Improve Multi-Object Tracking Using YOLOv5 and Deep-SORT for Chinese Yam Counting Model

Tianzhi Cao, Hiroshi Okamoto

Hokkaido University, Hokkaido 001-0021, Japan.

Abstract: Accurate counting information is important for crop yield and quality. Obtaining yield estimation of Chinese Yam is critical to improving productivity. Referring to the method of counting pedestrian flow in surveillance video and its accuracy, a method for counting the number of Chinese Yams harvested in the field is proposed by improving YOLOv5s detection combined with Deep-SORT tracking. In order to improve the recognition effect of the detector, the attention module CBAM is fused with the Neck part of the YOLOv5s network to improve the feature extraction ability of the network; CIoU Loss is used instead of GIoU Loss as the target bounding box regression loss function to speed up the bounding box regression rate while improving positioning Accuracy; use DIoU-NMS to replace NMS to improve the missed detection problem when the target is crowded.Adjust the structure of the Deep SORT appearance feature extraction network and retrain on the yam re-identification dataset to reduce the identity switching caused by target occlusion. Connect the improved YOLOv5s detector and Deep SORT, and set a virtual detection line in the video to count the number of yams.The experimental results show that the number of yams can be counted more accurately. Compared with the original algorithm, the improved YOLOv5s has an average accuracy rate of 2.1 percentage points. Combined with Deep-SORT tracking, the counting statistics accuracy rate reaches 92.7%.

Keywords: YOLOv5s; Deep-SORT; Attention Mechanism; Number Statistics; Yam Counting.

1. Introduction

In time, and low precision.In the current smart agriculture, the use of camera video to count crops is a research hotpot.We can learn from the method of using surveillance video to count pedestrian flow in intelligent transportation system,and improve algorithm and system.

The counting of Chinese Yams based on camera video usually includes two parts: Chinese Yam target detection and tracking.The traditional target detection algorithm use manual construction of target features, and then use classification algorithms to classify and judge whether the target exists.Typical algorithms such as Haar+AdaBoost`these algorithms need to perform sliding window operations in the image, the detection efficiency is low, and the resource consumption is large.Moreover, the artificially designed features have low robustness and poor generalization effect, which may easily lead to false detection and missed detection of Chinese Yam.With the continuous development of machine learning and GPU parallel computing technology, target detection has gradually changed from traditional method to the method based on deep learning, which are mainly divided into One-Stage structure and Two-Stage structure.Two-Stage algorithm first generates candidate areas during detection, and then classifies and calibrates based on the candidate areas. The accuracy is relatively high, and the representative models include the R-CNN series. The One-Stage algorithm does not need to generate candidate regions during detection, and directly regresses the target category and boundary. The detection speed is fast which representative models include SSD, Retina Net, and YOLO series.The Chinese Yam detection method based on deep learning can learn the characteristics of Chinese Yam from data, and has the advantages of fast detection speed and high accuracy.This article uses YOLOv5 in the YOLO model, released by Ultralytics, and divided into s, m, l, and x according to the weight of the model.Among them, the YOLOv5s model has the least amount of parameters (Params), the lowest amount of floating

point operations (FLOPs), but the fastest running speed, reaching an inference speed of 2 ms per image, and the mAP on the COCO verification set reaches 53.7%.

According to different initialization methods, target tracking algorithms can be divided into detection-based tracking(DBT) and detection-free tracking(DFT).DBT needs a detector to detect the target in each frame of image in advance, and then track the detected target, so the tracking effect depends on the detection effect;However, DFT needs to manually mark out the tracked target in the initial frame, the flexibility of the algorithm is relatively low, and it cannot track different targets that appear in subsequent frames.Object tracking can also be divided into online tracking and offline tracking according to different video frame processing methods.When online tracking processes each frame, the target is tracked according to the information in the current frame and the previous frame, and the tracking results of the previous frame cannot be modified according to the information of the current frame; offline tracking can use the data before and after the current frame to obtain the global optimal solution, but not Suitable for real-time applications in realistic scenarios. In the yam tracking task of camera video, considering real-time and flexibility, detection-based online tracking is the closest approach to practical application.

Based on the above work, the Chinese Yam detection based on deep learning has the advantages of fast speed and high accuracy, and it can show good detection results no matter in complex environmental background or on edge devices with low computing power.This paper proposes to use the lightweight target detection model YOLOv5s as a detector, combined with the Chinese Yam quantity statistics method of Deep-SORT target tracking, to achieve end-to-end detection and statistics.For the problem of yam recognition rate, the attention mechanism is integrated with the detection network to strengthen the ability of the model to extract features, so that the model pays more attention to the detected target itself. Use CIoU Loss to replace the original detection frame regression loss function to improve the problem of low positioning accuracy and slow regression speed of the target detection frame during training. Use DIoU-NMS to replace the original NMS to improve the missed detection caused by Chinese Yam occlusion or overlap. Input size adjustment to Deep-SORT feature extraction network and retraining on the Chinese Yam re-identification data set. Connect detector and tracker, tune parameter and test application on camera video.

The difficulty of counting the number of Chinese Yam in camera video lies in the blurred soil background, occlusion of Chinese Yams, and occasional scale change of targets in the video.These factors will lead to missed or false detection during Chinese Yam detection, or identity switching (IDs), affecting the final statistical results.In view of these difficulties, this paper uses the target detection algorithm based on deep learning to detect Chinese yams, combine the detection results with the online tracking DBT algorithm to track the targets, and finally uses the virtual vertical detection line in the video to complete the statistics of the number of Chinese Yams.

2.Materials and Methods

2.1 YOLOv5s target detection

In the DBT algorithm, the effect of the detector seriously affects the result of target tracking, and the speed of the detector and the size of the model are also the key to complete real-time target tracking.Since most of the farmland sites are embedded devices with low computing power, it is impossible to deploy large-scale detection models. In order to reduce computing costs and enhance practicability,in this paper, YOLOv5s, the smallest model in the YOLOv5 series, is selected as the basic model for Chinese Yam detection.

2.1.1YOLOv5s

The structure of YOLOv5s is mainly divided into four parts, Input,Backbone, Neck, Head, as shown in Figure 1.The Input mainly includes data preprocessing, including Mosaic data enhancement, adaptive image filling, and in order to apply to different data sets, YOLOv5s integrates adaptive anchor box calculations at the Input, so that when the data set is replaced, it automatically sets the initial anchor box size.The Backbone network extracts different levels of features from the image through deep convolution operations. It mainly uses the BottleneckCSP(bottleneck cross-stage partial) and the SPP(spatial pyramid pooling). The purpose of the former is to reduce the amount of calculation and improve the reasoning speed. The latter realizes the feature extraction of different scales on the

same feature map helps to improve the detection accuracy.The Neck layer includes a FPN(feature pyramid network) and a PAN(path aggregation network). FPN transmits semantic information from top to bottom in the network, and PAN transmits positioning information from bottom to top, and integrates information from different network layers in Backbone to further improve detection capability. As the final detection part, the Head output mainly predicts targets of different sizes on feature maps of different sizes.

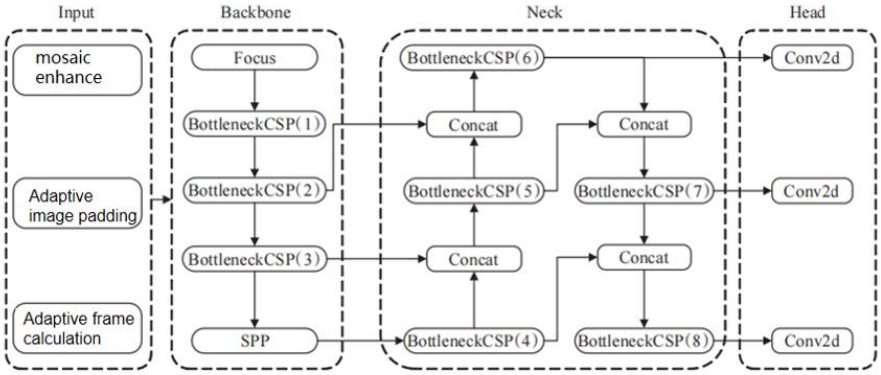

Fig.1 Structure of YOLOv5s

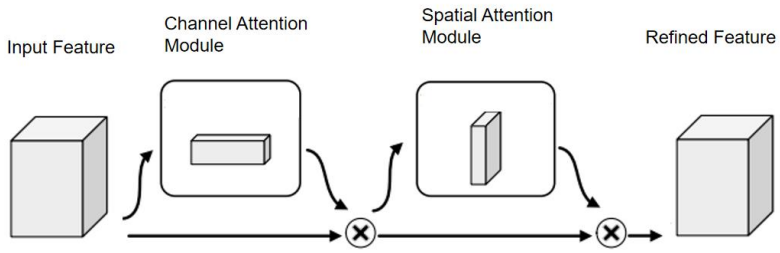

Fig.2 Structure of CBAM

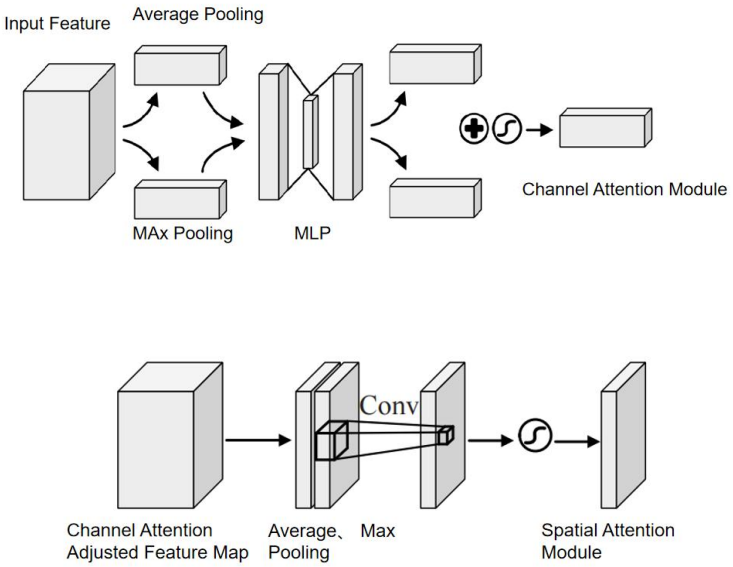

Fig.3 Structure of CAM and SAM

2.1.2YOLOv5s Fusion Attention Mechanism

In the field of computer vision, the effectiveness of the attention mechanism has been proven, and it has been widely used in classification, detection, and segmentation tasks. In the CNN network, the attention mechanism is used on the feature map to obtain the attention information available in the feature map, mainly including spatial attention and general attention information. The CBAM (convolution block attention module) pays attention to the spatial and channel information at the same time, and reconstructs the feature map in the middle of the network through two sub-modules, CAM (channel attention module) and SAM (spatial attention module), emphasizing the importance of features, suppressing general features, to achieve the purpose of improving the target detection effect, its structure is shown in Figure 2.

For the 3D feature map $F \in R^{C \times H \times W}$ of a certain layer in the CNN network, CBAM sequentially infers the 1D channel attention feature map Mc and the 2D spatial attention feature map Ms from F, and performs element-by-element correlation respectively. Multiply, and finally get the output feature map of the same dimension as F, as shown in formula (1).Where F represents the feature map of a network layer in the network, Mc(F) represents the channel attention reconstruction of F by CAM, Ms(F′) represents the spatial attention reconstruction of F′ by SAM on the result of channel attention reconstruction, $\otimes$ means element-wise multiplication.

$$\begin{cases} F' = Mc(F) \otimes F \\ F'' = Ms(F') \otimes F' \end{cases} \quad (1)$$

The structure of CAM and SAM is shown in Fig.3.Figure 3(a) shows the calculation process in CAM. Each channel of the input feature map F undergoes maximum pooling and average pooling at the same time, and the resulting intermediate vector passes through a multi-layer perceptron (MLP) , in order to reduce the amount of calculation, MLP only designs one hidden layer, and finally adds the feature vector output by MLP element-wise and performs Sigmoid activation operation to obtain the channel attention Mc. Figure 3(b) shows the calculation process of SAM. The feature map F' activated by Mc is subjected to maximum pooling and average pooling along the channel direction, and convolution operation is performed on the obtained intermediate vector, after the convolution result is activated by Sigmoid, the spatial attention Ms is obtained.

One of the most important functions of the attention mechanism is to reconstruct the feature map, highlighting the important information in the feature map while suppressing the general information. The most critical part of extracting features in the YOLOv5s network is the backbone. Therefore, this article combines CBAM after Backbone, before Neck network feature fusion, the reason for doing this between the two is that YOLOv5s has completed feature extraction in Backbone, and after Neck feature fusion, it predicts output on different feature maps, and CBAM performs attention here. The structure can play the role of linking the past and the future. The specific structure is shown in Figure 4.

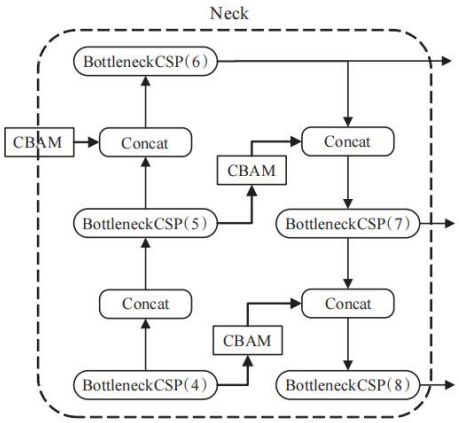

Fig.4 Structure of Neck integrating CBAM

2.1.3Loss function improvement

YOLOv5s uses GIoU Loss as the bounding box regression loss function to judge the distance between the predicted box(PB)and the ground truth(GT),such as formula (2):

$$\begin{cases} GIoU = IoU - \frac{A^c - U}{A^c} \\ L_{GIoU} = 1 - GIoU \end{cases} \quad (2)$$

In the formula, IoU represents the intersection ratio of PB and GT, A^c represents the area of the smallest rectangular box that includes PB and GT at the same time, U represents the union of PB and GT, and L_{GIoU} is the GIoU loss. The advantage of GIoU Loss is scale invariance, that is, the similarity of PB and GT has nothing to do with their spatial scale. The problem with GIoU Loss is that when PB or GT is completely surrounded by the opponent, GIoU Loss completely degenerates into IoU Loss. Because it relies heavily on IoU items, the convergence speed in actual training is too slow, and the accuracy of the predicted bounding box is relatively low.For these problems, CIoU Loss also considers the overlapping area of PB and GT, the distance between the center points, and the aspect ratio, such as formula (3):

$$\begin{cases} CIoU = IoU - \frac{\rho^2(b,b^{gt})}{c^2} - \alpha\gamma \\ L_{GIoU} = 1 - CIoU \end{cases} \quad (3)$$

In the formula, b and b^{gt} represent the center points of PB and GT, ρ^2(...) represents the Euclidean distance, c represents the shortest diagonal length of the smallest bounding box of PB and GT, α represents a positive balance parameter, γ represents the consistency of the aspect ratio of PB and GT. α and γ are defined as formula (4):

$$\begin{cases} \gamma = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \\ \alpha = \frac{\gamma}{(1-IoU)+\gamma} \end{cases} \quad (4)$$

which w^{gt} and h^{gt} represent the width and height of GT and PB, respectively.

Compared with the GIoU Loss and CIoU Loss used in YOLOv5s, the penalty items of PB, GT center distance and aspect ratio are added to the loss item, so that the network can ensure faster convergence of the prediction frame during training and get a higher return. For positioning accuracy, this paper uses CIoU Loss as the loss function of the Chinese Yam detection network.

2.1.4NMS non-maximum suppression improvement

In the prediction stage, NMS is usually used to remove redundant detection frames. The criterion for judging is the IoU ratio between a certain detection frame and the detection frame with the highest predicted score. When the IoU is greater than the set threshold, the predicted detection frame will be remove. In general scenarios, this method is effective, but in dense target environments, due to mutual occlusion between targets, the detection frames of different targets are very close, and the overlapping area is large, so they will be removed by NMS by mistake, resulting in target detection failure. In the camera video, the yam target is concentrated in the middle of the field in the image, which is a relatively dense and prone to occlusion scene. This paper uses DIoU as the NMS criterion to improve this problem.

DIoU considers the distance between the center points of two bounding boxes on the basis of IoU, as in formula (5):

$$DIoU = IoU - \frac{\rho^2(b,b^{gt})}{c^2} \quad (5)$$

As same as the formula(3), b and b^{gt} represent the center points of PB and GT, ρ^2(...) represents the Euclidean distance, c represents the shortest diagonal length of the smallest bounding box of PB and GT.

DIoU-NMS is defined as formula (6):

$$s_i = \begin{cases} s_i, DIoU(M, B_i) < \varepsilon \\ 0, DIoU(M, B_i) \geq \varepsilon \end{cases}$$

M represents a prediction frame with the highest prediction score, $B_i$represents the prediction frame that needs to be removed, s_i represents the classification score, and ε represents the threshold of NMS.DIoU-NMS considers the IoU while judging the distance between the center points of the two bounding boxes M and B. When the distance is far away, the prediction box will not be removed, but another target is detected, which helps to solve the mutual occlusion of the targets. The problem of missed detection in the case. This paper uses DIoU-NMS to replace the original NMS.

2.2Deep-SORT object tracking

Multi-target online tracking algorithm SORT (simple online and real-time tracking) uses Kalman filter and Hungarian matching, and uses the IoU between tracking results and detection results as a cost matrix to implement a simple, efficient and practical tracking paradigm.However, the defect of the SORT algorithm is that the association metric used is only effective when the uncertainty of the state estimation is low, so a large number of identity switching phenomena will occur during the execution of the algorithm, and tracking failures are relatively prone to occur.In order to improve this problem, Deep-SORT combines the target's motion information and appearance information as a correlation measure to improve the problem of target tracking failure.

2.2.1SORT algorithm process

The core of the sort algorithm is the Kalman filter algorithm and the Hungarian algorithm.The main function of the Kalman filter algorithm is to predict the motion variables of the next moment by the current series of motion variables, but the first detection result is used to initialize the motion variables of the Kalman filter. Simply speaking, the Hungarian algorithm is to solve the allocation problem, that is, to allocate a group of detection frames and the frames predicted by Kalman, so that the frames predicted by Kalman can find the detection frame that best matches itself, so as to achieve the effect of tracking.

The workflow of the whole algorithm is as follows(fig):

(1) Create the corresponding Tracks from the detected results of the first frame. Initialize the motion variable of the Kalman filter, and predict its corresponding frame through the Kalman filter.

(2) Perform IOU matching on the frame of the frame target detection and the frame predicted by Tracks in the previous frame, and then calculate the cost matrix through the IOU matching result (the calculation method is 1-IOU).

(3) Use all the cost matrices obtained in (2) as the input of the Hungarian algorithm to obtain linear matching results. At this time, we get three results. The first is Unmatched Track, we directly delete the mismatched Tracks; the second is Unmatched Detection, we initialize such Detection as a new Track; the third is the successful pairing of the detection frame and the predicted frame, which shows that We successfully tracked the previous frame and the next frame, and updated the corresponding Tracks variable through the Kalman filter for the corresponding Detection.

(4) Repeat steps (2)-(3) until the end of the video frame.

2.2.2Deep-SORT algorithm process

The workflow of the whole algorithm is as follows:

(1) Create the corresponding Tracks from the detected results of the first frame. Initialize the motion variable of the Kalman filter, and predict its corresponding frame through the Kalman filter. Tracks at this time must be unconfirmed.

(2) Perform IOU matching on the frame of the frame target detection and the frame predicted by Tracks in the first frame, and then calculate the cost matrix (the calculation method is 1-IOU) through the IOU matching result.

(3) Use all the cost matrices obtained in (2) as the input of the Hungarian algorithm to obtain linear matching results. At this time, we get three results. The first is Unmatched Tracks, we directly delete the mismatched Tracks ,because this Tracks is in an uncertain state, if it is in a definite state, it can only be deleted after reaching a certain number of times (default 30 times); the second is Unmatched Detections, We initialize such Detections as a new Tracks; the third is that the detection frame and the predicted frame are successfully paired, which means that we have successfully tracked the previous frame and the next frame, and the corresponding Detections are passed through Kalman Filter updates its corresponding Tracks variable.

(4) Repeat steps (2)-(3) until the confirmed Tracks appear or the video frame ends.

(5) Predict the boxes corresponding to the Tracks of the confirmed state and the Tracks of the uncertain state through the Kalman filter. Cascade match the frames of the confirmed Tracks with the Detections. Previously, the appearance features and motion information of the Detections were saved every time the Tracks were matched. By default, the first 100 frames were saved, and the appearance features and motion information were used for cascade matching with the Detections. This is done because the Tracks and Detections of the confirmed state are more likely to match.

(6) There are three possible results after cascade matching. The first one, Tracks matching, such Tracks update their corresponding Tracks variables through Kalman filtering. The second and third types are the mismatch between Detections and Tracks. At this time, the previously unconfirmed Tracks and the mismatched Tracks will be matched with Unmatched Detections one by one for IOU matching, and then the cost matrix will be calculated based on the IOU matching results, and its calculation method is 1-IOU.

(7) Repeat steps (5)-(6) until the end of the video frame.

2.2.3 State Estimation and Tracking Processing

Deep-SORT uses the result of the detector to initialize the tracker. Each tracker will set a counter. After Kalman filtering, the counter is accumulated. When the prediction result and the detection result successfully match, the counter is set to 0.If a tracker does not match a suitable detection result within a period of time, the tracker is deleted.Deep-SORT assigns a tracker to the new detection results in each frame. When the prediction results of the tracker match the detection results for 3 consecutive frames, it is confirmed that a new track has appeared, otherwise the tracker is deleted.

Deep-SORT uses an 8-dimensional state space (u, v, γ, h, ẋ, ẏ, γ̇, ḣ) to describe the state of the target and its motion information in the image coordinate system. u and v represent the center coordinates of the target detection frame, γ and h represent the aspect ratio and height of the detection frame respectively, and (ẋ, ẏ, γ̇, ḣ) represent the relative speed of the first four parameters in image coordinates. The algorithm uses a standard Kalman filter with a constant velocity model and a linear observation model, taking the bounding box parameters (u,v,γ,h) as direct observations of the object state.

2.2.4 Allocation Problem

Deep-SORT combines motion information and appearance information, and uses the Hungarian Algorithm to match prediction boxes and tracking boxes. For motion information, the algorithm uses the Mahalanobis distance to describe the degree of correlation between the Kalman filter prediction results and the detector results, such as the formula:

$$d^{(1)}(i,j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i)$$

In the formula, d_j and y_i represent the state vectors of the j-th detection result and the i-th prediction result respectively, and S_i represents the covariance matrix between the detection result and the average tracking result.

The Mahalanobis distance measures the standard deviation of the detection results from the average tracking results, taking into account the uncertainty of the state estimation, and can exclude low-probability associations.

When the uncertainty of target motion information is low, the Mahalanobis distance is a suitable correlation factor, but when the target is occluded or the camera perspective shakes, only using the Mahalanobis distance correlation will cause the target identity to switch. Therefore, consider adding appearance information, calculate the corresponding appearance feature descriptor r_j for each detection frame d_j, and set $//r_j//=1$. For each tracking track k, set the feature warehouse $R_k=\{r_k^{(i)}\}_{k=1}^{L_k}$, which is used to save the feature descriptors of the last 100 objects successfully associated, $L_k = 100$. Calculate the minimum cosine distance between the i-th tracking frame and the j-th detection frame, such as the formula:

$$d^{(2)}(i,j)=\min\{1-r_j{}^T r_k^{(i)} \mid r_k^{(i)} \in R_i\}$$

When $d^{(2)}(i,j)$ is less than the specified threshold, the association is considered successful.The Mahalanobis distance can provide reliable target location information in the case of short-term prediction, and the cosine similarity of appearance features can be used to restore the target ID when the target is occluded and reappears.In order to make the advantages of the two metrics complement each other, a linear weighting method is used to combine:

$$c_{i,j}=\lambda d^{(1)}(i,j)+(1-\lambda)d^{(2)}(i,j)$$

2.2.5 Deep Appearance Features

The original algorithm uses a residual convolutional neural network to extract the appearance features of the target, and trains the model on a large-scale pedestrian re-identification dataset, making it suitable for pedestrian detection and tracking. Since the original algorithm is only used for the pedestrian category, the input images are all scaled to 128×64, which is inconsistent with the aspect ratio of the yam target. In order to make the model suitable for vehicle feature extraction, this paper improves the network model. After adjustment the network input image size is 256×64.

3.Results and Discussion

3.1 Dataset and Experimental Environment

The dataset used in the experiment is mainly from the Michishita Hironaga Farm in Obihiro City, Hokkaido.Fix the gopro camera on a tractor running in parallel and at constant speed,taking a bird's-eye view of the Chinese Yams placed on the field. The angle is fixed, the lighting conditions and the environmental background are in good condition,as in Figure 5.Use the LabelImg tool to label and create a dataset.The experiment uses Pytorch as the software framework, and the model training hardware environment is Intel(R) Core(TM) i7−11800H (16 GB) and NVIDIAGeForce RTX 3060 Laptop GPU (6 GB).

Fig.5 Chinese Yam dataset collection scene

3.2 Parameter setting and evaluation index

Model training parameter settings: the input image size is 600×450, the number of iterations is 100, the batch size is 16, and the initial learning rate is 0.001.

The recall rate M_r, the average precision M_p, the average missed detection rate M_m, and the average false detection rate M_f are used as the evaluation criteria of the target detection model.

$$M_r = \frac{T_P}{T_P + F_N}$$

$$M_p = \frac{T_P}{T_P + F_P}$$

$$M_m = \frac{F_N}{F_N + T_P}$$

$$M_f = \frac{F_P}{F_P + T_N}$$

In the formula: T_P is the yam that is correctly detected; F_N is the yam that is not detected; F_P is the yam that is falsely detected; T_N is the yam that is not falsely detected.

Define the number of real trajectories in the t frame as g_t, the number of successfully matched trajectories in these trajectories is recorded as c_t, the matching cost of the i pair of successful matches is recorded as d_t^i, and the true trajectories that have not been successfully matched are recorded as m_t.The trajectory predicted by the model but not successfully matched in the t frame is recorded as fp_t. If there is an inconsistency in trajectory matching in two adjacent frames, it means that identity switching has occurred, and the number is recorded as mme_t. Based on these definitions, some comprehensive multi-target tracking model performance evaluation indicators can be generated.

$$1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t}$$

The number of identity switching IDs and the number of tracking frames per second Speed are used as the evaluation criteria of the tracking model.The statistical accuracy of Chinese Yam flow is used as the evaluation standard of the whole scheme.

3.3 YOLOv5s ablation and comparison experiment

The dataset in this paper randomly divides the ratio of 6:2:2 into the training set, validation set and testing set of the detection network.In order to verify the three improvement strategies for YOLOv5s proposed in this paper, an ablation experiment was carried out on the above data set to judge the effectiveness of each improvement point, and CBAM and CIoU Loss were added to the initial YOLOv5s in turn.The experiments do not use the pre-trained model, and the training process uses the same parameter configuration.

CBAM	CIoU Loss	Precision	Recall	AP@0.5
×	×	90.2	93.9	93.3
√	×	92.6	96.8	95.1
×	√	91.0	93.7	94.2
√	√	93.9	96.6	95.8

Table 1 Ablation of YOLOv5s

The first row of Table 1 indicates the basic performance of the original YOLOv5s on the data set, and the average detection accuracy is 93.3%.After introducing CBAM and CIoU Loss respectively, it can be seen that CBAM improves the detection results more significantly, Precision, Recall, and AP all have significant improvements, while the performance of CIoU Loss is slightly weaker.According to the analysis, this is related to the different functions of the two modules. The attention mechanism aims to improve the network's ability to extract important features, and the result is an increase in accuracy, while CIoU Loss speeds up the regression of the prediction frame and improves Regression accuracy, so there is only a small improvement in detection accuracy.After introducing CBAM and CIoU Loss at the same time, the detection network achieved the best results, and the average precision AP was increased by 2.5 percentage points compared with the original network.Visualize some test results of CIoU Loss and DIoU-NMS, as shown in Figure 6.Missed Chinese yams were detected while maintaining high detection accuracy.

Fig.6 CBAM+CIoU Loss+DIoU-NMS of ablation result

For horizontal comparison, this paper selects Faster R-CNN+FPN, YOLOv3+SPP, and mobilenetv2-YOLOv4 to train and test on the same data set, all using pre-trained models.The experimental results are shown in Table 2. It can be seen that the improved YOLOv5s in this paper is ahead of the other three networks in terms of detection rate FPS, weight size, and average precision AP. For the network weight size, the original YOLOv5s is 7.2 MB, and the improved network only increases 0.5MB.

Model Name	FPS	Weight Size/MB	AP@0.5
Faster R-CNN+FPN	12	311	88.7
YOLOv3+SPP	56	238	92.1
Mobilenetv2_YOLOv4	82	47.59	93.5
Proposed	120	7.7	95.8

Table 2 Comparison of different detection networks

3.4 Deep Appearance Feature Extraction Network Experiment

The adjusted re-identification network is trained on the data-set YaRi, the input image size is 256×64, and the other parameters remain unchanged.Connect the improved YOLOv5s and Deep-SORT after yam re-identification for testing, the results are shown in Table 3.

Model Name	IDs/time	Speed/Hz
SORT	65	60
Deep-SORT	28	33
YaRi Deep-SORT	22	31

Table 3 Chinese Yam tracking experiment

Since the SORT algorithm only uses motion features as the basis for target association, a total of 65identity switches occurred in the yam tracking of the above data. Compared with SORT, Deep-SORT reduced by 56%, and the model after the yam re-identification in this paper further reduces IDs, not only IDs are reduced to 22 times, but also the detection speed can reach 31Hz on the local test platform, which meet the standard of real-time detection.

3.5 Statistical experiment of Chinese Yam quantity

This paper draws on the statistical methods of pedestrian flow monitoring and vehicle flow monitoring. The difference is that the moving detection target and the fixed camera are converted into fixed detection target and moving camera.Use the method of setting the detection line in the video to count the traffic of Chinese yam.

The specific method is: set a red dot on the left border of the target tracking bounding box to represent the trajectory of the target,as Figure 7;and set a virtual detection line perpendicular to the direction of camera travel in the field road,as Figure 8;When the trajectory of a point representing the target tracking frame intersects the detection line, the total number of yam flows is accumulated, and the coordinates of the point are recorded.Because the detection target passes in one direction, there is no need for two-way counting statistics.The test results are shown in the figure 9.

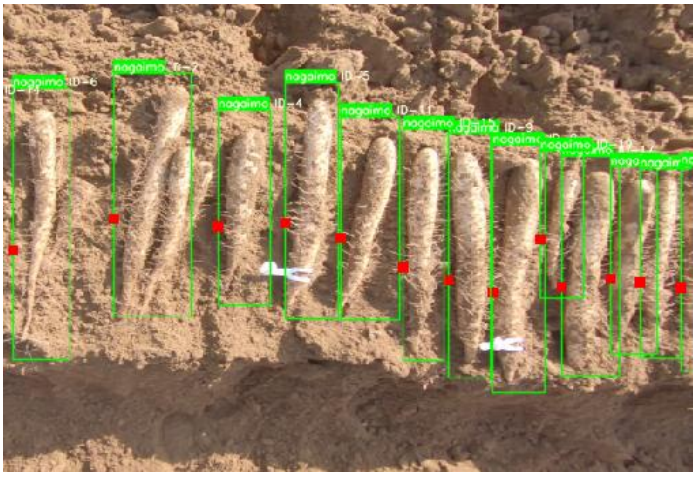

Fig.7 Target Tracking Bounding Box With Dot

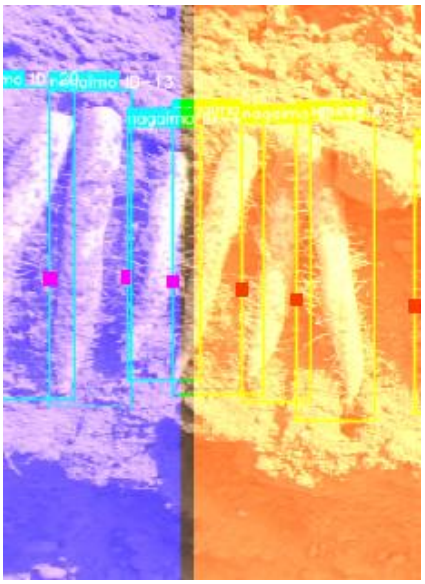

Fig.8 Virtual Vertical Detection Line

Fig.9 Statistical Results of Chinese Yam Traffic

The Chinese yam flow statistical test data collected in this paper is manually counted and compared with the experiment. The results are shown in Table 4. Each row of data represents the traffic flow counted by manual statistics, the original algorithm and the improved algorithm in the test video.From the results, it can be seen that the accuracy of the improved method proposed in this paper is higher than that of the original algorithm. Due to light problems and soil coverage on the surface of yams, some yams may be missed, which affects the accuracy of target tracking.

Model	Result
Manual Count	197
Original YOLOv5s+Deep-SORT	153
Proposed YOLOv5s+Deep-SORT	181

Table 4 Chinese Yam Number Statistics Experiment

4.Conclusion

In this paper, YOLOv5s is used as a detector, combined with the Deep-SORT target tracking method to count the number of Chinese Yams.The fusion of the attention mechanism CBAM and YOLOv5s effectively improves the accuracy of the detector; the CIoU Loss loss function and DIoU-NMS non-maximum suppression are used to replace the original GIoU Loss and ordinary NMS,

which further improves the positioning accuracy of the detector , and effectively improved the missed detection phenomenon in the Chinese Yam overlapping coverage scene.The original feature extraction network in Deep-SORT is used for input adjustment and re-identification training to make the algorithm more suitable for the application of taro crops such as Chinese Yams.The improved YOLOv5s detector is connected to the algorithm to conduct a statistical experiment on the number of Chinese yams. The results show that the algorithm proposed in this paper shows good statistical accuracy.

In the deployment of the model, it is difficult for the model to achieve the ideal speed for low-end device hardware without GPU and other hardware foundations. Therefore, it is necessary to compress the network structure on this basis to meet the needs of mobile terminals.

References

[1] Egi Yunus, Hajyzadeh Mortaza, Eyceyurt Engin. Drone-Computer Communication Based Tomato Generative Organ Counting Model Using YOLO V5 and Deep-Sort[J]. Agriculture Volume 12, Issue 9. 2022. PP 1290-1290.

[2] Fukudome Chiaki, Takisawa Rihito, Nakano Ryohei. etc. Analysis of mechanism regulating high total soluble solid content in the parthenocarpic tomato fruit induced by pat-k gene[J]. Scientia Horticulturae Volume 301, 2022.

[3] Xiang S, Wang SY, Xu M. etc. Correction: YOLO POD: a fast and accurate multi-task model for dense Soybean Pod counting [J]. Plant Methods Volume 19, Issue 1. 2023. PP 45-45.