh I-H. CSPNet: A new backbone that can enhance learning capability of CNN. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle: 2020; pp.1571–1580.

[26] Zhang Z, He T, Zhang H, Zhang Z Y, Xie J Y, Li M. Bag of freebies for training object detection neural networks. arXiv prepress, 2019.

[27] Wang G, Fang H B, Wang D Z, Yan J W, Xie B L. Ceramic tile surface defect detection based on deep learning. Ceramics International, 2022; 48(8): 11085–11093.

[28] Syed T N, Liu J Z, Zhou X, Zhao S Y, Yuan Y, Ahmed M S H, Ali L I. Seedling-lump integrated non-destructive monitoring for automatic transplanting with Intel RealSense depth camera. Artificial Intelligence in Agriculture, 2019; 3: 18–32.

[29] Xin Y X, Wang C Y. An image edge detection method based on Canny operator. Information & Computer, 2017; 18: 37–38, 41.

[30] He K M, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. In: 2017 IEEE International Conference on Computer Visio (ICCV), Venice: IEEE, 2017; pp.2980–2988.

[31] Bolya D, Zhou C, Xiao F Y, Lee Y J. Yolact: Real-time Instance Segmentation. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul: IEEE, 2019; pp.9157–9166.