Research on the Detection of Stellera Chamaejasme Flower Based on Deep Learning

Jieteng Jiang, Shuo Dong, Chunmei Li* and Yihan Ma
Department of Computer Technology and Application, Qinghai University, Xining, Qinghai, 810016, China

Keywords: Deep Learning, YOLOv3-SPP, Faster_Rcnn, SSD, Target Detection, Grassland Degradation.

Abstract: Under the dual impact of global climate change and human activities in recent decades, the grassland vegetation in the Three-River Source area is seriously degraded, and accurate evaluation of grassland is the primary condition for ecological protection. Therefore, using intelligent means to evaluate grassland is the first step in ecological protection. In this paper, the most widely used target detection algorithms Faster-RCNN, SSD and Yolov3-SPP are used to detect the degradation indicator grass species of Stellera chamaejasme flower. The experimental results are compared and analyzed, and the characteristics of the three target detection algorithms and their performance in the detection of degraded indicator grass species are discussed.

1 INTRODUCTION

In terms of target detection, The Region-based Convolutional Neural Network (RCNN) (He, Zhang, Ren, Sun 2016) successfully connects target detection and deep Convolutional network, and improves the accuracy of target detection to a new level. RCNN consists of 3 independent steps: candidate window generation, feature extraction, SVM classification and window regression. RCNN mainly uses the Selective Search method to generate many candidate windows. Then all the generated candidate windows are sent to the deep network at once to extract features. Finally, the SVM classifier is trained to classify all candidate windows and window regression. Since RCNN is divided into 3 independent processes, the detection efficiency is very low. Based on this situation, scholars have improved RCNN and proposed a scale Spatial Pyramid Pooling Net (SPPNet) and Fast Region Based Convolu- tional Neural Network (Fast-RCNN) (He, Zhang, Ren, Sun 2015). It does not send all the candidate windows to the network, just send the image to the deep network once, and then map all the candidate windows on a certain layer in the network, which greatly improves the detection speed of the model. Fast-RCNN (Girshick 2015) uses the candidate window network (Region Proposal Network, RPN), and generates candidate windows, useing the same structure as Fast-RCNN for classification and window regression. Faster-RCNN combines target detection into a unified deep network framework. Region Based Fully Convolutional Network (RFCN) (Ren, He, Girshick, Sun 2017) is further improved on this basis. The analysis found that the network layer after Region of Interest (ROI) pooling no longer has translation invariance, and the number of layers after ROI pooling will directly affect the detection efficiency. Therefore, RFCN designs a position-sensitive ROI pooling layer, and directly judges the results after this pooling, which greatly improves the detection efficiency. YOLO (You Only Look One) (Dai, Li, He & Sun 2016) and SSD (Single Shot Multibox Detector) (Redmon, Divvala, Girshick, Farhadi 2016) are proposed to improve the detection efficiency of target detection, and try to make target detection reach the level of real-time detection. SSD can improve the efficiency of target detection while maintaining detection accuracy, which is a win-win algorithm in terms of detection accuracy and detection efficiency. Compared with traditional target detection methods, target detection methods based on deep networks have obvious advantages in accuracy. First of all, a neural network is a network structure with self-learning function that simulates the human brain. The forward calculation of the deep network can be regarded as a process of continuously abstracting objects. The high level of the deep network (near the

Jiang, J., Dong, S., Li, C. and Ma, Y.
Research on the Detection of Stellera Chamaejasme Flower Based on Deep Learning.
DOI: 10.5220/0011768400003607
In *Proceedings of the 1st International Conference on Public Management, Digital Economy and Internet Technology (ICPDI 2022)*, pages 824-830
ISBN: 978-989-758-620-0

output layer) records more things. Second, the deep network structure can better fit large-scale training samples. The greater difficulty of target detection lies in the variability of target objects, which have different colors, sizes and shapes in different scenes. The deep network has a large number of parameters, which makes it have a strong learning ability. A large number of training samples are conducive to activating the deep network neurons, so that it can store and analyze the state of the target object in different colors, shapes, and environments in memory. Based on the above two points, we can see the inevitability of deep networks to achieve excellent results in the field of target detection, and the accuracy of target detection is much higher than that of ordinary methods. Target recognition in a complex background is a key and difficult problem in the task of target recognition. The main interference factors of grassland degradation indicator grass species identification in a complex background are weather, visibility, and the appearance of indicator grass species and edible forage grass. Traditional target recognition tasks in complex backgrounds use the CNN network structure, and are commonly used in handwritten character recognition (Liu, Anguelov, Erhan, Szegedy, Reed &Fu, et al. 2016, Yang, Jin, Tao, Xie, & Feng 2016), face recognition, behavior recognition and crop recognition, etc., and have achieved good results. But its shortcomings are also obvious, that is, the recognition speed is slower and the hardware requirements are higher. The popular YOLO algorithm is generally used for real-time target detection tasks, and its recognition speed is faster. The improved YOLOv3 algorithm has a higher recognition accuracy and a satisfactory effect on the recognition of small targets. It is now one of the more popular target recognition task algorithms in complex backgrounds.

2 INTRODUCTION TO THE MODEL ALGORITHM USED

2.1 Yolov3-SSP

The YOLOv3-SPP detection model adds a spatial pyramid module (SPP) on the basis of the YOLOv3 algorithm model to improve deep features. The spatial pyramid module uses a scale pool, which can output features of a fixed size without considering the size of the extracted feature map, and is mainly used to replace the entire connection layer. No matter how large the input is, the output size is the same, so the last layer of feature mapping is cropped proportionally, and then input to the pooling layer to output features, which can eliminate the problem of inconsistent input image sizes. The YOLOv3-SPP network structure diagram is shown in Figure1.

The loss function of the YOLOv3-SPP detection model algorithm is the same as the YOLOv3 loss function, which consists of three parts: coordinate error $Loss_{coord}$ (center coordinate error, width and height coordinate error), confidence error $Loss_{\mathrm{conf}}$ (including target confidence error $Loss_{obj}$, without target confidence Degree error $Loss_{noobj}$) and classification error $Loss_{class}$.

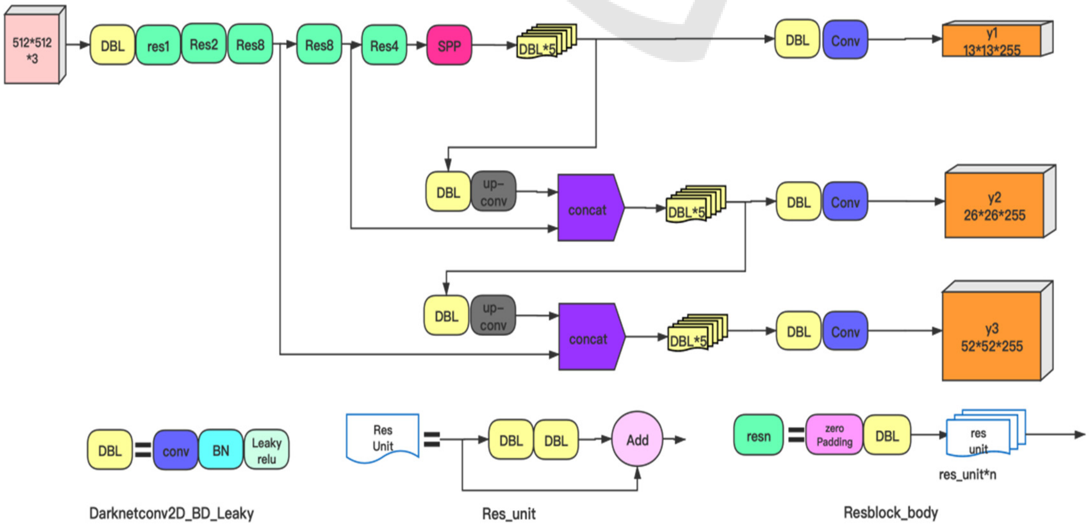


Figure 1: YOLOv3-SPP network structure diagram.

The calculation method for the center coordinate error of the prediction box is shown in formula (1).

$$\mathrm{Loss}_{\mathrm{coord}_{\mathrm{cp}}} = \lambda_{\mathrm{coord}} \sum_{\mathrm{i=0}}^{\mathrm{s}^2} \sum_{\mathrm{j=0}}^{\mathrm{B}} \mathrm{I}_{\mathrm{ij}}^{\mathrm{obj}} [(\mathrm{x_i} - \hat{\mathrm{x}}_\mathrm{i})^2 + (\mathrm{y_i} - \hat{\mathrm{y}}_\mathrm{i})^2] \quad (1)$$

The calculation method for the width and height coordinate error of the prediction box is shown in formula (2)-(8).

$$\mathrm{Loss}_{\mathrm{coord}_{\mathrm{wh}}} = \lambda_{\mathrm{coord}} \sum_{\mathrm{i=0}}^{\mathrm{s}^2} \sum_{\mathrm{j=0}}^{\mathrm{B}} \mathrm{I}_{\mathrm{ij}}^{\mathrm{obj}} [(\sqrt{w_i^j} - \sqrt{\widehat{w}_i^j})^2 + (\sqrt{h_i^j} - \sqrt{\hat{h}_i^j})^2] \quad (2)$$

$$Loss_{coord} = Loss_{coord_{cp}} + Loss_{coord_{wh}} \quad (3)$$

$$Loss_{obj} = \lambda_{obj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} [\hat{c}_i^j \log(\hat{c}_i^j) + (1 - \hat{c}_i^j) log(1 - \hat{c}_i^j)] \quad (4)$$

$$Loss_{noobj} = \lambda_{noobj} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{noobj} [\hat{c}_i^j \log(\hat{c}_i^j) + (1 - \hat{c}_i^j) log(1 - \hat{c}_i^j)] \quad (5)$$

$$Loss_{conf} = Loss_{obj} + Loss_{noobj} \quad (6)$$

$$Loss_{class} = \lambda_{class} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \sum_{c \in classes} p_i(c) log\ (\hat{p}_i(c)) \quad (7)$$

$$Loss = Loss_{coord} + Loss_{conf} + Loss_{class} \quad (8)$$

Among them, x_i y_i w_i h_i are the abscissa, ordinate, width, and height coordinates of the real box. $\hat{x}_i$ $\hat{y}_i$ $\hat{w}_i$ $\hat{h}_i$ are the abscissa, ordinate, width and height coordinates of the prediction box. I_{ij}^{obj}:indicates that if the detection box at (i, j) has a target, its value is 1, otherwise it is 0; I_{ij}^{noobj}: indicates that if the detection frame at (i, j) has no target, its value is 1, otherwise it is 0; $\hat{c}_i^j$: indicates the predicted value; B is the detection box; S is the grid size; $\hat{p}_i^j$:represents the accuracy rate of each category in the j-th prediction box of the i-th grid.

2.2 Faster RCNN

The difference between the Faster RCNN target detection algorithm used in this article and the previous RCNN series algorithms is that Faster RCNN has integrated feature extraction, candidate region extraction, bounding box regression, and category classification into one network, which greatly improves the comprehensive performance, especially in terms of detection speed.

Faster RCNN uses RPN network (Region Proposal Network) instead of selective search to recommend candidate regions. It can be trained end-to-end for the task of generating candidate detection boxes, and can predict the boundary and score of the target at the same time. The input is a picture, and the output is multiple candidate regions. The network structure of the Faster RCNN detection algorithm is shown in Figure 2.

The loss function of Faster RCNN includes two parts, namely regression loss $\mathrm{Loss_{reg}}$ (RPN position regression loss: anchor position fine-tuning, ROI position regression loss: continue to fine-tune ROI position) and classification loss $\mathrm{Loss_{cls}}$ (ROI classification loss: ROI category, RPN classification loss: whether anchor is gt). The loss function expression is shown in formula (9)- (13).

$$Loss_{cls} = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i{}^*) \quad (9)$$

$$L_{cls}(p_i, p_i{}^*) = -log\ (p_i p_i{}^* + (1 - p_i{}^*)(1 - p_i)] \quad (10)$$

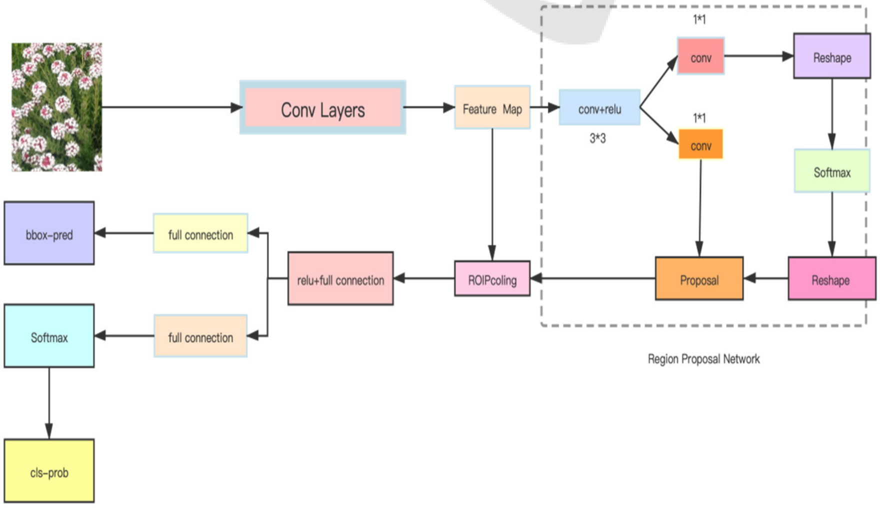


Figure 2: Faster RCNN network structure diagram.

$$Loss_{reg} = \frac{1}{N_{reg}} \sum_i p_i{}^* L_{reg}(t_i, t_i{}^*) \qquad (11)$$

$$L_{reg}(t_i, t_i{}^*) = R(t_i - t_i{}^*) \qquad (12)$$

$$p_i{}^* = \begin{cases} 0 & ,negative\ label \\ 1 & ,positive\ label \end{cases} \qquad (13)$$

Among them, p_i is the probability that the anchor point prediction is the target; N_{cls} is the minimum batch; $p_i{}^*$ is the true value, $t_i = \{t_x, t_y, t_w, t_h\}$ is the coordinate vector of the box corresponding to the positive anchor point, $\mathrm{L_{reg}(t_i, t_i{}^*)}$ is the regression loss, R is Smooth L1 function, $\mathrm{L_{cls}(p_i, p_i{}^*)}$ is the classification loss of the two categories (target vs. non-target).

Faster RCNN is a classic representative two-stage target detection algorithm. After continuous improvement and perfection, the detection accuracy is higher than the previous Fast RCNN target detection algorithm and single-stage target detection algorithms such as YOLO series target detection algorithms. The shortcomings of the Faster RCNN target detection algorithm are particularly obvious. First of all, it cannot detect targets in real time. For the task of detecting targets in real time at this stage, this is not enough. Second, the calculation of the FasterRCNN target detection algorithm is more complicated than that of the YOLO series of target detection algorithms.

2.3 SSD

In 2016, the single-shot detector (SSD) network model was proposed by Wei Liu et al.The algorithm is based on the single-stage target detection algorithm of deep learning. The SSD target detection algorithm uses multi-scale fusion to improve the detection accuracy and solve the problem of insufficient detection accuracy of YOLO. The SSD detection speed is also better than the two-stage target detection algorithm in the same period. Its main idea is to sample densely and uniformly at different positions of the image, and it borrows from the on the concept of Anchor in Faster RCNN. When sampling, the predicted target bounding box is passed through a priori boxes of different scales and aspect ratios, and then CNN extracts the features, and then classifies and regresses them directly. SSD uses a pyramid-structured feature layer group for target detection. The SSD network structure diagram is shown in Figure 3.

The SSD loss function is composed of the position loss of the corresponding search box and the category confidence loss. The specific loss function expression is shown in formula (14)- (21).

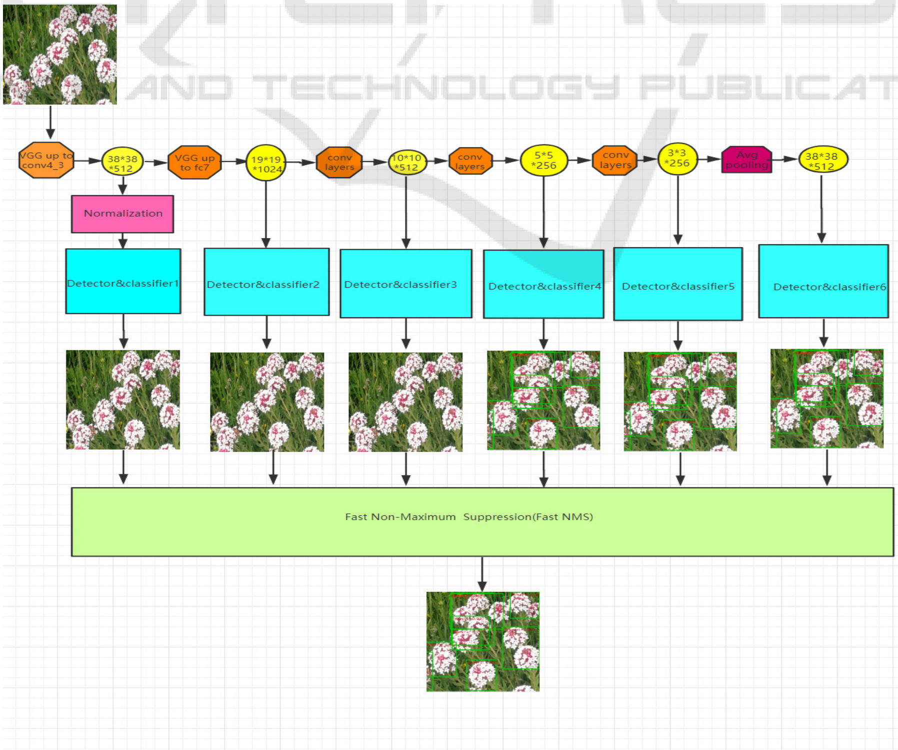


Figure 3: SSD network structure diagram.

$$\text{Loss}(x,c,l,g) = \frac{1}{N}(Loss_{conf}(x,c) + \alpha Loss_{loc}(x,l,g)) \quad (14)$$

$$Loss_{loc}(x,l,g) = \sum_{i\in Pos}^{N} \sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k} smooth_{L1}(l_i^m - \hat{g}_j^m) \quad (15)$$

$$\hat{g}_j^{cx} = (g_j^{cx} - d_i^{cx})/d_i^w \quad (16)$$

$$\hat{g}_j^{cy} = (g_j^{cy} - d_i^{cy})/d_i^h \quad (17)$$

$$\hat{g}_j^{w} = \log\left(\frac{g_j^w}{d_j^w}\right) \quad (18)$$

$$\hat{g}_j^{h} = \log\left(\frac{g_j^h}{d_j^h}\right) \quad (19)$$

$$Loss_{conf}(x,c) = -\sum_{i\in Pos}^{N} x_{ij}^{p} \log(\hat{c}_i^p) - \sum_{i\in Neg} \log(\hat{c}_i^0) \quad (20)$$

$$where\ \hat{c}_i^p = \frac{exp(c_i^p)}{\sum_p exp(c_i^p)} \quad (21)$$

Among them, i refers to the serial number of the search box, j refers to the serial number of the real box, p refers to the category serial number, and p=0 represents the background. $\mathrm{x}_{\mathrm{ij}}^{\mathrm{p}}$： the predicted box i and the real box j match with respect to the category p. The higher the probability prediction of p, the smaller the loss. $\hat{\mathrm{c}}_{\mathrm{i}}^{\mathrm{p}}$： The i-th search box corresponds to the predicted probability of category p.

3 COMPARATIVE EXPERIMENT ON TARGET DETECTION OF STELLERA CHAMAEJASME FLOWER

The YOLOv3-SPP target detection algorithm based on the Pytorch deep learning framework is implemented in the grass degraded indicator grass species Stellera chamaejasme flower. It is proved that the target detection of the grassland degradation indicator grass species based on the convolutional neural network is feasible. This chapter uses the Faster RCNN target detection algorithm and the SSD target detection algorithm to detect the Stellera chamaejasme flower, and the detection results of the two algorithm models are compared and analyzed with the Yolov3-SPP algorithm model to explore the target detection algorithm suitable for this data.

3.1 Data Preparation

In this chapter, the data set of the comparison experiment based on the convolutional neural network for the detection algorithm of the Stellera chamaejasme flower is the self-built detection data set of the Stellera chamaejasme flower. The training set contains 18,000 pictures, which is used for the training of the convolutional neural network. The test set contains 2000 pictures, which is used for the evaluation work of the Stellera chamaejasme flower target detection model. In order to observe the model training work better and more intuitively, 200 images format data are set as a validation set for model evaluation during the training.

3.2 Experimental Setup

The operating system used in this experiment is Ubuntu 18.04, the hardware configuration is 8 GeForce GTX 1080Ti GPUs (16G memory), the integrated development tool PyCharm is used, the Python version is 3.6, the VNC Viewer remote login software. The network development framework used is Pytorch The target detection model is YOLOv3-SPP detection model, Faster RCNN detection model and SSD detection model, and the feature extraction network uses VGG16.The number of detection categories is set to the number of categories that need to be detected, which is 1. The value of the learning rate is 0.001, the Batch-Size size of the training network model is 16, and the number of training rounds of the model is set to 100. The training parameters, models and training logs are saved during the training process of the model, so as to better and more intuitively monitor the network training results during the training process of the detection model.

3.3 Experimental Process

The contrast test process of the Stellera chamaejasme flower detection algorithm based on the convolutional neural network is as follows:

- YOLOv3-SPP, SSD, Faster RCNN all use the VOC 2007 format degradation indicator grass species target detection data set, according to the parameter settings of each algorithm of this experiment.
- Call the processed data training set for model training of the three algorithms.
- Perform model evaluation on the YOLOv3-SPP, SSD, Faster RCNN model trained with this data set, and input 1200 600*800 images used for the target detection of the Stellera chamaejasme flower target into the trained Among the three models. And develop a unified evaluation standard.
- Select 200 600*800 images to be detected outside of the selected training data set and test data and input them into the detection model to obtain the result images that have

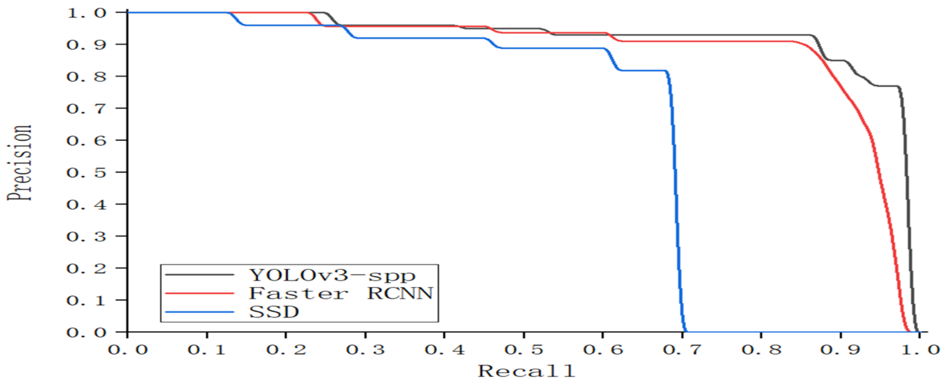
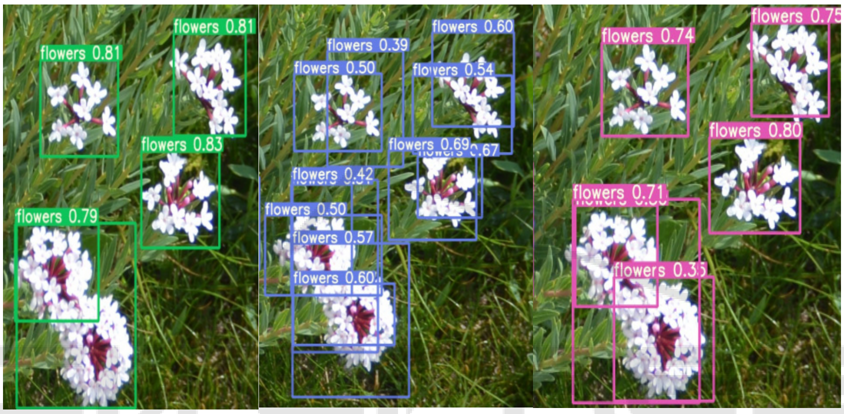


Figure 4: The PR curve of the comparison test of the target detection of the Stellera chamaejasme flower.



YOLOv3-SPP SSD Faster RCNN

Figure 5: Comparison of model detection effects.

- been detected by the YOLOv3-SPP model, the SSD model, and the Faster RCNN model.
- According to the comparison of the obtained evaluation indicators and the comparison of the results of the same detection pictures, analyze the experimental phenomenon and get the reasons for the experimental results.

3.4 Analysis of Experimental Results

The size of the test set is 2000 pictures with a size of 600*800 and the corresponding annotation files of each picture. After the model evaluation of the test set, the test results of YOLOv3-SPP, SSD, FasterRCNN are shown in Figure 4.

Since there is only one detection category, the mAP value is the AP value, which can be known by analyzing the above test results that the accuracy, recall, and ap values of YOLOv3-spp are higher than those of SSD, Faster_rcnn algorithm. It can be seen that in the detection task of degenerative indicator grass species, the detection effect of YOLOv3 is more accurate than that of SSD and Faster_rcnn algorithm. Because the same size of the image is passed, the clarity of the detected image will be affected by resize, which in turn affects the training effect of the detection model. However, YOLOv3-ssp has more SSP layers than SSD and faster-rcnn networks, which makes the pictures of the input network clearer, which in turn makes the algorithm model of YOLOv3-spp better than the other two algorithm models. Input 200 images to be detected with a size of 600*800 to the YOLOv3-spp, faster_rcnn, ssd network, and the detection results are shown in the following figure 5.

It can be seen from the comparison graph of the detection results that the yolov3-spp network adds the spp layer, which makes the model training more complete. Yolov3-spp compared with faster_rcnn, the edge detection effect of the ssd algorithm is better. However, the three detection models all have defects. If the flower type of the Stellera chamaejasme flower is poor, the detection effect is poor. The main reason is that the labeling of the data set is not perfect, the amount of data for the poor pattern of the Stellera chamaejasme flower is small, and the color of the

Stellera chamaejasme flower is closer to the background color in a complex background, which means that the detection effect is poor.

4 CONCLUSION

In this paper, three algorithms, namely Faster-RCNN, SSD and Yolov3-SPP, were used to detect the degradation indicator grass species of Stellera chamaejasme flower, and the experimental results were compared and analyzed to discuss the characteristics of the three target detection algorithms and their performance in the detection of degradation indicator grass species. Finally, it is concluded that the Yolov3-SPP algorithm is superior to the Faster-Rcnn and SSD.

REFERENCES

Dai, J. , Li, Y. , He, K. , & Sun, J.(2016) R-fcn: object detection via region-based fully convolutional networks.In: Advances in Neural Information Processing Systems. Cambridge.pp. 379-387.

Girshick,R., Donahue,J., Darrell,T., Malik,J.(2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington. pp. 580-587

Girshick,R.(2015)FastR-CNN.In:Proceedings of the IEEE International Conference on Computer Vision. Washington. pp. 1440-1448.

He,K.,Zhang,X.Y.,Ren,S.Q.,Sun,J. (2016) Deep Residual Learning for Image Recognition.In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).Washington.pp. 770-778.

He,K.,Zhang,X.Y.,Ren,S.Q.,Sun,J.(2015)Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. In:IEEE Transactions on Pattern Analysis and Machine Intelligence . Washington.37(9):1904-1916.

Liu,W., Anguelov ,D., Erhan, D., Szegedy, C., Reed, S., & Fu, C.Y., et al. (2016) SSD: Single Shot MultiBox Detector.I n: European Conference on Computer Vision. New York.pp. 21-37.

Ren,S.Q., He,K.M., Girshick,R.,Sun,J.(2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks.In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Washington. pp.1137-1149.

Redmon, J., Divvala, S.,Girshick ,R., Farhadi ,A.(2016) You Only Look Once: Unified， Real-Time Object Detection.In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Washington.pp. 779-788.

Yang, W., Jin,L., Tao, D.,Xie, Z.,& Feng,Z. (2016) Dropsample: a new training method to enhance deep convolutional neural networks for large-scale unconstrained handwritten chinese character recognition. Pattern Recognition, 58:190-203.