American Journal of Biochemistry and Biotechnology

Original Research Paper

Biformer Attention and ASF-YOLO for Cordyceps Sinensis Target Recognition

[1]Ru Yang, [2]Peng Wu and [2]Zhentao Qin

[1]*School of Civil and Architecture Engineering, Panzhihua University, Panzhihua, China*
[2]*School of Mathematics and Computer Science, Panzhihua University, Panzhihua, China*

Article history
Received: 11-06-2024
Revised: 04-09-2024
Accepted: 18-09-2024

Corresponding Author:
Zhentao Qin
School of Mathematics and Computer Science,
PanzhihuaUniversity,
Panzhihua, China
E-mail: 309507443@qq.com

Abstract: Cordyceps sinensis, a highly valued traditional Chinese medicine, faces challenges in collection due to inefficiencies in manual searching, strenuous labor, and the impact of subjective expertise. The integration of deep learning into Cordyceps sinensis identification is an unexplored area. To alleviate the manual labor and enhance the precision and speed of identifying Cordyceps sinensis, a novel detection approach that combines attention mechanisms with the ASF-YOLO model has been developed. This approach replaces the Spatial Pyramid Pooling Fast (SPPF) with a Context Augmentation Module (CAM) and swaps the original C3 model with a lighter model, C3-Faster, which is based on FasterNet. Additionally, it incorporates the Bi-level Routing Attention (BiFormer) mechanism and a Context Integration module to better detect smaller targets and increase accuracy. For the detection of tiny Cordyceps sinensis targets against intricate backgrounds, a novel fusion framework, ASF-YOLO, which leverages attention scale sequence fusion, has been introduced to boost detection accuracy further. Through experimental verification, the average accuracy rate (MAP) for Cordyceps sinensis can reach 99.2, 0.6% higher than that of traditional YOLOv5. The enhanced YOLOv5 boasts an average detection accuracy of up to 95% and it could identify some cordyceps sinensis that could not be identified by traditional YOLOv5.

Keywords: Identification of Cordyceps, Attention Mechanism, ASF-YOLOs, Deep Learning

Introduction

Cordyceps sinensis, recognized for its significant therapeutic properties in traditional Chinese medicine, is typically sought through laborious and inefficient manual processes. As deep learning technology evolves, especially the application of the YOLO series algorithm in image recognition, new ideas have been provided for the automatic detection of cordyceps. As a member of the YOLO series, YOLOv5 has become a hotspot in this field for its fast and accurate characteristics, providing a new method for the automated search of Cordyceps. However, due to the small size of cordyceps, the area where cordyceps is located is usually complex. When using the traditional YOLOv5 algorithm to detect small targets in complex areas, low detection accuracy, weak robustness, and limited recognition ability would occur. To solve these problems, Wang and Zhang (2023) introduced advanced background suppression techniques, such as attention mechanisms and feature enhancement techniques, to improve the adaptability and robustness of the algorithms for complex backgrounds. However, the detection speed was slow. The Feature Pyramid Network (FPN) can markedly enhance the detection capabilities for minor targets, however, this progress leads to heightened computational requirements and complexity within the model's architecture (Lin *et al*., 2017). Deformable Convolutional Networks(DCN) allow for adaptive adjustment of the shape of the kernel size tailored to the contours of the targets, thereby improving the detection ability of irregular minor targets, but adding additional calculations (Dai *et al*., 2017). Wang *et al*. adopted multi-scale feature fusion in High-Resolution Network (HRNet) to enhance the detection of minor targets, yet introduce greater model complexity and substantial computational demands (Wang *et al*., 2022).

Based on these studies, a new target recognition method that integrates the Biformer attention mechanism and the ASF-YOLO model is proposed. The inclusive Context Enhancement Module (CAM) is used to extract context information from different receiving domains using extended convolution and then integrate it into FPN to improve the context information of small objects. The streamlined C3-Faster model is utilized to reduce the model's parameters and computational load, simplifying its structure while maintaining its detection performance. In addition, an attention mechanism and context-aggregation module of Context Aggregation are introduced to enhance the precision of detecting minor targets. At the same time, an ASF-YOLO system that employs attention-scaled sequence merging is incorporated, specially designed for detecting small targets of Cordyceps sinensis with high detection accuracy.

Materials and Methods

YOLOv5 Model

YOLOv5 is a popular object detection model, which is the fifth generation of the YOLO series. The YOLO series of algorithms is renowned for swift performance and ease of deployment. YOLOv5 has made multiple improvements and optimizations based on this to enhance the efficacy and precision of object detection (Wang *et al*., 2023).

YOLOv5 is an efficient target detection model that provides various variants according to different needs and computing resources to achieve a trade-off between rapidity and precision. The smallest model, YOLOv5s, is used and its architecture is depicted in Fig. (1).

Improved YOLOv5 Model

To enhance the precision and efficiency of the YOLOv5 model, FasterNet is adopted as the Backbone (Chen *et al*., 2023), by integrating FasterNet with the original C3 to create the C3-faster module, which then replaces the standard C3. Subsequently, the attention mechanism, known as Bi-Level Routing Attention (Zhu *et al*., 2023), is incorporated. This mechanism enables a more adaptable distribution of computational resources and enhances the model's capacity to understand visual content. The context enhancement module CAM and the third Concat fusion method (Lee *et al*., 2019) employ various dilation rates in convolutions to gather contextual data across varying receptive fields, obtain reliable semantic information, identify characteristics of diminutive objects that have limited pixel count and to bolster the detection of minor targets. In addition, a Context Integration module based on the proposal is implemented. This module can effectively summarize the overall visual context to boost the effectiveness of object detection and segmentation in images. The concept of Context Integration has been introduced by Liu *et al*. (2024). Finally, the ASF-YOLO framework, which leverages attention-based sequence fusion, is incorporated (Kang *et al*., 2024). The SSFF module is integrated to enhance the network's ability to extract features across various scales, while the TPE module amalgamates feature maps across scales to enrich the detail level. It offers supplementary data for segmenting small targets, this subsequently improves the model's ability to identify smaller objects. The architecture of the enhanced YOLOv5 model is presented in Fig. (2).

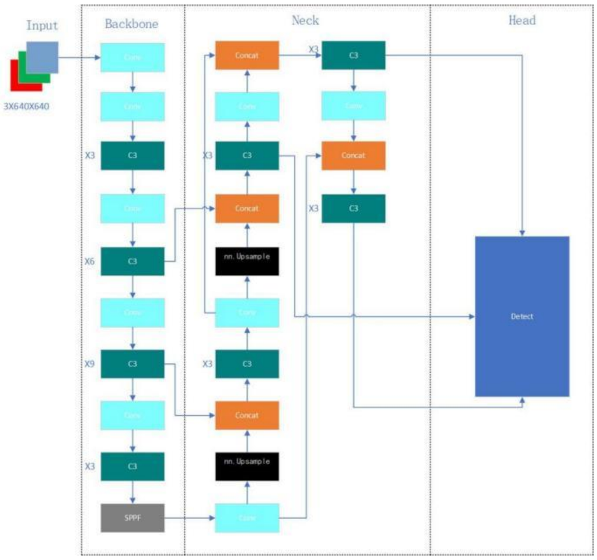

Fig. 1: Yolov5 network structure

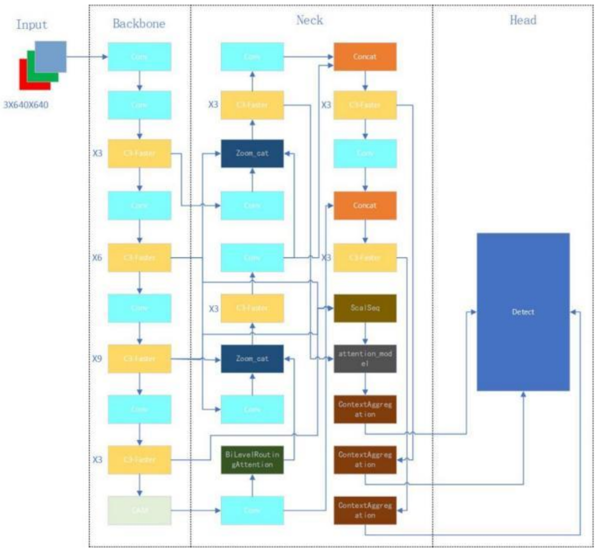

Fig. 2: Network structure of the improved YOLOv5

C3-Faster

To enhance network velocity and curtail computational demands, a new Partial Convolution (PConv) method is introduced. Based on this, FasterNet is proposed as a more efficient alternative. FasterNet, as the backbone, is integrated with C3 to substitute the original C3 module (Park *et al*., 2022; Du *et al*., 2024).

PConv requires conventional convolution to be applied to only a subset of input channels designated for spatial feature extraction, while the rest are preserved. To support orderly or sequential memory recall, either the starting or ending set of continuous channels is utilized to encapsulate the entire feature map's representation. FLOPs are:

$$h \times w \times k^2 \times c_p^2 \quad (1)$$

For a standard $r = 1/4$, the computational complexity of a PConv is just 1/16 of that of a standard Conv. Moreover, PConv requires minimal memory access, namely:

$$h \times w \times 2c_p + k^2 \times c_p^2 \approx h \times w \times 2c_p \quad (2)$$

For $r = 1/4$, it is only 1/4 of the conventional Conv h denotes the vertical dimension of the input image; w denotes the horizontal dimension of the input image; k signifies the dimensions of the convolution kernel's square and C_p denotes the channel count of the mask utilized in partial convolution and the channel count of the mask in the partial convolution operation.

FasterNet is then used as the Backbone. The architecture comprises four levels of hierarchy, each potentially starting with an embedding or a merge layer for spatial reduction and channel expansion. Each level contains a succession of FasterNet blocks, which demand fewer memory accesses and achieve higher FLOPS.

BiFormer

The BiFormer model is a vision Transformer that integrates dual-level routing attention mechanisms, which improves detection ability through a dual-layer routing attention mechanism. In the BiFormer framework, every image patch is linked to a corresponding location router. These location routers direct image patches to higher and lower levels in accordance with defined criteria. Upper-layer routing uses a rough routing policy to determine which areas are likely to contain targets. After identifying areas that may contain targets, more refined routing is carried out within those areas (Sun *et al*., 2022). This new dynamic sparse attention mechanism is realized by double-layer routing. With the dual-layer routing attention mechanism, BiFormer can dynamically allocate computing resources to the most important areas in the image, thereby reducing unnecessary computation and improving detection efficiency.

Two-layer routing attention consists of a zone-level routing step and a tag-level attention step. The central concept is to eliminate the least pertinent key-value pairs at a broader regional scale. First, a regional association graph is filtered and refined, making sure that every node retains solely the top k links. Each area should concentrate on the areas of the initial k paths. Once the area of concern is determined, attention will be applied one by one. It is an important step in assuming that key-value pairs are spatially dispersed.

Bi-level Routing Attention is implemented. First, the input feature map X is divided into G×G regions and queries P, keys U, and values N are generated by linear mapping:

$$P = X^r W^q, U = X^r W^k, N = X^r W^v \quad (3)$$

where, W^q, W^k, and W^v are the projected weights of query, key, and value respectively.

The dot product of region-level query Q^r and region-level key K^r is calculated and the correlation adjacency matrix A^r is obtained:

$$P^r = Q^r \left(K^r\right)^T \quad (4)$$

Then, the *topkIndex* is applied to identify the k most pertinent areas within the matrix A^r and generate the routing index matrix I^r for each region:

$$I^r = topkIndex\left(A^r\right) \quad (5)$$

The *topkIndex* function is used to select the largest k number from a set of values.

According to the routing index matrix I^r, the information is aggregated from the original keys K and values V to obtain the aggregated keys K^g and values V^g:

$$K^g = assemble\left(K, I^r\right), V^g = assemble\left(V, I^r\right) \quad (6)$$

Then the attention is calculated using the assembled keys K^g and values V^g, as well as the original query Q, to obtain the attention output. Finally, a local context enhancement item $GCE(V)$ is introduced and added to the attention output to obtain the final output feature O:

$$O = Attention\left(Q, K^g, V^g\right) + GCE\left(V\right) \quad (7)$$

The BiFormer architecture consists of a stack of multiple BiFormer blocks and can be configured differently depending on specific tasks and needs. Use overlapping patch embedding or patch merging modules to reduce input spatial resolution while increasing channels. BiFormer block is the basic building unit of BiFormer consisting of several sub-layers. This encompasses a self-attention module and a feedforward neural network sub-module. The self-attention module uses a two-step routing attention process to adaptively focus on the most relevant key-value pairs based on the queries. Non-linear transformation and feature extraction of attention output is carried out by multi-layer perceptron in the complex sublayer of the feedforward neural network. This combination gives BiFormer the adaptability and expressiveness to perform well in different computer vision tasks. The intricate design of the BiFormer block is depicted in Fig. (3).

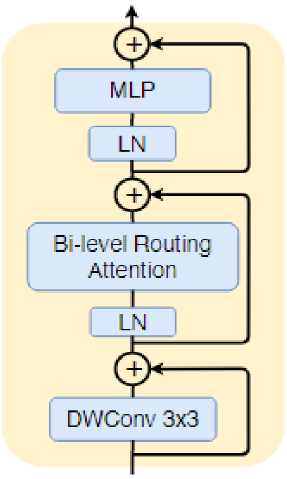

Fig. 3: Detailed structure diagram of Biformer block

CAM

CAM is a module that combines enhanced context information to aid in the identification of minor objects. CAM employs dilated convolutions with varying dilation rates to capture context across different receptive fields and incorporates a top-down FPN to enhance contextual richness (Li *et al*., 2023). The structure of CAM is shown in Fig. (4). To capture semantic details across various receptive fields, convolution is carried out on C5 with different expansion rate void convolution. The nuclear size is 3×3 and the expansion rate is 1, 3, 5.

Three fusion strategies are considered: Summation Fusion, Adaptive Fusion, and Concatenation Fusion. Here, the Concatenation Fusion approach is selected. Its structure is shown in Fig. (5).

Context Integration

Context Integration Block is a building block used in deep learning models to enhance feature representation by aggregating global context information from images. It functions to seize the overarching spatial context at each tier of the feature pyramid. The design of the Integration Block allows it to function as a modular addition that seamlessly fits into current convolutional neural network frameworks. Context Integration blocks can help address performance degradation due to scaling changes, low contrast, and cluster distribution. In this way, the Context Integration Block can provide richer context information to the model, thus enhancing the precision and dependability of segmentation tasks and bolstering the detection of minor targets. Fig. (6) is the implementation process diagram context integration block.

ASF-YOLO

The Attentional Scale Sequence Fusion based You Only Look Once (ASF-YOLO) is a deep learning model for cell instance segmentation. It is based on the YOLO framework for extension and improvement (Yang and Qiu, 2024). ASF-YOLO is composed of two principal elements: a Scale Sequence Feature Fusion component and a Triple Feature Encoder component which can offer supplementary details for the segmentation of minor objects. Then, using the channel attention mechanism, the feature information of the SSFF and TFC modules is fused to further improve the accuracy of instance division. The layout of the ASF-YOLO is depicted in Fig. (7).

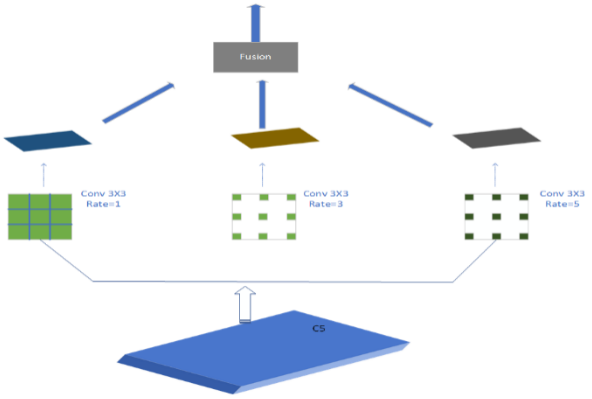

Fig. 4: Structure diagram of CAM

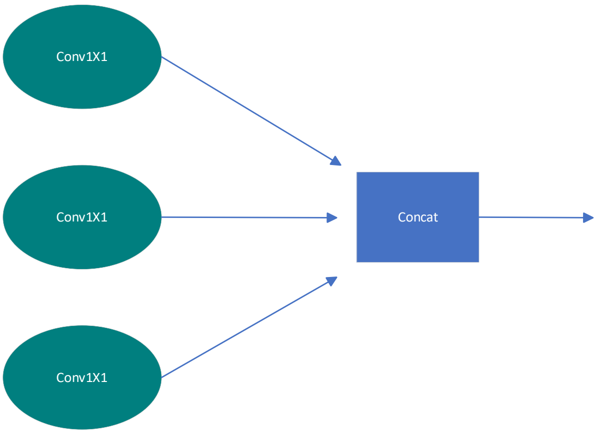

Fig. 5: Structure diagram of concatenation fusion

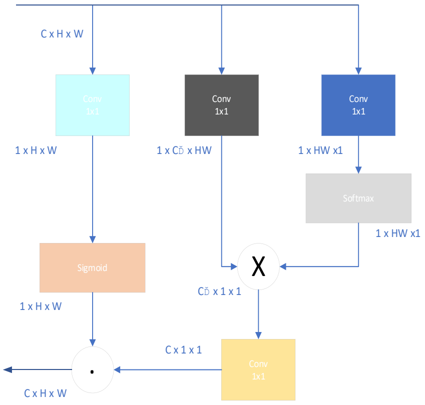

Fig. 6: Implementation process diagram of context integration block

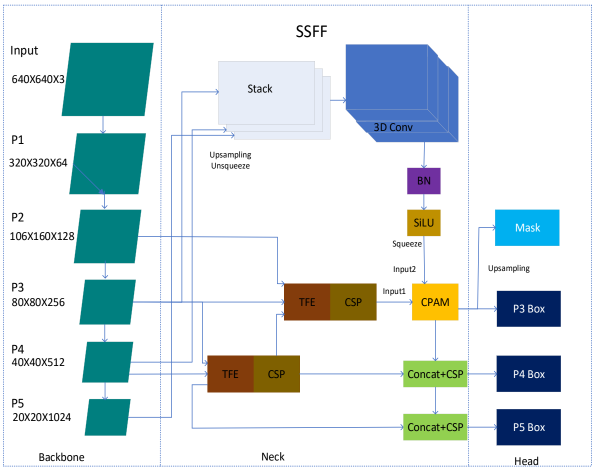

Fig. 7: Overall structure of ASF-YOLO

SSFF: The SSFF unit represents an innovative approach to scale sequence feature fusion, adept at the intricate details from deep feature maps combined with the granular details from shallow feature maps. Despite variations in image dimensions during down sampling, the scale-invariant features remain constant. Its purpose is to address multi-scale challenges by consolidating feature maps across scales, thereby improving the detection of targets of varying dimensions and forms. The schematic diagram of the SSFF unit's operation is depicted in Fig. (8).

Implementation of SSFF module. (1). Extract 1×1 convolution from the backbone network at scales P4 and P5 and adjust the channel count to 256, followed by upsampling using nearest neighbor interpolation to adjust it to the same size as the P3 level; (2). Employ the unsqueeze function to expand each feature dimension in a 3D tensor to a 4D tensor; (3). Merge the adjusted 4D feature maps along the channel dimension to create a 3D feature map for subsequent convolutional processing; (4). Employing 3D convolution, 3D batch normalization, and SiLU activation to perform scale sequence feature extraction.

Scale-space constructs are used to generate the images of different scales, which are then used as inputs to the SSFF module. The scaled image as input for SSFF can be obtained as follows:

$$F_o(\mathrm{a},b) = G_o(a,b) \times f(a,b) \tag{8}$$

$$G_\sigma(a,b) = \frac{1}{2\pi\sigma^2} e^{-(a^2+b^2)/2\sigma^2} \tag{9}$$

where, $f(a,b)$ represents a 2D input image with a width of a and a height of b; $f_o(a,b)$ is generated by a series of convolutions using a 2D Gaussian filter $G_\sigma(a,b)$ andσ is the standard deviation of a 2D Gaussian filter used for convolution.

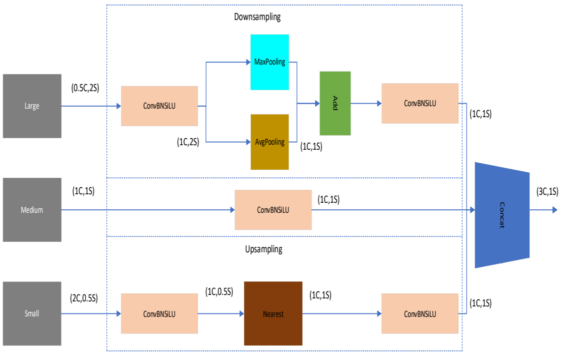

Fig. 8: TFE module implementation structure diagram

TFE: This is used to enhance small target detection by processing feature maps of varying scales to capture fine details of minor targets (Hui *et al*., 2024).

Implementation of TFE module. (1). It first accepts feature maps of various scales large, medium, and small as input. (2). For compact feature maps, convolutional layers are utilized to normalize the channel dimensions as well, followed by nearest-neighbor up-sampling to retain local features richness of low-resolution images and prevent the loss of small target feature information; (3). Stitch the modified large, medium, and small feature maps along the spatial dimension to create three feature maps of identical dimensions; (4). Perform a convolution on the concatenated feature graph and then concatenate it on the channel dimension to obtain the output feature map of the TFE module. The resultant feature map from the TFE module can be obtained by the following methods:

$$F_{TFE} = Concat(F_l, F_m, F_s) \tag{10}$$

where, F_l, F_m, and F_s respectively represent large, medium, and tiny feature maps. F_{TFE} is obtained by concatenating F_l, F_m, and F_s, with a resolution identical to that of F_m and with three times the number of channels.

The implementation of the TFE module's configuration is illustrated in Fig. (8).

CPAM: The CPAM module aims to extract representative feature information from different channels, fuse fine-grained feature details with multi-scale feature data, and extract more representative feature information. The CPAM module includes a channel attention network and a location attention network. The channel attention network receives the feature map (Input1) from the TFE module and uses the channel attention block of the Squeeze and Excitation Network (SENet) to generate channel weights. Feature images are output (Input2) from the channel network and Scale Sequence Feature Fusion (SSFF) module, overlaid with positional annotations, to extract key positional information for each Chinese caterpillar fungus.

Implementation of CPAM Module

Channel attention: 1D convolution is used to capture local channel interactions after global average pooling without dimensionality reduction. Where *k* indicates the scope of local interactions across channels, meaning the count of neighboring channels engaged in the attention prediction for a given channel. *k* has the following relationship with the number of channels *C*:

$$C = \psi(k) = 2^{(\gamma \times k - b)} \tag{11}$$

To facilitate greater layers with a greater number of channels, the dimension of the one-dimensional convolution kernel is adjusted using a defined function to influence cross-channel interactions. This kernel size denotes:

$$k = \Psi(C) = \left| \frac{log_2(C)) + b}{Y} \right|_{odd} \tag{12}$$

Odd: It takes an odd number. γ is set to 2 and bto 1.

The result is merged with the features from the SSFF (Input 2) to serve as input to the location network.

Location attention: Retention of spatial structure information of feature maps by pooling in horizontal (pw) and vertical (ph) directions:

$$p_w = \frac{1}{H}\,_{0 \le j \le H} E(w, j)$$

$$p_h = \frac{1}{W}\,_{0 \le i \le W} E(i, h) \tag{13}$$

Here, *w* and h denote the horizontal and vertical dimensions of the input feature map, respectively; *E*(*w*,*j*) and *E*(*i*,*h*) correspond to the values at the coordinates (*i*,*j*) within the input feature map.

The position attention indices are derived from concatenation and convolution operations and are represented as follows:

$$P(a_w, a_h) = Conv[Merge\,(p_w, p_h)] \tag{14}$$

where Conv stands for 1×1 convolution and Merge stands for convolution.

Segmentation of attention features to generate location-dependent feature map pairs. Feature map pairs are represented as follows:

$$s_w = Split(a_w) \tag{15}$$

$$s_h = Split(a_h) \tag{16}$$

S_w and S_h the width and height of the split output respectively.

The final output of CPAM:

$$F_{CPAM} = E \times s_w \times s_h \tag{17}$$

where, E denotes the channel and spatial attention weights, which is shown in Fig. (9).

The experiment selected the Cordyceps sinensis identification area in Xiangcheng County, which is located in the western part of Sichuan Province, China, on the southeastern edge of the Qinghai-Tibet Plateau, at the southwestern extremity of Ganzi Prefecture in Sichuan Province, in the middle-northern section of the Hengduan Mountains, spanning from 99°22′ to 100°04′ East longitude and from 28°34′ to 29°39′ North latitude. The dataset contains 525 images of cordyceps and is split into training and testing subsets at a ratio of 8:2.

The environment configuration for this experiment is as follows:

Intel(R)Core(TM)i7-8570H CPU, 2.20 GHz, NVIDIA GTX1050Ti, 16.0 GB, Windows 10, Python, and Pytorch.

Results and Discussion

Experimental Verification

This experiment used a self-made labeled cordyceps dataset, which was enhanced by adding noise, adjusting brightness, rotating, translating, and flipping. The dataset contains 525 images of cordyceps and is split into training and testing subsets at a ratio of 8:2. Part of the sample diagram is shown in Fig. (10).

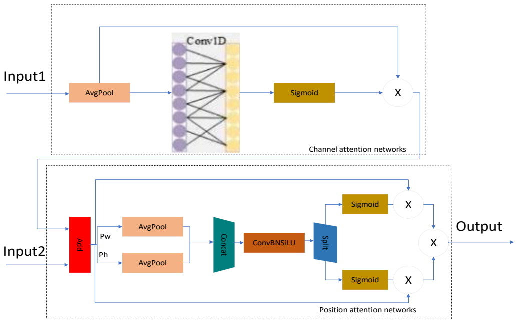

Fig. 9: Structure diagram of CPAM module

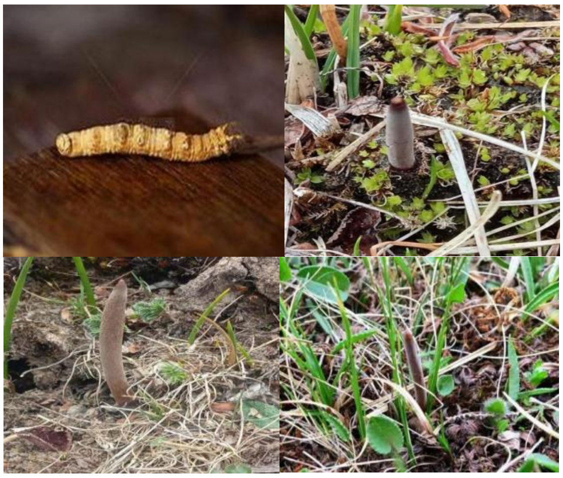

Fig. 10: Sample picture

Experimental Parameters

The image dimensions are 640×640 pixels; the design has undergone 100 iterations; each batch consists of 8 samples; the initial learning rate is 0.01, complemented by a weight decay of 0.0005.

Experimental Results

The results of cordyceps image detection are presented in Fig. (11). As shown in Fig. (11), the detection box confidence is on average 0.95 or higher.

Compared with traditional YOLOv5 and other improved YOLOv5, the improved YOLOv5 network can identify some cordyceps targets that cannot be identified by traditional YOLOv5, other improved YOLOv5 or other networks. Fig. (12) shows the comparison and detection diagram.

In this experiment, under the same dataset, the AP of the enhanced YOLOv5 is employed to assess the performance, as shown in Fig. (13). The map value of the improved algorithm is 99.2%, which is higher than that of other algorithms. Table (1) presents the MAP comparison among various algorithms. Table (1) indicates that the proposed method has achieved substantial enhancements in Mean Average Precision (MAP) compared to traditional object detection algorithms, with a 4% increase in MAP over the traditional YOLOv5 algorithm.

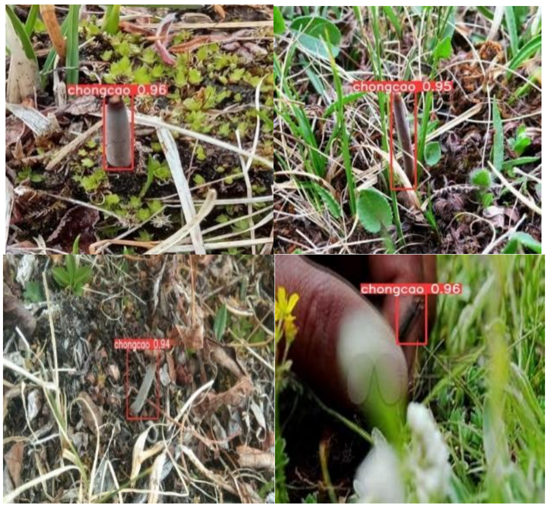

Fig. 11: Test results

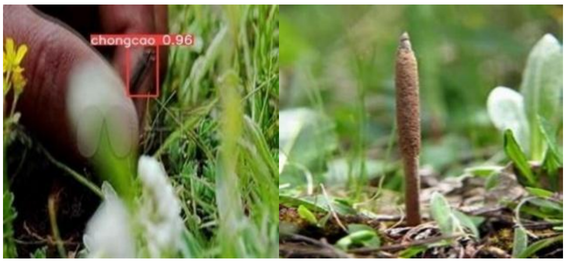

Fig. 12: Comparison and detection diagram

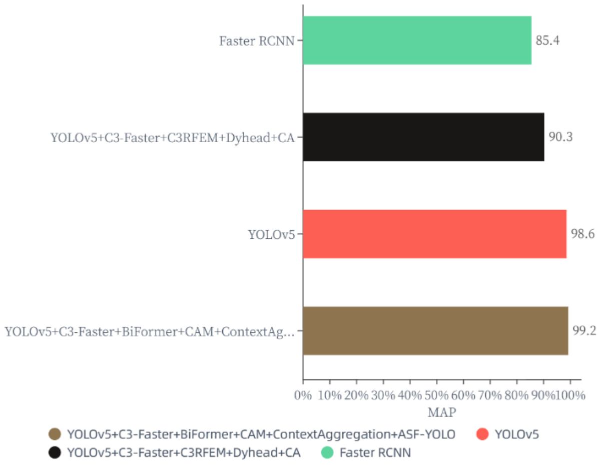

Fig. 13: Comparison of MAP values of different algorithms

Table 1: MAP comparison of different algorithms

Object detection algorithms	MAP (%)
Faster R-CNN	85.4
YOLOv5+C3-Faster+C3RFEM+Dyhead	90.3
Yolov5	98.6
YOLOv5+C3-Faster+BiFormer+CAM+ContextAggregation+ASF-YOLO	99.2

Conclusion

The target recognition method for Cordyceps sinensis based on ASF-YOLOv5is proposed which can integrate the Biformerattention mechanism and ASF-YOLO framework to significantly improve the detection accuracy of small targets. On the self-made cordyceps sinensis dataset, the method has an average accuracy (MAP) of 99.2% and an average accuracy of actual detection of 95%. This can verify its high efficiency and reliability in the identification of cordyceps sinensis. In contrast to the conventional YOLOv5 algorithm, this approach exhibits distinctive strengths in detecting challenging Cordyceps targets, opening up a new path for the automatic detection of Cordyceps sinensis. Although the results of this study are encouraging, there is still potential for further improvement. Given the diversity of the growth environment of cordyceps and its varying forms, especially under the influence of diverse backgrounds, this may affect the detection ability. Therefore, the elasticity of this algorithm requires further improvement.

Future studies can focus on a wider range of cordyceps samples to enhance the generalization ability and adaptability of the model. In addition, the introduction of cutting-edge image processing technologies, such as image super-resolution processing, is expected to further improve the recognition accuracy of small targets.

Acknowledgment

Thank you to the publisher for their support in the publication of this research article. We are grateful for the resources and platform provided by the publisher, which have enabled us to share our findings with a wider audience. We appreciate the efforts of the editorial team in reviewing and editing our work, and we are thankful for the opportunity to contribute to the field of research through this publication.

Funding Information

This study was sponsored by Key Scientific Research Project of Panzhihua University Funds (No. 41671432) and National Innovation Project for College Students (Nos.S2020011360016 and S201911360092).

Author Contribution

Ru Yang: Conceived and executed the experiments, processed the data, and composed the manuscript.

Peng Wu: Contributed to gathering materials pertinent to the experimental work.

Zhentao Qin: Planned the experimental procedures and provided revisions to the draft.

Ethics

The authors declare their responsibility for any ethical issues that may arise after the publication of this manuscript.

Conflict of Interest

The authors state that there are no conflicts of interest.

References

Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C.-H., & Chan, S.-H. G. (2023). Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 12021–12031. https://doi.org/10.1109/cvpr52729.2023.01157

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., & Wei, Y. (2017). Deformable Convolutional Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 764–773. https://doi.org/10.1109/iccv.2017.89

Du, Y., Liu, X., Yi, Y., & Wei, K. (2024). Incorporating bidirectional feature pyramid network and lightweight network: a YOLOv5-GBC distracted driving behavior detection model. *Neural Computing and Applications*, *36*(17), 9903–9917. https://doi.org/10.1007/s00521-023-09043-5

Hui, Y., You, S., Hu, X., Yang, P., & Zhao, J. (2024). SEB-YOLO: An Improved YOLOv5 Model for Remote Sensing Small Target Detection. *Sensors*, *24*(7), 2193. https://doi.org/10.3390/s24072193

Kang, M., Ting, C.-M., Ting, F. F., & Phan, R. C.-W. (2024). ASF-YOLO: A novel YOLO model with attentional scale sequence fusion for cell instance segmentation. *Image and Vision Computing*, *147*, 105057. https://doi.org/10.1016/j.imavis.2024.105057

Lee, K., Ko, J. G., &Yoo, W. (2019). An Intensive Study of Backbone and Architectures with Test Image Augmentation and Box Refinement for Object Detection and Segmentation. *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, 673–677. https://doi.org/10.1109/ictc46691.2019.8939591

Li, D., Ren, H., Wang, G., Wang, S., Wang, W., & Du, M. (2023). Coal gangue detection and recognition method based on multiscale fusion lightweight network SMS-YOLOv3. *Energy Science & Engineering*, *11*(5), 1783–1797. https://doi.org/10.1002/ese3.1421

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., &Belongie, S. (2017). Feature Pyramid Networks for Object Detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 936–944. https://doi.org/10.1109/cvpr.2017.106

Liu, Y., Li, H., Hu, C., Luo, S., Luo, Y., & Chen, C. W. (2024). Learning to Aggregate Multi-Scale Context for Instance Segmentation in Remote Sensing Images. *IEEE Transactions on Neural Networks and Learning Systems*, 1–15. https://doi.org/10.1109/tnnls.2023.3336563

Park, S., Yeo, Y.-J., & Shin, Y.-G. (2022). PConv: simple yet effective convolutional layer for generative adversarial network. *Neural Computing and Applications*, *34*(9), 7113–7124. https://doi.org/10.1007/s00521-021-06846-2

Sun, J., Gao, H., Wang, X., & Yu, J. (2022). Scale Enhancement Pyramid Network for Small Object Detection from UAV Images. *Entropy*, *24*(11), 1699. https://doi.org/10.3390/e24111699

Wang, C., & Zhang, Z. (2023). YOLOv7: A New Era in Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. https://doi.org/10.1109/CVPR52729.2023.00721

Wang, J., Long, X., Chen, G., Wu, Z., Chen, Z., & Ding, E. (2022). U-HRNet: Delving into Improving Semantic Representation of High Resolution Network for Dense Prediction. In *arXiv:2210.07140*. https://doi.org/10.48550/arXiv.2210.07140

Yang, H., &Qiu, S. (2024). A Novel Dynamic Contextual Feature Fusion Model for Small Object Detection in Satellite Remote-Sensing Images. *Information*, *15*(4), 230. https://doi.org/10.3390/info15040230

Ye, X., Gao, S., & Li, F. (2023). HB-YOLOv5: Improved YOLOv5 based on hybrid backbone for infrared small target detection on complex backgrounds. *Earth and Space: From Infrared to Terahertz (ESIT 2022)*, 1250505. https://doi.org/10.1117/12.2664934

Zhu, L., Wang, X., Ke, Z., Zhang, W., & Lau, R. (2023). BiFormer: Vision Transformer with Bi-Level Routing Attention. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10323–10333. https://doi.org/10.1109/cvpr52729.2023.00995