IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 December 2023, accepted 24 December 2023, date of publication 26 December 2023,
date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347588

RESEARCH ARTICLE

Combining YOLOv7-SPD and DeeplabV3+ for Detection of Residual Film Remaining on Farmland

DEQI HUANG AND YANGTING ZHANG
College of Electrical Engineering, Xinjiang University, Ürümqi 830017, China
Corresponding author: Yangting Zhang (zyt990501@126.com)

This work was supported in part by the Key Technology Research Project through the Key Research and Development Program of the Autonomous Region on Total Recycling and Clean Conversion of Agricultural Residual Film under Grant 2022B02038.

ABSTRACT Aiming at the problems of low pickup rate of residual film recycling machine, low recognition of background soil and residual film left in farmland under complex farmland environment, and mutual occlusion between classes, we propose a method combining YOLOv7-SPD target detection and Deeplabv3+ image segmentation. YOLOv7-SPD was first introduced to recognize and locate the residual film left in the farmland, and the detected residual film image was passed to the image segmentation algorithm, and the segmented image was processed to calculate the area of the residual film in the farmland. By improving the loss function, fusing the Coordinate Attention (CA) mechanism, and introducing the Space-to-Depth (SPD) module and Atrous Separable Convolution (ASConv) to improve the accuracy of the leftover film detection of farmland residual film. The experimental results show that the average detection precision of the final improved model recall is 87.62% and the average precision is 93.72%, which are 4.93% and 2.53%; The mIOU and F1 of the image segmentation model reached 91.55% and 94.77%, respectively, which is more significant. This research result demonstrates the potential of this algorithm in practical applications related to agricultural residue management and field cleanliness assessment, providing certain technical support to improve the recovery rate of residual film recycling machines and realizing the accuracy and efficiency of detection.

INDEX TERMS Agricultural residual film leftover, attention module, YOLOv7, DeeplabV3+.

I. INTRODUCTION

In modern agricultural technology, mulch is used in agricultural production, which can improve the yield of agricultural products. However, with the expansion of the use of mulch film, the problems of difficulty in recycling and incomplete recycling have gradually emerged, leading to the increasingly prominent dilemma of mulch film residues in mulched farmland. Residual film leftovers can cause ecological pollution in farmland and affect crop yields [1]. Exposed to the natural environment for a long time, plastic films will decompose into tiny particles and enter the soil, water, and atmosphere, causing irreversible damage to the ecosystem, and the long-term accumulation will lead to a decline in the quality of the soil, affecting the growth and development of crops [1]. It destroys the soil ecological balance and puts soil microbial, animal, and plant diversity at risk, and edible crop products contaminated with residual films pose a risk to populations [3]. In this situation, we need to take measures to solve the problem of film residues in agricultural fields.

Currently, residual film recycling relies on residual film recycling machines and manual pickup for residual film recycling [3]. The current residual film recycling machine technology is not mature enough to efficiently clean and recycle all types of agricultural residual film. The recovery rate of existing residual film recycling machines ranges from 82% to 92%, and different soil environments, different materials, thicknesses, and sizes of residual films will have an impact

The associate editor coordinating the review of this manuscript and approving it for publication was Tai Fei.

on the recovery rate of residual film recycling machines [5]. For the smaller residual film left over from the farmland, the use of mechanical equipment processing cannot be completely recovered, and it is necessary to manually pick up the supplemental collection. However, manual pickup is difficult, time-consuming, and laborious, and the recovery efficiency is low. Therefore, it is particularly important to modernize and efficiently process the recovery of residual film in agricultural fields.

Nowadays, with the help of advanced vision systems and machine learning algorithms that can accurately identify and localize the residual film in farmland, the efficiency and quality of residual film recycling can be improved to ensure efficient and accurate recycling of residual film. Jiang et al [6] proposes a traditional target detection method to locate the residual film in farmland, using the minimum outer rectangle method to determine the target area of the residual film and obtain the matching function, but at the same time, some of the residual film information is also lost, which increases the localization error. With the rapid development of deep learning technology, deep learning-based computer vision algorithms are widely used in real-world scenarios instead of the traditional manual selection of features. The common target detection algorithms are YOLO series [7], SSD [8], EfficientDet [9], R-CNN [10], Mask R-CNN [11], and Faster R-CNN [12]. The image segmentation algorithms are U-Net [13], Deeplab [14], PSP-Net [15], Transformer [16] and Segformer [17]. Deep learning-based computer vision detection methods can learn more complex feature representations, thus improving the accuracy of detection. In recent years, many scholars have studied agricultural neighborhoods using deep learning methods. Zhai et al. [18] proposed a method based on a U-Net model to establish an improved U-Net model for evaluating residual film contamination, but it does not apply to the detection of small targets in proximity. Qiu et al. [19] proposed a method for predicting residual film content in the plow layer of cotton fields based on UAV imaging and deep learning. The UAV collected images of residual film on the soil surface of the cotton field, manually sampled to obtain the residual film content of the plow layer, and constructed a residual segmentation model of the cotton field image. Zhang et al. [20] proposed a recognition method of residual film in farmland, applying the improved Faster R-CNN convolutional neural network to the recognition and detection of residual film in farmland, which meets the requirements of recognizing residual films. Zhou et al. [21] proposed the MFFM Faster R-CNN algorithm to solve the problem of multi-scale variation of residual film and enhance the feature extraction capability.This method is improved compared with the traditional algorithm, but the training steps are more cumbersome, the neighboring windows have large repetitive information, and there are more invalid regions, which leads to a large amount of computation and slow detection speed. Zhou et al. [21] proposed a method for lightweight detection of farmland trash and identification and detection of residual film, but there are more types of farmland trash, and the accuracy of the target detection of residual film is not high. Chen et al. [23] proposed the YOLOv7-WFD algorithm to enhance the ability of the model to extract target features. Chen et al. [23] proposed a method based on an attention mechanism for detecting residual film in farmland, which provides strong technical support for accurately recognizing residual film in farmland. However, it was only able to detect the residual film and did not do the segmentation process of the residual film. Yan et al. [25] designed a DeepLabV3 based algorithm to solve the problem of low detection accuracy of semantic segmentation algorithms used in remote sensing image cropland segmentation. Wu et al. [26] proposed a lightweight MobileNetV2_DeepLabV3 image segmentation network with good dam detection. Meng et al. [27] proposed a spatio-temporal convolutional neural network model for detecting pineapple fruits by using Shift Window Transformer fused region convolutional neural network model.

In the research of agricultural residual film recognition, the traditional algorithms increase the localization error, and the detection speed is slow using two-stage target detection methods. With the development of technology, most studies use UAV imaging and remote sensing images for farmland residual film detection, and there is no more research on residual film small target detection, based on this, this study selects YOLOv7 and DeeplabV3+ detection network as the base model, and optimizes and improves the network model for the problem of complex farmland environment and more background interferences, and improves the detection of residual film in farm-land by applying deep learning to the Detection and identification of residual film in farmland, to solve the problem of detecting residual film fragments left in farmland, and to complete the identification, localization, and calculation of residual film area in farmland, which is of some significance for residual film recycling. The final research results were assessed on actual farmland images in Yuli County to verify the effectiveness of the algorithm.

The contributions of this study are summarized as follows:

(1) A single-stage YOLOv7-SPD-based residual film leftover detection method for farmland is proposed, which can distinguish the residual film from cotton in the complex cotton field environment and complete the identification and localization of the residual film in the farmland with high accuracy.

(2) A method of image segmentation of residual film in farmland is proposed, in which the detected residual film in farmland is further segmented, and the area of the residual film is obtained by extracting pixel points, which provides certain technical support for improving the recovery rate of residual film recycling machine.

II. RELATED WORKS

YOLOv7 takes the input image extracts features through a backbone network and then performs target detection at multiple scales through a feature pyramid network [26]. This has the advantage of capturing features of both small-scale and large-scale targets, improving the accuracy and recall

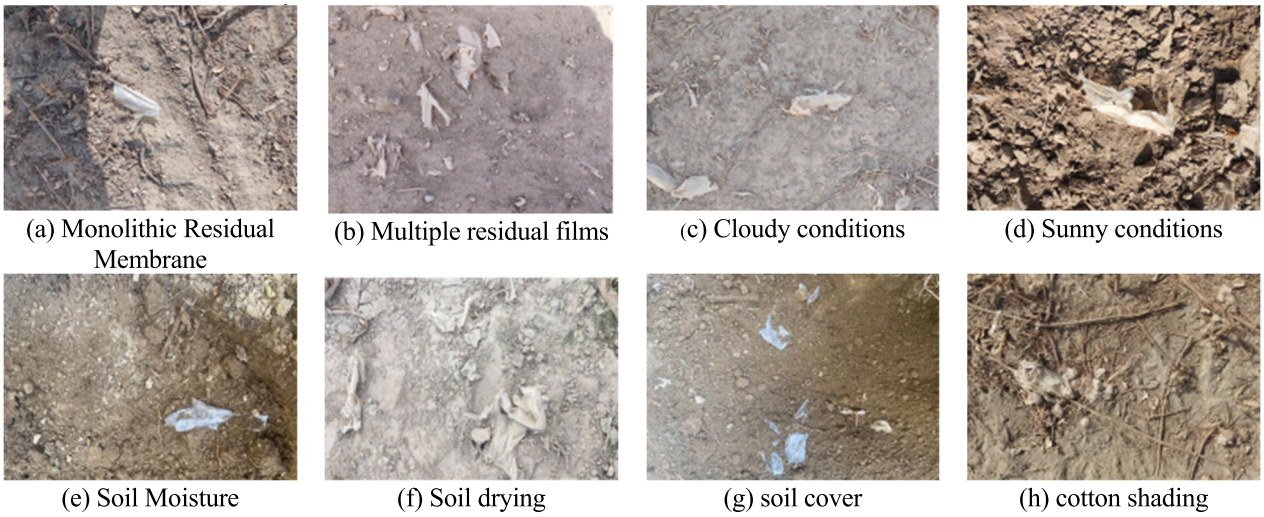


FIGURE 1. Sample data on leftover film from agricultural fields.

of detection. Null convolution, on the other hand, expands the receptive field by introducing cavities or holes in the convolution kernel so that the network can better understand the contextual information of the target, thus improving the performance of target detection.

DeeplabV3+ is a semantic segmentation model that is based on deep convolutional neural networks and achieves the task of image segmentation by classifying the input image at the pixel level [29]. DeeplabV3+ employs two main techniques to improve the performance: Dilated convolution and Atrous Spatial Pyramid Pooling (ASPP). Null convolution increases the receptive field while maintaining the size of the feature map, allowing the network to better capture the details in the image. This is useful for segmentation of small objects or details in an image. ASPP, on the other hand, extends the network's field of view by using different sampling rates and utilizes null convolution to obtain a wider range of contextual information. This helps to improve the network's segmentation results at various scales, especially when segmenting large targets or objects. DeeplabV3+ also introduces two other important techniques: Image Pyramid Pooling and Conditional Random Field. Image Pyramid Pooling generates input features of different scales, thus improving the network's segmentation of multi-scale objects.

III. MATERIALS AND METHODS

A. EXPERIMENTAL EQUIPMENT AND DATA

In this paper, the residual film left on the surface of farmland after the work of the 11JCM-300 residual film recycling machine in a cotton field in Yuli County, Bayin Mongolia Autonomous Prefecture, Xinjiang were taken as the research object, and the pictures of residual film left on the surface of farmland with different weather, different light conditions, different degree of shade, and different number of residual film pieces were collected, and a total of 3,349 images of residual film were collected, with an image resolution of 3,072 pixels×4,096 pixels as shown in Figure 1, which is a sample data of residual film left on the surface of farmland under the complicated situation of farmland environment with different humidity, light, and degree of shade.

The data were analyzed, and the problems of a small target and uneven sample size of sample data were found, due to the complex environment of the farmland, there are more cotton residues of farmland disturbances, and cotton and mulch film have more similar characteristics. To better identify the small target of residual film leftover film, cotton was labeled. The number of farmland residual film and cotton categories are 8340 and 6280, respectively. To improve the performance of the model and reduce the risk of overfitting, augmenting the dataset helps the model to learn to better recognize and detect the specific target of residual film in agricultural fields. In this paper, the data augmentation ratio in mosaic mixup was used to augment the dataset. Training samples with higher diversity are generated to help the model learn and generalize better, resulting in 5580 images of leftover agricultural residual film.

B. YOLOv7-SPD FARMLAND RESIDUAL FILM LEFTOVER FILM DETECTION MODEL CONSTRUCTION

In this study, a YOLOv7-SPD farmland residual film leftover film detection method is proposed to address the problems of complex farmland environment and low pickup rate of residual film recycling machine. Aiming at the characteristics of farmland residual film fragments with small areas in the land, irregular shapes, complex boundary features, and color close to the soil, we can improve the accuracy of small target detection by improving the network model and loss function. First, the CA attention is introduced into the Spatial Pyramid Pooling Connected Spatial Pyramid Convolution (SPPCSPC) structure of YOLOv7, which enables the model to better recognize and distinguish the residual film leftover film and soil by increasing the sensory field of the model.

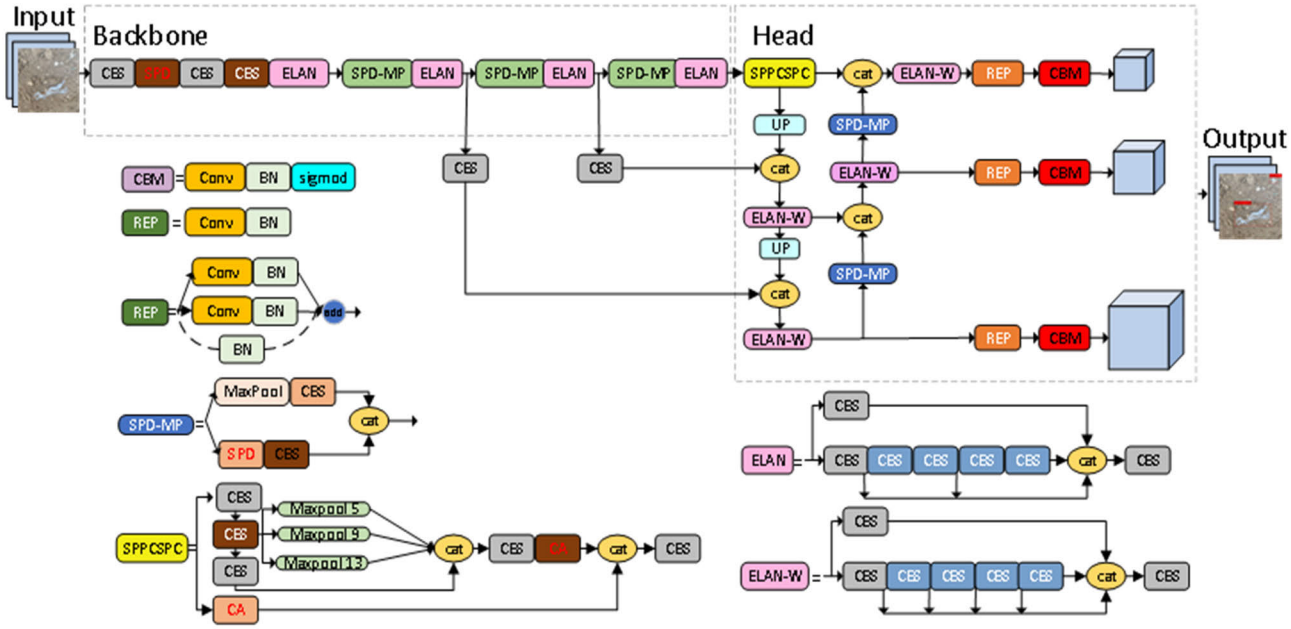


FIGURE 2. Structural diagram of the improved YOLOv7.

This attention mechanism can adaptively weight the channels of the feature map to improve the expression of the features. Second, in the improved MP-1 and MP-2 structures, the CBS in the lower half of the MP structure was replaced with the SPD structure. The purpose of this is to retain more residual membrane leftover features, avoid feature loss, and enhance the detection ability of small targets. With these improvements, cotton and residual film small targets can be effectively differentiated and the accuracy of detection can be improved. These improvements enable the YOLOv7-SPD model to better handle complex target features and increase the sensitivity and recognition ability for small targets. The structure of the improved model is shown in Figure 2.

1) NORMALIZED WEIGHTED DISTANCE LOSS FUNCTION

Since the residual film in farmland accounts for a relatively small proportion of the land area and the film accounts for a relatively small proportion of the image, the original IoU (Intersection over Union) metric is not suitable enough, so the rectangular box similarity metric NWD (Normalized Weighted Distance) is used to detect the similarity between two rectangular boxes for small targets in the farmland environment [30]. The NWD metric uses the normalized Wasserstein distance to measure the similarity between two rectangular boxes, which considers the weights and distributions of the pixels and therefore captures the similarity between the bounding boxes of small targets more accurately. By introducing the NWD metric, the similarity metric problem of small targets in the farmland environment can be effectively solved to improve the accuracy and stability of farmland image analysis.

NWD measures the similarity of rectangular boxes by modeling them as Gaussian distributions and calculating the distribution distance. For bounding boxes with small targets, there are usually some background pixels present because the shape of real objects is often not strictly rectangular. To accurately assign weights to different pixels within the bounding box, a two-dimensional Gaussian distribution can be utilized. This method assigns the highest weight to the pixel at the center of the bounding box, while gradually decreasing the weight as the distance from the center increases.

The similarity between two rectangular boxes can be transformed into a distribution distance by comparing the Gaussian distributions associated with each box. The second-order Wasserstein distance between two 2D Gaussian distributions, $\mu_1 = N(m_1, \Sigma_1)$ and $\mu_2 = N(m_2, \Sigma_2)$, μ_1 and μ_2 is calculated as follows:

$$W_2^2(\mu_1, \mu_2) = ||m_1 - m_2||_2^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_2^{1/2}\Sigma_1\Sigma_2^{1/2})^{1/2}) \quad (1)$$

The formula for this loss function is as follows:

$$NWD(N_a, N_b) = \exp(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C}) \quad (2)$$

$W_2^2(N_a, N_b)$ is the distance metric and C being a constant that is closely related to the data set, NWD = 1 when the two rectangular boxes overlap completely, and NWD = 0 when the two rectangular boxes are infinitely far apart, thus it has the nature of IoU to some extent.

The NWD loss function weights the errors of each category and divides all the errors by the Euclidean distance between the true value and the mean value to overcome the problem of error comparison since the leftover film of agricultural residuals and cotton have different value domains from each other.

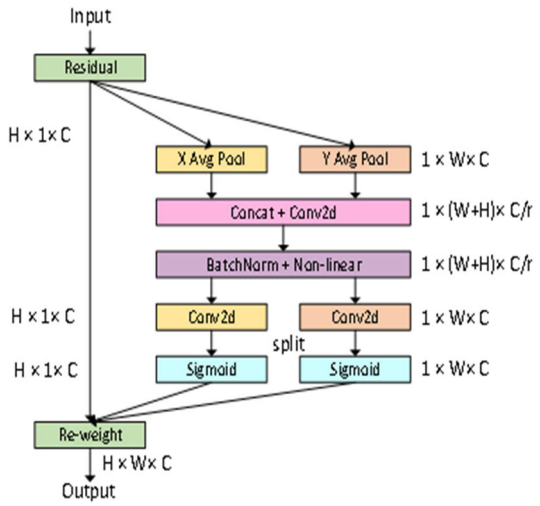


FIGURE 3. Structure of the coordinate attention mechanism.

2) COORDINATE ATTENTION MECHANISM

As the color characteristics of the farm residual film, cotton, and soil background color are very similar, and the maximum pooling residual edge structure in the SPPCSPC structure of YOLOv7 leads to the backward transmission of various types of information, which reduces the phenomenon of disappearing gradient but also brings more useless information backward transmission. Therefore, the introduction of the CA attention mechanism [31] at the position of residual edges and maximally pooled outputs can help the model to focus on the most relevant information, which can improve the performance and accuracy of the model.

The commonly used attention mechanisms are Efficient Channel Attention(ECA) [32], Convolutional Block Attention(CBAM) [33], Selective Kernel(SK) [34], Efficient Pyramid Split Attention(EPSA) [35], etc. However, the ECA module lacks long-range dependencies, there is a performance bottleneck problem for complex images, and irrational use of the ECA module can easily lead to overfitting. CBAM and SK modules can focus on important regions, but they focus excessively on local features, increasing network computation and complexity. The EPSA module has many model parameters, which does not apply to cross-domain tasks, and the model computation complexity is large. CA can adaptively learn the important weight of each channel, which helps to improve the model's ability to focus on important features and improve the discriminative degree of feature expression. The Coordinate Attention mechanism, on the one hand, can inhibit the backward transfer of useless gradient information, and can obtain semantic feature information over a longer distance, increase the model sensory field, and thus improve the model detection accuracy.

X Avg pool and Y Avg pool are the global average pooling in the width and height directions, respectively, as shown in Figure 3. The feature maps are spliced and fed into the convolution module with a shared convolution kernel of 1×1,

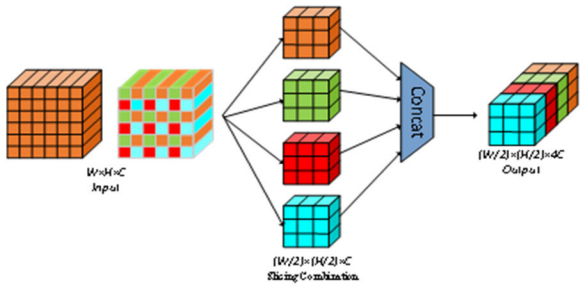


FIGURE 4. SPD structure diagram.

and the attention weights of the feature maps in height and width are obtained after batch normalization and Sigmoid activation function, respectively.

3) SPACE-TO-DEPTH

The shape and size of the leftover film in agricultural fields are different, and the boundary features are not obvious. When the leftover film is obscured by other objects or the background, the YOLOv7 network may not be able to capture the complete target information, resulting in the loss of fine-grained information, which may lead to false detection or missed detection. To solve the problem that the boundary features of small targets of residual film in farmland are not obvious, and more residual film features can be extracted, the Space-to-Depth (SPD) module is added to replace the original CBS module, and the SPD-MP structure is constructed for downsampling, and the specific operation flow is shown in Figure 4. By Slicing Combination, more features can be retained, and more information about the target features can be passed to the next layer to prevent feature loss, thus improving the model accuracy.

In this way, SPD can retain more fine-grained information about the residual film during downsampling and fuse features at different scales and locations to improve the performance of the processing task of modeling residual film left over from agricultural fields.

C. SEMANTIC SEGMENTATION BASED ON DEEPLABV3+ RESIDUAL FILM REMAINDER

To make DeeplabV3+ have better performance in terms of boundaries and details of farmland residual film, the improvement of the DeeplabV3+ model is shown in Figure 5, which incorporates a better loss function and introduces Strip Pooling (SP), and finally uses the null separable convolution to improve the feature extraction capability and enhance the model's ability to sense the boundaries of the farmland residual film.

1) STRIP POOLING STRUCTURE

Due to the inconspicuous edge portion of the leftover agricultural residual film, to reduce the loss of feature information, multi-scale feature fusion (MSFF) is performed, which improves the feature extraction capability by up sampling

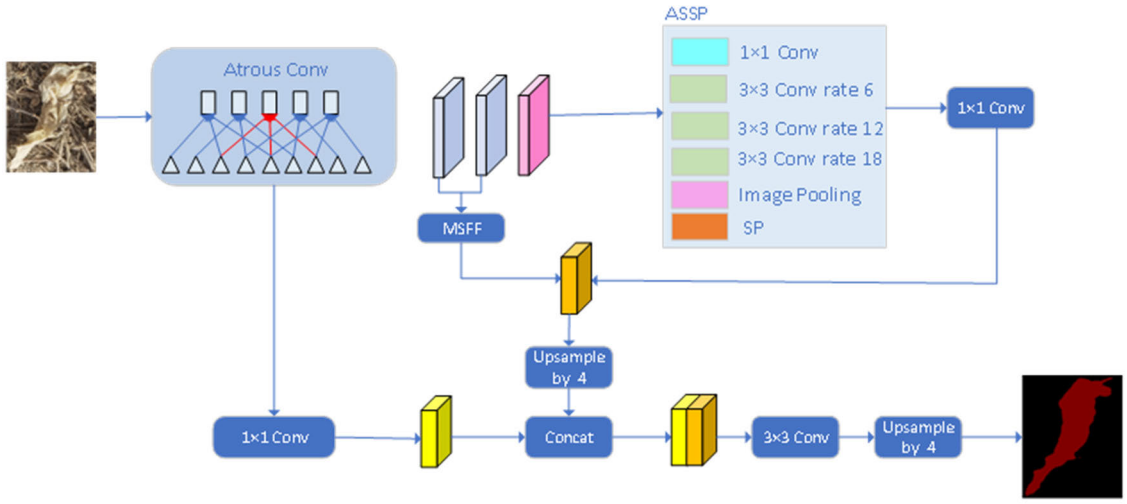


FIGURE 5. Improvement of DeeplabV3+ structure.

the features at several different sizes and introduces SP [36]. In traditional convolutional neural networks, the use of a pooling layer reduces the size of the feature map by down-sampling to capture higher-level features. In strip pooling, the feature map is divided into multiple strips, and then the pooling operation is performed on the features within each strip, as shown in Figure 6. Different from the traditional uniform division, strip pooling dynamically adjusts the width and position of the strips according to the content in the feature map, which can better preserve the object boundary and detail information. Strip pooling can enhance the model's ability to perceive the boundaries of complex objects in the image, which makes it possible to predict object contours when performing residual film segmentation more accurately.

The feature map $x \in R^{W\times H\times C}$ is pooled by a horizontal pooling kernel and a vertical pooling kernel for a feature map. Then it is fed into a one-dimensional convolution with a convolution of 3. The features are decoded and expanded to obtain $y^{H}_{c,i,j} \in R^{W\times H\times C}$ and $y^{W}_{c,i,j} \in R^{W\times H\times C}$ finally summed for feature fusion to obtain:

$$y_{c,i,j} = y^{H}_{c,i,j} + y^{W}_{c,i,j} \tag{3}$$

2) ATROUS SEPARABLE CONVOLUTION

The original ASPP module in DeeplabV3+ uses null convolution with different null rates to obtain the sensory fields at different scales of the image, to improve the feature extraction capability and the model accuracy, but this operation has many redundant operations, which will obtain many redundant feature maps, thus increasing the model parameters, so we propose a null separable convolution instead of the null convolution in ASPP, which can reduce the number of parameters in the model without loss of accuracy and reduce the number of model parameters. The input channel of each void convolution in ASPP is Cin and the output channel is Cout. The void convolution directly uses the convolution

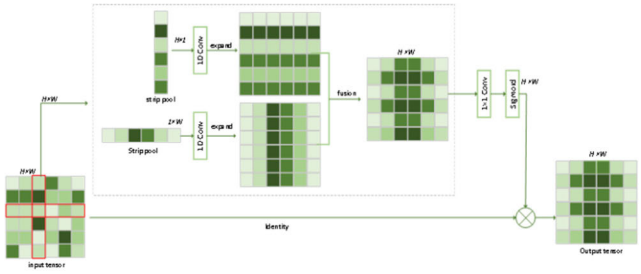


FIGURE 6. Strip pooling module.

with different void rates to get the output result. The cavity separable convolution first uses the grouped convolution with different cavity rates to reduce the number of parameters and then uses the convolution with convolution kernel 1 to change the number of channels to get the result with the output channel as Cout. Advantages such as large receptive fields, reduced number of parameters, improved computational efficiency, and preservation of feature map size are realized.

D. EXPERIMENTAL CONDITION

The experiment uses Pycharm to improve the YOLOv7 network model, the experimental environment is under Windows 10, 64-bit operating system, the CPU is Intel(R)Core (TM)i7-12700H, the RAM is 32GB, and the GPU adopts a 12GB graphics card of NVIDIA GeForce RTX 3080Ti, which has both pixel fill rate and graphics memory bandwidth compared with the other processors in terms of image processing. The GPU is an NVIDIA GeForce RTX 3080Ti with a 12GB graphics card, which has a better pixel fill rate and memory bandwidth than other processors in image processing. Cuda version 11.3 can better accelerate diversified workloads, and the programming language is Python 3.7. During the target detection training process, the input size of the model is 640 × 640 pixels, and the training batch size is 12, and the model

TABLE 1. Performance comparison of different detection algorithms.

Model	Computation(GFLOPs)	Precision(%)	Recall(%)	mAP(%)
YOLOv4	141.917	81.60	32.41	46.61
Faster R-CNN	401.714	69.53	83.09	65.24
YOLOv5s	16.485	81.98	77.25	73.29
YOLOX	26.760	88.06	82.05	82.46
YOLOv7	106.130	91.19	85.78	82.69
YOLOv8	111.665	80.90	74.40	81.90
Ours	96.589	93.72	92.12	87.62

is trained using stochastic gradient descent, (SGD) is used as the optimizer, and by continuously iterating this process, it helps the target detection model to gradually find better parameter configurations on the training data to improve performance and accuracy. The initial learning rate is 0.1, the minimum learning rate is 0.0001, and a total of 270 rounds of iterations. Image segmentation initial learning rate 0.007, minimum learning rate is 0.00007, epoch is 200.

E. MODEL EVALUATION INDICATORS

The target detection evaluation metric is primarily employed to assess the performance and accuracy of the target detection algorithm in detection tasks, providing a comprehensive evaluation of the model's performance. To evaluate the performance of the farmland residual film leftover film recognition and detection network, the main choice of P, Recal, F1, and mAP. As an evaluation indicator, the formula is shown below:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{4}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{5}$$

$$F_1 = 2\frac{PR}{P + R} \times 100\% \tag{6}$$

In the given context, T_P refers to the count of samples accurately identified with residual film, F_P represents the count of samples incorrectly identified with residual film, and FN denotes the count of samples where residual film remains undetected.

The mAP is a comprehensive evaluation metric to meas ure the balance between the accuracy and recall of the target detection algorithm on different categories.

$$mAP = \frac{\sum_{i=1}^{N} \int_0^1 P(R)dR}{2} \times 100 \tag{7}$$

$\sum_{i=1}^{N} \int_0^1 P(R)dR$ is the average precision AP, which is calculated by plotting the Precision-Recall curve and calculating the area under the curve. In the calculation, the residual film left over from agricultural fields as well as cotton is evaluated independently and finally, the average precision mean is calculated.

The mean intersection over union (mIOU) is a fundame ntal and significant metric utilized in the evaluation of image segmentation models. It serves as a crucial indicator for asses sing the performance and accuracy of semantic segmentation tasks. mIoU is the result of averaging the Intersection-over-U nion values predicted for each category and its formula is sho wn below.

$$mIOU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{8}$$

where i denotes the true value, j denotes the predicted va lue, and Pij denotes the prediction of i as j.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. COMPARISON OF PERFORMANCE OF DIFFERENT

1) YOLOV7-SPD ALGORITHM PERFORMANCE COMPARISON

To verify the superiority of YOLOv7 and its improved model over other target detection models, this paper examines and evaluates the performance of various widely adopted target detection algorithms on the current dataset. By conducting a comprehensive analysis, the study aims to assess the effectiveness and suitability of these algorithms for detecting targets in the given dataset, including the Two-stage target detection algorithm, Faster R-CNN, and the single-stage target detection algorithms, YOLOv4 and YOLOv5s, as well as YOLOX. The field-collected data set of residual film left over from the farmland was used for training and validation and evaluated in the test set. set for training and validation and evaluated in the test set. The comparison results shown in Table 1.

The comparison shows that the real-time detection speed of the five single-stage YOLO series target detection models is faster than that of the two-stage Faster R-CNN target detection model; the Faster R-CNN model has a higher recall, is computationally complex and time-consuming, and has lower average precision than that of the YOLO series, which may not be able to satisfy real-time requirements in real farmland environments; whereas in the single-target detection YOLO series, due to the increase in the number of model parameters of the YOLOv7 network, the computational amount also increases, and the computational amount of the model in this paper is higher than that of YOLOv5s and YOLOX, but compared with the original YOLOv7 model, it has been reduced by 9.541GFLOPs. compared with the current mainstream models YOLOv4, Faster R-CNN, YOLOv5s and YOLOX, the detection average accuracy mAP is higher and more time-consuming, and the average accuracy mAP

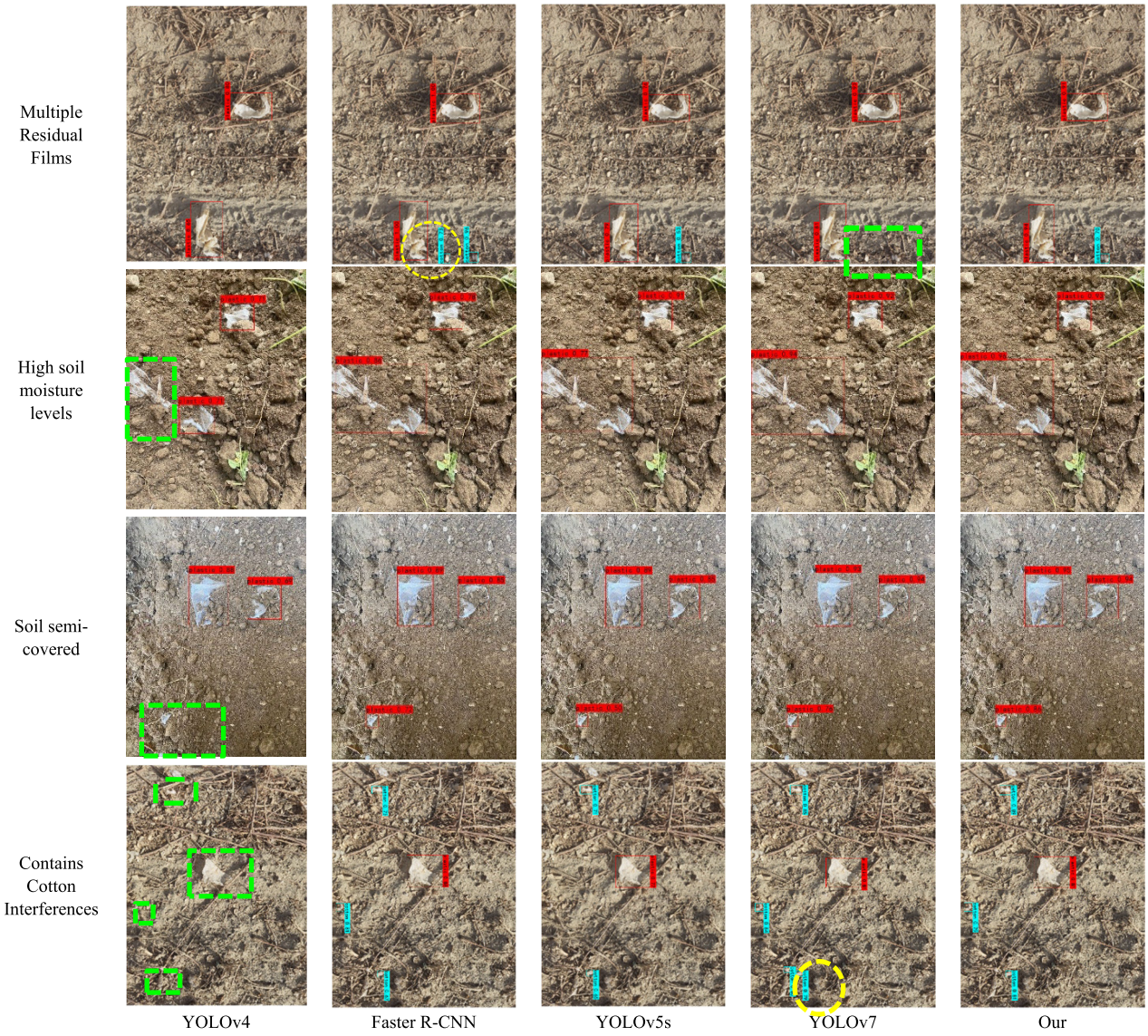


FIGURE 7. Effectiveness of different models in identifying residual film left in farmland.

is lower than that of the YOLO series, which may not meet the real-time requirements in real agricultural environments. Compared with the current mainstream models YOLOv4, Faster R-CNN, YOLOv5s, YOLOX, and YOLOv8, the detection average accuracy mAP is 41.01%, 22.38%, 14.33%, 5.16%, and 5.72% higher, respectively, which has a high comprehensive performance.

The improved model reduces the occurrence of missed mispickups compared to other models, and the average precision mean of detection is also improved compared to other models, as shown in Figure 7. The red thin rectangular box indicates the target area of residual film left in the farmland given by the network, the sky blue is the target area of cotton residue, the yellow circle is the residual target of the wrong pickup, and the green dashed rectangular box is the target of missed detection. The detection results show that the improved algorithm can accurately identify farmland residual film leftovers under complex environments such as different light, different soil moisture, cotton shading, soil half-masking, and residual film small targets. It shows that the improved model has a better effect on recognizing the leftover film of farmland residual film under the complex environment and has a certain high efficiency, which can provide a theoretical basis for the research and development of the residual film replenishment mechanical device of the intelligent residual film recycling machine.

2) COMPARISON OF DEEPLABV3+ MODEL ALGORITHM PERFORMANCE

To verify the superiority of DeeplabV3+ and its improved model over other target detection models, this section investigates the performance of Unet, PSPnet, HRNet, Segformer,

TABLE 2. Performance comparison of different segmentation algorithms.

Model	Precision(%)	Recall(%)	F1(%)	mIOU(%)
U-Net	93.71	85.74	89.97	85.15
PSP-Net	92.31	91.66	91.98	87.56
HR-Net	92.58	91.82	92.20	87.88
Segformer	92.52	87.97	90.16	85.19
Our	94.33	95.06	94.77	91.55

DeeplabV3+, and improved DeeplabV3+ algorithms in the current dataset, trained and validated with a field-collected dataset of residual leftover film from agricultural fields, which was and tested in the test set. The comparison results are shown in Table 2.

As can be seen from Table 2, although the U-Net model is slightly higher in accuracy, it has a larger model size and parameter count and thus requires larger memory resources during training and inference. For some complex scenes and small targets, due to its rougher up-sampling method, it may not be able to capture more detailed boundary and shape information, resulting in limited accuracy of segmentation results. In contrast, DeeplabV3+ adopts feature extraction methods such as null convolution and ASPP module, which can effectively deal with the occlusion problem of leftover agricultural residual film. The improved DeeplabV3+ improves the mIOU by 6.4%, 3.99%, 3.67%, and 6.36%, compared to U-Net, PSP-Net, HR-Net, and Segformer, respectively. The improved DeeplabV3+ has a recall of 95.06% and F1 of 94.77%, with better overall performance.

B. YOLOv7-SPD MODEL VALIDITY ANALYSIS

1) EFFECT OF ATTENTION MECHANISMS ON DETECTION MODELS

Since the original YOLOv7 network model cannot effectively capture the key information of the residual film in farmland and is easy to miss the detection, different attention mechanisms are introduced to improve the model to solve this problem. We selected four mainstream attention mechanism modules, CA attention, SE attention, ECA attention mechanism, and CBAM attention, and added them to YOLOv7 for ablation experiments. Through these experiments, the effect of attention mechanisms on model performance can be verified. The incorporation of attention mechanism modules facilitates the model in capturing farm residual film feature information with greater precision and effectively suppresses the transmission of irrelevant information. This enables the model to focus more selectively on the relevant features, enhancing its ability to accurately detect and analyze farm residual film, thus reducing the leakage detection rate. By comparing the effects of different attention mechanism modules, their contribution to the task of detecting leftover agricultural residual film can be evaluated. These ablation experiments help to improve the model's effectiveness in detecting residual film on farmland in practical applications, which in turn improves efficiency and reduces the waste of resources in the agricultural production process.

The CA, SE, ECA, and CBAM modules are sequentially added to the SPPCSPC structure of YOLOv7 to obtain the YOLOv7+CA, YOLOv7+SE, YOLOv7+ECA and YOLOv7+CBAM networks. The algorithmic precision, recall, F1, and AP values under different attention mechanisms are compared respectively, as shown in Table 3.

As can be seen from Table 1, the introduction of the CA Attention mechanism results in a significant improvement in the model's detection performance, leading to the highest average accuracy achieved. The CA module enables the model to focus on relevant features by selectively enhancing the importance of informative channels while suppressing the irrelevant ones. This enhancement in feature representation helps the model to accurately detect and classify farm residual film, ultimately improving the overall performance metrics. In terms of the detection precision of individual categories, compared to YOLOv7+SE, YOLOv7+ECA, and YOLOv7+CBAM, the average precision in detecting cotton was improved by 1.15%, 0.32%, and 0.54%, respectively; in the detection of residual leftover film in agricultural fields, the average precision was improved by 0.04%, 1.01%, and 1.03%, respectively; compared to the original YOLOv7 model, the model introducing CA attention mechanism has improved in F1, recall, precision and average precision. The experiment shows that the YOLOv7+CA model can effectively improve the accuracy of detection and has a certain superiority.

2) YOLOV7-SPD DETECTION MODEL ABLATION EXPERIMENTS

In order to validate the effectiveness of the final model, this section investigates the impact of the NWD loss function, attention mechanism and space-for-depth module for downsampling on the performance of the model, and analyzes the model recall and mean average precision mean mAP performance metrics under different influencing factors, with details of the specific data shown in Table 4. Each improvement has a certain improvement on the model effectiveness, and the final improved model has the highest recall and mAP value in the detection of residual film leftover film in agricultural fields, reaching 92.12% and 87.62%, respectively, which indicates that the improvement measures in this paper have a positive effect on the model and Effectively improves the precision of recognizing residual film left in farmland.

C. IMPROVED DEEPLABV3+ MODEL VALIDITY ANALYSIS

1) IMPACT OF LOSS FUNCTION ON SEGMENTATION MODELS

CE loss is a commonly used loss function for classification tasks to measure the difference between the model output and the true label. The performance of the model is evaluated by calculating the cross entropy between the true label and the model's prediction. The smaller the cross-entropy is, the

TABLE 3. Experimental results on the effect of attentional mechanisms on detection models.

Model	F1(%)		Recall (%)		Precision (%)		Average precision (%)	
	Cotton	Remnants of agricultural film	Cotton	Remnants of agricultural film	Cotton	Remnants of agricultural film	Cotton	Remnants of agricultural film
YOLOv7	70	88	60.78	85.78	82.08	91.19	73.89	91.48
YOLOv7+CA	74	91	65.59	89.76	85.01	92.16	78.04	93.45
YOLOv7+SE	73	91	65.06	90.01	84.32	91.80	76.89	93.41
YOLOv7+ECA	74	91	67.00	89.44	83.53	92.60	77.72	92.44
YOLOv7+CBAM	73	90	62.30	87.25	87.18	92.51	77.50	92.42

TABLE 4. YOLOv7-SPD detection model ablation experiments.

Model	AP(%)	Recall(%)	mAP(%)
YOLOv7	91.48	85.78	82.69
+ NWD loss	93.26	89.36	84.29
+Coordinate Attention	93.45	89.76	85.75
+SPD-MP module	94.95	92.12	87.62

TABLE 5. Impact of loss function on segmentation model.

Model	Precision(%)	Recall(%)	F1(%)	mIOU(%)
CE Loss	92.13	92.83	92.48	88.23
CE Loss + Dice Loss	92.88	93.52	93.20	89.30

TABLE 6. Segmentation model ablation experiment test results.

Model	Precision (%)	Recall (%)	F1 (%)	mIOU (%)	Number of parameters(M)
DeepLav3	92.88	93.52	93.20	89.30	5.813
+ MSFF	94.58	93.53	94.05	90.63	5.906
+Enchce ASPP(SP)	94.44	94.90	94.72	91.52	6.137
+ASconv	94.33	95.06	94.77	91.55	4.180

closer the model's prediction results are to the true label, the smaller the loss is, and the better the model's performance is. However, it is sensitive to outliers, and its contribution to the overall loss is larger when there are extremely incorrect prediction probabilities, which may cause the model to pay too much attention to these abnormal samples and ignore other more valuable information. And when the sample distribution not balanced, i.e., the number of samples in some categories is much larger than others, the CE loss function may lead to poorer prediction of the model for a few categories.

The leftover films of farmland residual films are of varied sizes and occupy different areas, to balance the relationship between positive and negative samples, the Dice Loss loss function is introduced. Dice loss is a metric function used to evaluate the similarity of two samples, which takes into account the local similarity between the prediction results and the real labels when calculating the loss, not only focusing on the overall classification accuracy, but also making the neighboring pixels in the image category relationships remain consistent.

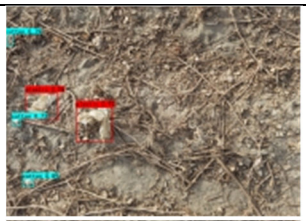
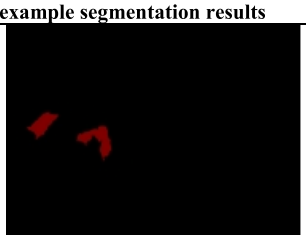
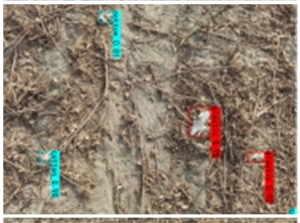
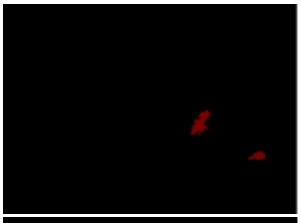
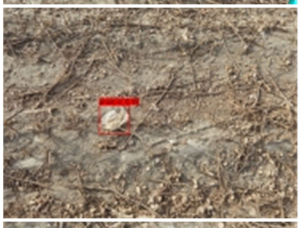
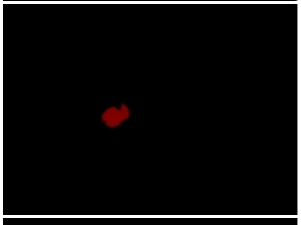
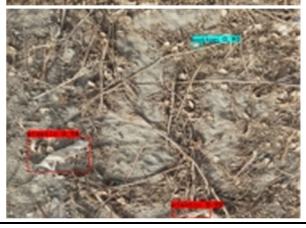
As can be seen from Table 5, Dice Loss considers the local similarity between the predictions and the true labels when calculating the loss, rather than focusing only on the overall classification accuracy. This allows Dice Loss to better preserve spatial consistency, i.e., consistency in the category relationships of neighboring pixels in an image. The addition of Dice Loss improves Precision by 0.75%, Recall by 0.69%, F1 by 0.72%, and mIOU by 1.07%, and the segmentation effect of the model is greatly improved.

2) SEGMENTATION MODEL ABLATION EXPERIMENT TEST RESULTS

To verify the validity of the segmentation model, different influencing factors were added to analyze the model PRECISION, RECALL, F1, mIOU, and the number of parameters in different cases. The specific data details are shown in Table 6. It is shown that the improved segmentation model of residual film left in farmland not only has certain accuracy but also has real-time performance.

The improved DeeplabV3+ further improves the performance of DeeplabV3 in the field of semantic segmentation by introducing new modules and techniques for the task of target segmentation of agricultural residual film leftovers in complex farmland environment scenarios. By introducing a multi-scale image input method and fusing features at multiple resolutions, the feature fusion methods used include feature summation, feature average pooling, and feature adaptive fusion. To improve segmentation accuracy, the

TABLE 7. Detection and segmentation results of residual film left over from farmland example.

YOLOv7-SPD test results	Improved DeeplabV3+ model example segmentation results	Number of Pixel Points	Area of residual film（dm^2）
		165757	1.26463
		66654	0.50853
		80531	0.61440
		105784	0.80707

improved DeeplabV3+ adds a cavity-separable convolution module to the ASPP module, and after the complete convolution is performed, the convolution kernel with different expansion rates is used to perform more upscaling operations on the feature integration as shown in Figure 8, and the final model can obtain richer and more accurate feature representations of the residual film left in the farmland.

D. SEGMENTATION RESULTS OF DETECTINGd RESIDUAL FILM LEFTOVER FILM EXAMPLES IN FARMLAND

The untrained farmland residual film images collected in the field from the complex farmland environment in Bazhou are fed into the proposed model for detection, recognition, segmentation, and computation to observe its ability to detect farmland residual film in real situations. After the proposed YOLOv7-SPD model, it can effectively distinguish the leftover agricultural residual film from cotton, and accurately locate and recognize the residual film, and then processed by the improved DeeplabV3+ model, each pixel is scanned and extracted sequentially from the upper-left corner of the image, and the coordinates are represented by *(xi,yi)*, The total pixels N and the area of residual film after segmentation of the residual film left in the farmland were calculated as shown below.

$$N = \sum_{x=1}^{U} \sum_{y=1}^{V} f(x, y) \tag{9}$$

$$\text{Area of residual film} = \frac{N}{3072 \times 4096} \times 0.8 \times 1.2 \tag{10}$$

3072 and 4096 are the picture pixels of farmland residual film leftovers, and 0.8 and 1.2 are the length and width occupied by the real of the captured image in meters, respectively.

The segmentation results of farmland residual film instance detection are shown in Table 7, and the results show that in the complex farmland environment, the improved model has high feature extraction capability for the detection of farmland residual film of different sizes and irregular shapes and cotton residue, and the detection accuracy is high, and the improved DeeplabV3+ model has better instance segmentation effect, which is enough to accurately segment the boundaries of the farmland residual film. and calculate the area of residual film by the number of pixel points, which makes our proposed model useful in the application scenario of residual film in

FIGURE 8. Comparison of Residual Film Segmentation Characteristics. (a) Original Residual Film; (b) Original DeeplabV3+ model segmentation effect;(c) Improved DeeplabV3+ model segmentation effect.

farmland, realizes the accuracy of detection, and provides technical support for improving the recognition of residual film in farmlan

V. CONCLUSION

Aiming at the problems of complex farmland environment and low pickup rate of residual film recycling machine, this study proposes YOLOv7-SPD for the detection of residual film leftover in farmland and combines YOLOv7-SPD and DeeplabV3+ to identify, localize, segment, and compute the area of residual film leftover in farmland, which realizes the high efficiency and accuracy of the detection. The proposed YOLOv7-SPD has the highest recall and mAP values of 92.12% and 87.62%, respectively, in the detection of residual film left over from farmland, which indicates that the improvement measures in this paper have a positive effect on the model and effectively improve the recognition accuracy; in the segmentation of residual film left over from farmland. The improved DeeplabV3+ has 94.33% precision, 95.06% recall, 94.77% F1 and 91.55% mIOU on the segmentation of leftover film from agricultural residues and reduces the number of references.The experimental results demonstrate that the model exhibits excellent accuracy and real-time performance, establishing its practicality as an algorithm for effectively detecting residual film in agricultural fields. The model's ability to precisely identify and distinguish residual film contributes to efficient and reliable detection, offering valuable support for agricultural practices. This research outcome showcases the potential of the algorithm in practical applications related to residue management and field cleanliness assessment in agriculture. In future research, these datasets are not enough for complex and diverse farmland environments, we will collect more complex farmland residual film images, continue to optimize the computational volume and detection speed of the model to make the model lighter and more accurate, and develop and research the farmland residual film residue leftover film recognition and detection system to make this application more practical.

REFERENCES

[1] L. Yang, T. Heng, X. He, G. Yang, L. Zhao, Y. Li, and Y. Xu, "Spatial–temporal distribution and accumulation characteristics of residual plastic film in cotton fields in arid oasis area and the effects on soil salt transport and crop growth," *Soil Tillage Res.*, vol. 231, Jul. 2023, Art. no. 105737.

[2] S. Li, F. Ding, M. Flury, and J. Wang, "Dynamics of macroplastics and microplastics formed by biodegradable mulch film in an agricultural field," *Sci. Total Environ.*, vol. 894, Oct. 2023, Art. no. 164674.

[3] D. Wang, Y. Xi, X.-Y. Shi, Y.-J. Zhong, C.-L. Guo, Y.-N. Han, and F.-M. Li, "Effect of plastic film mulching and film residues on phthalate esters concentrations in soil and plants, and its risk assessment," *Environ. Pollut.*, vol. 286, Oct. 2021, Art. no. 117546.

[4] K. Koskei, A. N. Munyasya, Y.-B. Wang, Z.-Y. Zhao, R. Zhou, S. N. Indoshi, W. Wang, W. K. Cheruiyot, D. M. Mburu, A. B. Nyende, and Y.-C. Xiong, "Effects of increased plastic film residues on soil properties and crop productivity in agro-ecosystem," *J. Hazardous Mater.*, vol. 414, Jul. 2021, Art. no. 125521.

[5] J. W. Qiao, "The operating quality of retrieving machines for film residue," *Farm Products Process. Agricult.*, vol. 14, pp. 97–100, Jul. 2022.

[6] S. Q. Jiang, H. D. Zhang, and Y. J. Hua, "Research on location of residual plastic film based on computer vision," *J. Chin. Agricult. Mechanization*, vol. 37, no. 11, pp. 150–154, 2016.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.

[8] A. C. Berg, C. Y. Fu, C. Szegedy, D. Anguelov, D. Erhan, S. Reed, and W. Liu, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 21–37.

[9] M. Tan, R. Pang, and Q. V. Le, "EfficientDet: Scalable and efficient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10778–10787.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Maliket, "Rich feature hierarchies for accurate object detection andsemantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[11] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, Oct. 2015, pp. 234–241.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," 2014, *arXiv:1412.7062*.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, and L. Joneset, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017. [Online]. Available: https://doi.org/10.48550/arXiv.1706.03762

[17] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 12077–12090.

[18] Z. Zhai, X. Chen, R. Zhang, F. Qiu, Q. Meng, J. Yang, and H. Wang, "Evaluation of residual plastic film pollution in pre-sowing cotton field using UAV imaging and semantic segmentation," *Frontiers Plant Sci.*, vol. 13, Sep. 2022, Art. no. 991191.

[19] F. Qiu, Z. Zhai, Y. Li, J. Yang, H. Wang, and R. Zhang, "UAV imaging and deep learning based method for predicting residual film in cotton field plough layer," *Frontiers Plant Sci.*, vol. 13, Oct. 2022, Art. no. 1010474.

[20] X. J. Zhang, S. Huang, W. Jin, J. S. Yan, Z. L. Shi, X. C. Zhou, and C. S. Zhang, "Identification method of agricultural film residue based on improved faster R-CNN," *J. Hunan Univ. Natural Sci.*, vol. 48, no. 8, pp. 161–168, 2021.

[21] T. Zhou, Y. Jiang, X. Wang, J. Xie, C. Wang, Q. Shi, and Y. Zhang, "Detection of residual film on the field surface based on faster R-CNN multiscale feature fusion," *Agriculture*, vol. 13, no. 6, p. 1158, May 2023.

[22] J. J. Xing, D. J. Xie, R. B. Yang, X. R. Zhang, W. B. Shun, and S. B. Wu, "Lightweight detection method for farmland waste based on YOLOv5s," *Trans. Chin. Soc. Agricult. Eng.*, vol. 38, no. 19, pp. 153–161, 2022.

[23] J. Chen, J. Zhu, Z. Li, and X. Yang, "YOLOv7-WFD: A novel convolutional neural network model for helmet detection in high-risk workplaces," *IEEE Access*, vol. 11, pp. 113580–113592, 2023, doi: 10.1109/ACCESS.2023.3323588.
[24] Y. Lin, J. Zhang, Z. Jiang, and Y. Tang, "YOLOv5-atn: An algorithm for residual film detection in farmland combined with an attention mechanism," *Sensors*, vol. 23, no. 16, p. 7035, Aug. 2023.
[25] Y. Yan, Y. Gao, L. Shao, L. Yu, and W. Zeng, "Cultivated land recognition from remote sensing images based on improved deeplabv3 model," in *Proc. China Autom. Congr. (CAC)*, Xiamen, China, Nov. 2022, pp. 2535–2540.
[26] Z. Wu, Y. Tang, B. Hong, B. Liang, and Y. Liu, "Enhanced precision in dam crack width measurement: Leveraging advanced lightweight network identification for pixel-level accuracy," *Int. J. Intell. Syst.*, vol. 2023, pp. 1–16, Sep. 2023.
[27] F. Meng, J. Li, Y. Zhang, S. Qi, and Y. Tang, "Transforming unmanned pineapple picking with spatio-temporal convolutional neural networks," *Comput. Electron. Agricult.*, vol. 214, Nov. 2023, Art. no. 108298.
[28] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainablebag-of-freebies sets new state-of-the-art for real-time object detectors," 2022, *arXiv:2207.02696*.
[29] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.
[30] J. Wang, C. Xu, W. Yang, and L. Yu, "A normalized Gaussian Wasserstein distance for tiny object detection," 2021, *arXiv:2110.13389*.
[31] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," 2021, *arXiv:2103.02907*.
[32] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
[33] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*. Cham, Switzerland: Springer, 2018, doi: 10.1007/978-3-030-01234-2_1.
[34] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.
[35] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, "EPSANet: An efficient pyramid squeeze attention block on convolutional neural network," in *Proc. Asian Conf. Comput. Vis.*, Dec. 2022, pp. 1161–1177.
[36] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4002–4011.

DEQI HUANG received the B.S. degree from the Huazhong University of Science and Technology, and the master's degree in computer science and the Ph.D. degree in control science and engineering from the Wuhan University of Technology. He is currently an Assistant Professor with the College of Electrical Engineering, Xinjiang University. His research interests include machine learning, intelligent transportation systems, and image processing.

YANGTING ZHANG was born in Xinjiang, China, in 1999. She received the bachelor's degree in engineering. She is currently pursuing the master's degree in electronic information with Xinjiang University. Her research interests include pattern recognition intelligent systems and computer vision.