An Improved YOLOv8-Based Rice Pest and Disease Detection

Yang Yang[1], Jianlin Zhu[1(✉)], Bo Yang[1], Xiao Zhang[1,2], and Jin Huang[3]

[1] South-Central Minzu University, a. College of Computer Science; b. Hubei Provincial Engineering Research Center for Intelligent Management of Manufacturing Enterprises, Wuhan 430074, China
Jianlin.Zhu@mail.scuec.edu.cn

[2] School of Computer Science and Technology, University of Science and Technology of China, Hefei 230026, China

[3] School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China

Abstract. While rice pests and diseases significantly impact crop yields, existing deep learning methods for their detection face challenges with accuracy and deployment complexity. Addressing these issues, this study proposes the YOLOv8-HSFPN, an advanced detection framework. Firstly, it features an innovative High-level Select Feature Pyramid Network (HSFPN) neck network that effectively integrates high-level and low-level feature sets for enhanced feature fusion. Secondly, the addition of a deformable self-attention module further refines the model's adaptability to the varying shapes and locations of targets, dynamically adjusting to the salient features. The proposed model has undergone comparative and ablation studies alongside YOLOv8, YOLOv9, and YOLOv5, confirming its improved accuracy and streamlined deployment. This integration results in a robust detection model that not only marks a significant leap in accuracy, evidenced by a 3% empirical increase over the standard YOLOv8, but is also remarkably compact. At a mere 3.97MB, this substantial 49.87% size reduction compared to its predecessors renders it exceptionally suitable for devices with limited computational resources, thereby enhancing its viability in practical, real-world applications.

Keywords: Rice Pest and Diseases Detection · YOLOv8 · HSFPN · Deformable Self-Attention · Feature Fusion

1 Introduction

In recent years, the escalating severity of the climate has led to an increased frequency of extreme weather events. These events have inflicted significant damage on agricultural production, and concurrently, there has been an upward trend

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-82024-3_8.

N. Magnenat-Thalmann et al. (Eds.): CGI 2024, LNCS 15340, pp. 96–107, 2025.
https://doi.org/10.1007/978-3-031-82024-3_8

in the incidence of crop pests and diseases. [1] The need for swift and precise detection of crop pests and diseases has grown in importance. Traditionally, identification has depended on manual examination by teams of agricultural experts [2], typically assembled from plant protection departments. Nevertheless, manual identification methods hinge heavily on the expertise of the examiner, presenting several drawbacks, including protracted processing times, diminished detection efficiency, and a pronounced subjectivity in outcomes.

This study proposes an enhanced pest detection model, YOLOv8-HSFPN, which leverages a deep learning approach to integrate multilevel feature fusion and deformable self-attention mechanisms [3]. The HSFPN supersedes the conventional neck network by merging a feature fusion module that synergizes high-level and low-level feature sets. Additionally, it incorporates a deformable self-attention module designed for the adaptive calibration and weighting of input features [4]. The dataset utilized in this research comprises images captured within an experimental shed, which have undergone various transformations such as flipping, scaling, and cropping to augment the data, ultimately resulting in the finalized dataset.

2 Related Work

Deep learning, a pivotal subset of machine learning [5], alongside artificial intelligence and other advanced information technologies, paves the way for the evolution of modern agriculture, facilitating the progression towards intelligent farming. Despite the broad utilization of deep learning, research in the domain of crop pest detection remains relatively scarce [6,7].

Liangliang Tian et al. [8] proposes a novel apple leaf disease detection method called VMF-SSD (V-space-based Multi-scale Feature-fusion SSD),and experimental results showed that the VMF-SSD method achieves 83.19% mAP and obtains the detection speed of 27.53 FPS on the test set. Furthermore, Yuanjia Zhang et al. [9] enhanced detection capabilities by implementing the Efficient Channel Attention (ECA) mechanism, the hard-Swish activation function, and the Focal Loss function. Post-experimentation, this model demonstrated a mean Average Precision (mAP) of 94.04%, with its detection accuracy and speed meeting the demands for real-time applications.

3 Methodology

Currently, the majority of deep learning algorithms for pest detection operate on a single-stage basis, offering rapid detection speeds. However, these tend to fall short in accuracy when compared to multi-stage algorithms. This study leverages the HS-FPN from the MFDS-DETR to enhance the feature fusion layer and refine the original Neck component [3]. Within the YOLOv8 framework, the Neck serves as an intermediary between the backbone and head networks, facilitating feature fusion and processing, thereby augmenting both detection efficiency and accuracy. This paper proposes the YOLOv8-HSFPN model, as depicted in Fig. 1, an advancement over the existing YOLOv8. The HSFPN supersedes the original Neck network by integrating a feature fusion module that synergizes high-level and low-level features. Additionally, a variable row self-attention module is implemented to adaptively adjust and weight input features, culminating in simultaneous enhancements in detection speed and accuracy [3].

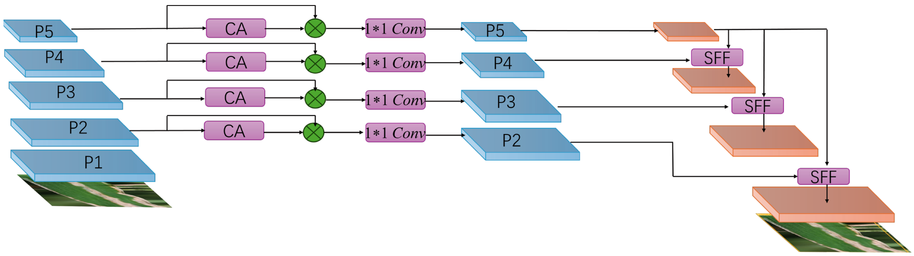

Fig. 1. The structure of YOLOv8-HSFPN

3.1 The Basic Principle of HSFPN

The structure of HSFPN is illustrated in Fig. 2, the architecture is comprised of two principal components: the feature selection module and the feature fusion module. The feature selection module leverages channel attention (CA) and dimension matching (DM) mechanisms to selectively filter feature maps across various scales [3]. Utilizing pooling operations, such as global average pooling and global maximum pooling, coupled with weight computation, this module proficiently isolates vital information within each channel. The feature fusion module, on the other hand, integrates the refined low-level features with high-level features employing the selective feature fusion (SFF) mechanism. High-level features undergo expansion and are scaled using bilinear interpolation or transposed convolution [11], facilitating their amalgamation with low-level features to bolster the model's feature representation capabilities. Collectively, these modules enable HS-FPN to adeptly address the challenges of multiscale detection, thereby enhancing both the precision and robustness of the detection process.

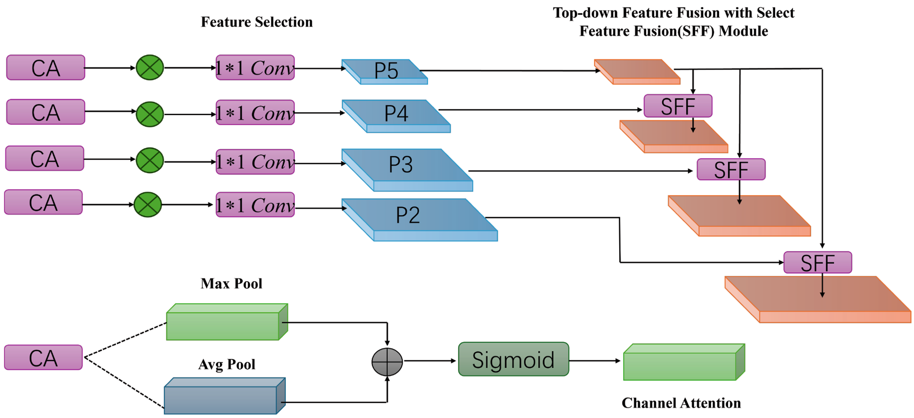

Fig. 2. The structure of HSFPN

3.2 The Feature of Selection Module

As depicted in Fig. 2, the Channel Attention (CA) module and the Dimension Matching (DM) module are crucial components in the detection process, significantly influencing its efficacy. Initially, the CA module processes the input feature map, denoted as $f_{in} \in R^{C\times H\times W}$, where C, H and W represent the number of channels, height, and width of the feature map, respectively. This map is subsequently merged with features derived from global maximum pooling and global average pooling operations [12]. Utilizing the Sigmoid activation function, the CA module calculates the weight of each channel, resulting in a weighted channel feature map, $f_{CA} = R^{C\times 1\times 1}$. Pooling operations serve multiple purposes: they reduce the feature map's dimensionality, compress features, minimize parameter computation, and confer translation, rotation, and scale invariance to the model. Maximum pooling is designed to capture the most salient features by dividing the input feature map into distinct, non-overlapping regions and selecting the maximum value from each region [13]. This operation not only diminishes the feature map's size but also accentuates crucial spatial details. Conversely, average pooling aims to condense feature dimensions while preserving the overall feature landscape [14]. Similar to maximum pooling, it segments the feature map into non-overlapping regions, selecting the average value of each segment as the output, thereby maintaining the statistical integrity of the features [15]. The integration of the CA module with pooling strategies thus ensures the extraction of representative information with minimal loss. Before proceeding to feature fusion, dimensional matching across varying scales of feature maps, which possess disparate channel counts, is imperative. The DM module accomplishes this by employing 1×1 convolution to standardize the channel count of each scale's feature map to 256.

3.3 SFF Module

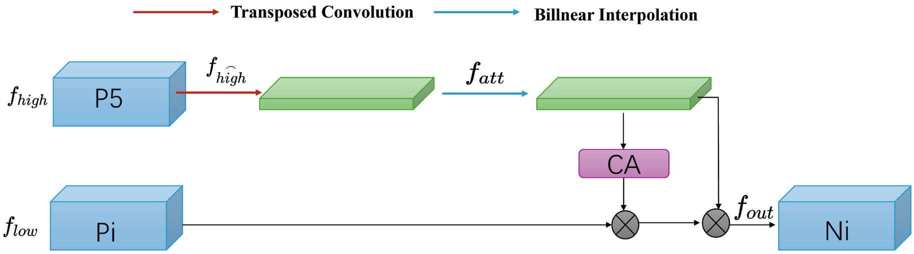

Fig. 3. The structure of SFF

The Selective Feature Fusion (SFF) module adeptly isolates crucial semantic information from low-level features by employing high-level features as guidance weights [3]. The structure of module is illustrated in Fig. 3. This method of feature fusion not only bolsters the module's efficiency but also its accuracy. Considering an input of high-level features $f_{high} \in R^{C\times H\times W}$ and an input of low-level features $f_{low} \in R^{C\times H_1\times W_1}$, the high-level input undergoes processing with a stride of 2 and a 3 convolution kernel, resulting in a modified feature size of $f_{\widehat{high}} \in R^{C\times 2H\times 2W}$. To align the dimensions of high-level and low-level features, bilinear interpolation is applied to either up-sample or down-sample the high-level features, yielding $f_{att} \in R^{C\times H_1\times W_1}$. Subsequently, with dimensionally consistent features, the high-level features are transformed into attention weights via the CA module to refine the low-level features. The culmination of this process is the fusion of the selectively filtered low-level features with the high-level features, thereby enriching the model's feature representation and producing an output of $f_{out} \in R^{C\times H_1\times W_1}$. The equations that follow delineate the feature selection and fusion process:

$$f_{att} = BL(T - Conv(f_{high})) \tag{1}$$

$$f_{out} = f_{low} * CA(f_{att}) + f_{att} \tag{2}$$

During image resampling, a synergistic approach involving transposed convolution (often referred to as inverse convolution) and bilinear interpolation was employed to reconstruct high-resolution feature maps [3,16]. Bilinear interpolation, noted for its straightforwardness and efficiency, operates directly on pixel values to facilitate the image scaling process. Transposed convolution offers distinct advantages: firstly, it adjusts data through learnable parameters, not only increasing the feature map's dimensions but also reconstructing the input data in a convoluted fashion. This is achieved by executing convolutional operations on an upsampled feature map, which involves padding the convolution kernel with zeros; secondly, it can generate varied positions within the output image by sampling disparate regions of the input image, thereby effectively addressing issues of non-uniform sampling.

3.4 The Deformable Self-Attention Module

As represented in Fig. 4, the deformable self-attention module is bifurcated into two pivotal components: the offset module and the attention module [3]. Initially, vectors are transmuted into feature mappings prior to their integration into the Offset Module. This process involves generating input query vectors based on the coordinates of a designated reference point. Subsequently, a linear transformation is applied to the query vector to ascertain the offset Δp_q, while a similar method is applied to the input feature mapping to derive the content feature mapping, and then bilinear interpolation is used to realize the output $offset_{value}$. Within the attention module, the transformational process commences with a linear alteration of the input query vector. Following this, the application of the Softmax function facilitates the creation of a weight vector for each offset [4]. Outputs derived from each offset in the offset module are then amalgamated with their respective weight vectors [3], culminating in the aggregation known as $Sample_{value}$. Attention headers associated with each reference point are subsequently concatenated, forming the composite vector $Sample_{output}$. The process concludes with a linear transformation of the sampled output vectors to procure the final output values. The equations delineated below elucidate this comprehensive process:

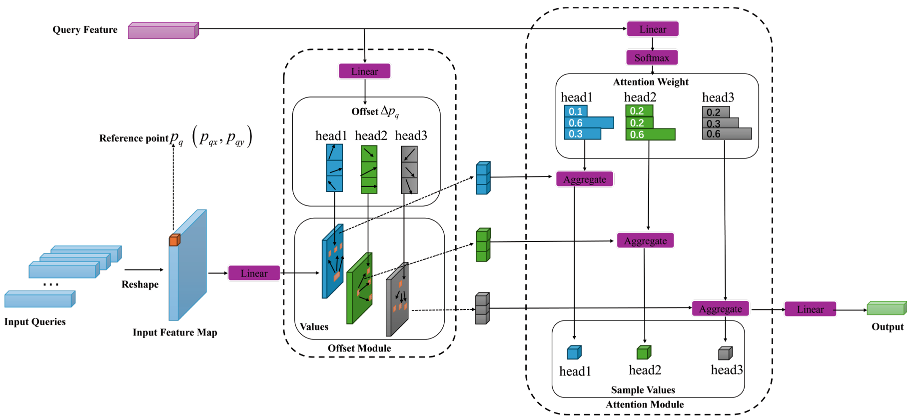

Fig. 4. Deformable Self-Attention Module

$$Weight = Soft\max(WQ) \tag{3}$$

$$Sample_{\nu alue} = \sum_{k=1}^{K} offset_{\nu alue} * Weight \tag{4}$$

$$\begin{aligned} Sample_{output} = Concat(&Sample_{\nu alue}^{1}, \\ &\cdots, Sample_{\nu alue}^{H}), \end{aligned} \tag{5}$$

$$Output = W * Sample_{output} \tag{6}$$

The Deformable Self-Attention Module is engineered to concentrate on critical regions within datasets, overcoming the constraints inherent in conventional attention mechanisms that process extensive contextual information. This module exhibits the capability to dynamically modulate its focus based on the input data, thereby encapsulating a broader spectrum of task-specific feature information. Such an adaptive, data-centric attention mechanism significantly enhances the model's efficacy in processing images characterized by intricate structures and variability.

4 Experiment

4.1 Datasets

This study utilized a comprehensive dataset comprising 11292 images, segmented into 9,868 for training, 949 for validation, and 475 for testing. The dataset encompasses four primary pests and diseases: Leaf Spot Disease, Brownspot, Tungro, and Bacterial Blight, with respective counts of 2658, 3631, 1979, and 3024. To closely mimic the variability encountered in natural settings, the dataset underwent several augmentation techniques, including image flipping and adjustments in brightness and darkness, among other enhancements.

Table 1. The dataset used in this paper.

Categories	Train Set	Validation Set	Test Set
Leaf Spot Disease	2237	313	108
Brownspot	3261	275	95
Tungro	1692	163	124
Bacterial Blight	2678	198	148

Fig. 5. Diseases and pests plant morphology

Table 1 shows the number of images in the training, test and validation sets in the dataset, Fig. 5 depict the plant morphology of four distinct rice pests and diseases, From left to right are Leaf Spot Disease, Brownspot, Tungro, and Bacterial Blight.

4.2 Experimental Environment

To enhance training efficiency, the model leverages the GPU-accelerated version of the PyTorch deep learning framework. Code debugging was conducted using PyCharm 2022.1.3. The software environment comprises Windows 11 Professional 22H2, Python 3.9, CUDA 11.8, NVIDIA Driver 551.76, and CuDNN 8.7. The model was developed with PyTorch 2.2, OpenCV 4.9, TensorBoard 2.16, and TorchSummary 1.5. For hardware, an NVIDIA GeForce RTX 2060 graphics card was utilized.

4.3 Implementation Details

To substantiate the efficacy and real-time capabilities of the enhanced model presented in this study for the detection of crop pests and diseases, an experimental evaluation was conducted. The performance of the proposed model was benchmarked against the YOLOv5 model, with both YOLOv8-HSFPN.pth and YOLOv5-best.pth models undergoing training. The retraining procedure was configured with an epoch count of 500, signifying a comprehensive 500 training cycles. Throughout the model's training phase, vigilant monitoring of the loss function's value was maintained. A stabilization of the loss function over a sequence of training cycles, without substantial decrement, may suggest the attainment of the model's optimal state. Any further training beyond this juncture could potentially lead to overfitting.

Table 2. Model specific parameters during training.

Parameter name	Parameter value
Size of training set images	416 × 416
Validation set image size	416 × 416
Optimizer	SGD
Iterations	500
Initial learning rates	0.01
Batch size	4

To accommodate various regression samples and enhance the efficacy of distinct detection tasks, a linear interval mapping approach has been employed to reformulate the IoU loss [17], thereby facilitating more refined edge regression. The formulation is as follows:

$$IoU^{focaler} = \begin{cases} 0, \quad IoU < d \\ \frac{IoU-d}{u-d}, d \ll IoU \ll u \\ 1, IoU > u \end{cases} \tag{7}$$

where $IoU^{focaler}$ is the reconstructed focaler-IoU and IoU is the original value.$[d, u] \in [0, 1]$. By adjusting the values of d and u, it is possible to make

$IoU^{focaler}$ different from the regression sample. The definition of its loss is shown below:

$$L_{Focaler-IoU} = 1 - IoU^{focaler} \tag{8}$$

Applying the focaler-IoU loss to the existing IoU-based bounding box regression loss function, we get $L_{Focaler-GIoU}$, $L_{Focaler-DIoU}$,$L_{Focaler-CIoU}$, $L_{Focaler-EIoU}$, and $L_{Focaler-SIoU}$,The calculation formula is as follows.

$$L_{Focaler-GIoU} = L_{GIoU} + IoU - IoU^{Focaler} \tag{9}$$

$$L_{Focaler-DIoU} = L_{DIoU} + IoU - IoU^{Focaler} \tag{10}$$

$$L_{Focaler-CIoU} = L_{CIoU} + IoU - IoU^{Focaler} \tag{11}$$

$$F_{Focaler-EIoU} = L_{EIoU} + IoU - IoU^{Focaler} \tag{12}$$

$$L_{Focaler-SIoU} = L_{SIoU} + IoU - IoU^{Focaler} \tag{13}$$

5 Experimental Results

5.1 Comparative Tests of Different Models

The experiments of this paper's algorithm with YOLOv5, YOLOv8 and YOLOv9 illustrate the effectiveness of the algorithm proposed in this paper for detecting rice pests and diseases. Secondly, the comparative analysis of ablation experiments shows that the HSFPN proposed in this paper improves the neck of the original network better than the original one.

Table 3. Comparative tests with different models.

Model name	mAP50	mAP50-95	Recall	Model weights
YOLOv9	60.4%	29.4%	62.6%	6.27MB
YOLOv8	60.50%	29.10%	62.30%	5.95MB
YOLOv5	58.90%	26.00%	65.80%	13.70MB
YOLOv8-HSFPN(ours)	62.40%	30.30%	63.20%	3.97MB

The experimental findings indicate that the proposed YOLOv8-HSFPN model achieves a mAP50 of 62.4% and a mAP50-95 of 30.3%, surpassing the performance of both the YOLOv8 and YOLOv5 models. Although the recall rate of 63.2% is slightly lower than that of the YOLOv5 model, the model proposed in this study is significantly more lightweight, with reductions in weight size by 33.2% and 71.0% compared to YOLOv8 and YOLOv5, respectively. This reduction in size renders the model suitable for deployment on edge devices with limited computational capabilities, thus aligning with the practical demands of agricultural production.

5.2 Ablation Experiments

Through ablation experiments, it is demonstrated that the model proposed in this paper has an advantage in accuracy, and the improved neck network also works better than other neck networks. The experimental results are shown in Table 4.

Table 4. Results of ablation experiments.

Model name	mAP50	mAP50-95	Recall
VanillaNet + BiFPN	59.0%	28.3%	62.2%
Slim-Neck	62.1%	29.4%	61.0%
HATHead	61.1%	29.2%	64.2%
SENetV2	61.9%	29.9%	65.4%
DySample	60.9%	29.4%	63.5%
RepHead	61.7%	29.9%	63.3%
AFPN	61.6%	29.1%	62.7%
CGNet	61.4%	29.2%	63.7%
HSFPN(ours)	62.40%	30.30%	63.20%

5.3 Real Image Detection

The experimental outcomes presented in Figs. 6, 7, and 8 illustrate the detection capabilities of each model under actual field conditions. Notably, the YOLOv5 model fails to detect the first image, whereas the YOLOv8 model identifies one instance. The YOLOv8-HSFPN model excels by detecting two instances with a confidence level of 0.63, which surpasses the confidence levels achieved by the prior detections.

Fig. 6. YOLOv8-HSFPN

Fig. 7. YOLOv8

Fig. 8. YOLOv5

5.4 Limitation

Although the model presented in this paper successfully fulfills its primary objective of pest and disease detection, it exhibits certain limitations that warrant attention for future enhancement. The datasets employed herein are exclusively derived from images captured within experimental greenhouses, which diverge significantly from natural crop growth conditions. Although these images may yield satisfactory performance in validation sets, they often fall short in real-world testing. Future work should, therefore, prioritize the acquisition of more diverse images from actual field conditions.

6 Conclusion

In this paper, we propose a novel High-level Select Feature Pyramid Network (HSFPN) that substitutes the Neck component of the traditional YOLOv8 architecture. Enhancements incorporating multilevel feature fusion and deformable self-attention mechanisms facilitate more efficient image feature extraction. The resultant model is not only more lightweight but also exhibits improvements in detection accuracy, while concurrently requiring fewer computational resources. Such reductions enable deployment on plant protection devices with computational constraints. The model proposed in this paper outperforms the compared models on both mAP50 and mAP50-95. Real-time detection of pests and diseases empowers agronomists to monitor crop health precisely and administer treatments as needed.

Acknowledgments. This work was supported in part by the Fund for Academic Innovation Teams and Research Platform of South-Central Minzu University (Grant Number: XTZ24003, PTZ24001), Knowledge Innovation Program of Wuhan-Basic Research

(Project No.: 2023010201010151), and the Research Start-up Funds of South-Central Minzu University under grant YZZ18006, and the Spring Sunshine Program of Ministry of Education of the People's Republic of China under grant HZKY20220331.

References

1. Watt, M.S., et al.: Early prediction of regional red needle cast outbreaks using climatic data trends and satellite-derived observations. Remote Sens. **16**(8), 1401 (2024)
2. Xia, Y., et al.: Detection of surface defects for maize seeds based on YOLOv5. J. Stored Prod. Res. **105**, 102242 (2024)
3. Chen, Y., et al.: Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. Comput. Biol. Med. **170**, 107917 (2024)
4. Nguyen, D.K., Ju, J., Booij, O., Oswald, M.R., Snoek, C.G.: Boxer: box-attention for 2D and 3D transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4773–4782 (2022)
5. Patterson, J., Gibson, A.: Deep Learning: A Practitioner's Approach. O'Reilly Media, Inc. (2017)
6. Teixeira, A.C., Ribeiro, J., Morais, R., Sousa, J.J., Cunha, A.: A systematic review on automatic insect detection using deep learning. Agriculture **13**(3), 713 (2023)
7. Liu, J., Wang, X.: Plant diseases and pests detection based on deep learning: a review. Plant Methods **17**, 1–18 (2021)
8. Tian, L., et al.: VMF-SSD: a novel v-space based multi-scale feature fusion SSD for apple leaf disease detection. IEEE/ACM Trans. Comput. Biol. Bioinform. (2022)
9. Zhang, Y., Ma, B., Hu, Y., Li, C., Li, Y.: Accurate cotton diseases and pests detection in complex background based on an improved YOLOX model. Comput. Electron. Agric. **203**, 107484 (2022)
10. Jiang, M., Wang, Y., Guo, M., Liu, L., Yu, F.: UPDN: pedestrian detection network for unmanned aerial vehicle perspective. In: Computer Graphics International Conference, pp. 27–39. Springer (2023)
11. Shahbaz, A., Jo, K.H.: Deep Atrous spatial features-based supervised foreground detection algorithm for industrial surveillance systems. IEEE Trans. Ind. Inf. **17**(7), 4818–4826 (2020)
12. Panda, M.K., Sharma, A., Bajpai, V., Subudhi, B.N., Thangaraj, V., Jakhetiya, V.: Encoder and decoder network with ResNet-50 and global average feature pooling for local change detection. Comput. Vis. Image Underst. **222**, 103501 (2022)
13. Scherer, D., Müller, A., Behnke, S.: Evaluation of pooling operations in convolutional architectures for object recognition. In: International Conference on Artificial Neural Networks, pp. 92–101. Springer (2010)
14. Yuan, X., Qiao, Z., Meyarian, A.: Scale attentive network for scene recognition. Neurocomputing **492**, 612–623 (2022)
15. Jiao, X., Chen, Y., Dong, R.: An unsupervised image segmentation method combining graph clustering and high-level feature representation. Neurocomputing **409**, 83–92 (2020)
16. Lu, W., Song, Z., Chu, J.: A novel 3D medical image super-resolution method based on densely connected network. Biomed. Sig. Process. Control **62**, 102120 (2020)
17. Zhang, H., Zhang, S.: Focaler-IoU: more focused intersection over union loss. arXiv preprint arXiv:2401.10525 (2024)