Mathematical Modeling and Algorithm Application
ISSN: 3006-0842 | Vol. 8, No. 1, 2026

Research on Feature Fusion Algorithms Based on the YOLO Model

Hu Li, Lei Dong*, Siyuan Zhao

School of Mechanical Engineering, Tianjin University of Technology and Education, Tianjin 300222, China.

* **Corresponding author**: Lei Dong

Abstract: As a crop of both medicinal value and illicit abuse potential, the rapid and accurate identification of illegally cultivated poppies is critical for global counter-narcotics operations. However, in complex natural environments, poppy plants exhibit high morphological similarity to other vegetation, resulting in substantial false positives and false negatives in existing detection models. To address this, we propose a feature fusion algorithm that integrates YOLOv9 with HSV color space filtering. The method employs empirically defined HSV thresholds to pre-screen poppy-specific color features, generating color attention masks. These masks are channel-aligned and fused pointwise with multi-scale deep features extracted by YOLOv9, achieving complementary enhancement between low-level color priors and high-level shape semantics. Evaluated on a custom drone aerial poppy dataset, the proposed model reduces the false positive rate from 18.5% to 3.2% while maintaining high detection efficiency. This work presents an effective feature fusion strategy to mitigate misclassification in small-target aerial detection within complex scenes, offering practical value for enhancing the automation and precision of anti-drug surveillance.

Keywords: Poppy Identification, YOLOv9, HSV Color Filter, Small Object Detection, Feature Fusion.

1. Introduction

Poppy cultivation holds significant medicinal and economic value, yet it is also a primary source of illicit narcotics, leading to strict global controls. Rapid and accurate detection of illegal plantations is therefore crucial for public health and anti-narcotics operations. However, traditional satellite remote sensing is often ineffective against small-scale, dispersed plots, making drone-based detection a critical research focus. This approach, however, faces major challenges: poppies exhibit growth-stage variation and natural camouflage, requiring robust feature discrimination; they are often hidden in complex environments, leading to occlusion and lighting issues that challenge feature extraction; and law-enforcement applications demand a careful balance between detection recall and precision to minimize both missed targets and false alarms.

Many scholars have used object detection for aerial image recognition with drones. Rominger[1] et al. successfully identified and counted individual low-growing opium poppy plants at a height of 40 metres by improving the YOLOv5 model. However, their research also indicated that, at higher flight altitudes, missed detections and false positives caused by excessively small target sizes were the main sources of error. Xinyun Zhang[2] et al. introduced a multi-scale attention mechanism into YOLOv5 to improve the model's ability to detect small targets at different scales, thereby enhancing detection accuracy in complex planting environments. To further reduce false positive rates, Qian Wang[3] et al. designed a two-stage detection workflow. First, they employed lightweight YOLOv5s for rapid initial screening. Then, they used a DenseNet121 network for fine-grained classification verification of candidate regions to effectively eliminate visually similar distractors. Matija Radovic[4] et al. achieved a detection accuracy of 97.8% using a CNN approach on aerial imagery. However, the insufficient speed of this approach indicates that YOLO-based algorithms are better suited to real-time processing in autonomous drone missions. Subsequently, research shifted towards optimising and improving the YOLO framework. Zhang[5] et al. proposed the YOLOHLA network, which enhances model generalisation through dual-branch feature extraction and achieves an average accuracy of 88.2%. The aforementioned research has made significant strides in improving recognition accuracy when utilising drone platforms and deep learning for poppy identification. Nevertheless, it continues to encounter significant challenges in distinguishing between morphologically similar plants.

Although existing algorithms have enhanced aerial target detection, they remain prone to false positives and negatives when discriminating targets of similar morphology. To address this, we propose a novel feature extraction model that integrates YOLOv9 with HSV-based color filtering to improve the detection accuracy of small drone-based targets. The primary contributions are: 1) a color filtering module that performs preliminary screening using empirically derived poppy color thresholds to isolate key chromatic features; and 2) a feature fusion mechanism that effectively combines the extracted color attributes with YOLOv9's shape-based semantic features. This integrated approach strengthens the model's discriminative capacity between poppies and visually similar crops, thereby reducing both false positive and false negative rates.

2. Algorithm Model

2.1. YOLOv9 Network Model

The YOLOv9 model, introduced in 2024, advances the YOLO (You Only Look Once) series by tackling information bottlenecks through two key innovations: Programmable Gradient Information (PGI) and a Generalized Efficient Layer Aggregation Network (GELAN). PGI ensures reliable gradient flow for weight updating by providing complete input information to the objective function. It consists of three components: a main inference branch, an auxiliary reversible branch to mitigate deep network degradation, and multi-level

auxiliary information to reduce error accumulation in deep-supervised, multi-branch architectures.

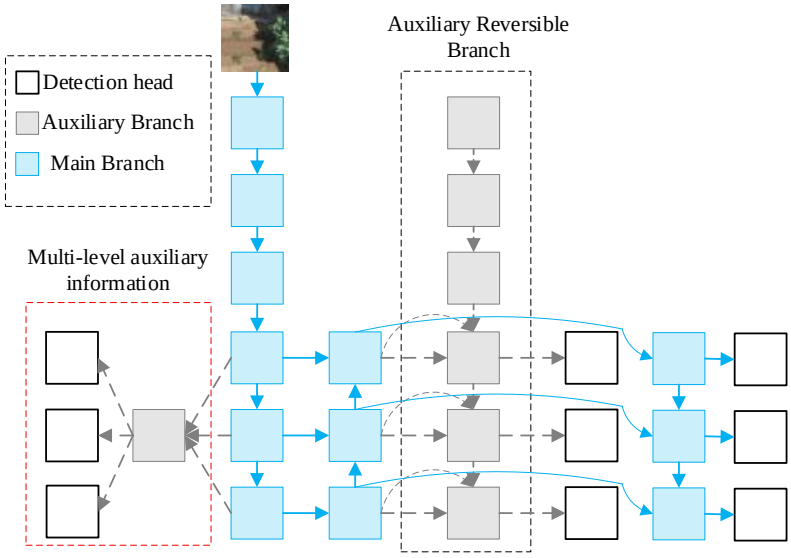


Figure 1 Programmable gradient infographic

GELAN combines the strengths of CSPNet and ELAN to design a flexible and efficient feature fusion architecture. This approach reduces the number of parameters in the new model, enhances feature reuse and improves its feature representation capabilities. Compared to other approaches, GELAN achieves superior parameter utilisation by using only traditional convolutional operators to generalise the functionality of ELAN, which relies solely on stacked convolutional layers, to a new architecture capable of utilising any computational block.

2.2. HSV Color Filtering Method

2.2.1. Color Space Distribution

The HSV color space models color in a manner more consistent with human perceptual intuition than the standard RGB model, by distinctly separating its three fundamental attributes: hue, saturation, and value (from which its name derives)[6]. This separation facilitates intuitive color analysis and manipulation in computer vision tasks. The structure of the HSV color space is illustrated in Figure 2.

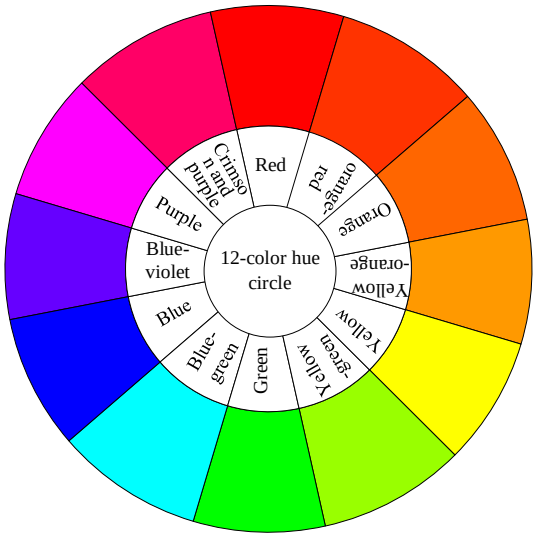


Figure 2 Flat Color Ring Distribution Chart

In the HSV color model, hue (H) is defined on a circular scale from 0° to 360°, where 0°, 120°, and 240° correspond to red, green, and blue, respectively, with intermediate angles representing continuous color transitions. Saturation (S), ranging from 0% to 100%, measures color purity, with higher values indicating greater vividness. Value (V), also ranging from 0% to 100%, represents perceived lightness, where 0% is pure black and 100% approaches white under low saturation.

2.2.2. Feature Fusion Model

A binary mask constitutes a fundamental and pivotal data structure within the fields of computer vision and image processing. The marking of specific regions within an image is achieved by the use of simple binary values of 0 and 1. Within the domain of image processing, a binary mask is defined as a matrix that is equivalent in dimensions to the original image. Each pixel position possesses a value of either 0 or 1, signifying whether that location belongs to the region of interest. This concise representation renders binary masks a crucial tool for tasks such as image segmentation, object detection, and image editing.

In this study, binary masks are typically generated based on specific thresholding criteria. First, RGB images are

converted to the HSV color space. This conversion is crucial because the HSV space effectively separates color information from luminance information, making color features relatively stable. Based on the unique color characteristics of poppy plants, appropriate threshold ranges are established across the three dimensions of hue (H), saturation (S), and value (V). Pixels meeting all channel thresholds are labeled as 1, forming the foreground region; pixels failing to meet the conditions are labeled as 0, constituting the background region. This process effectively establishes a color-based pixel-level classification. A schematic illustration of the procedure is shown in Figure 3.

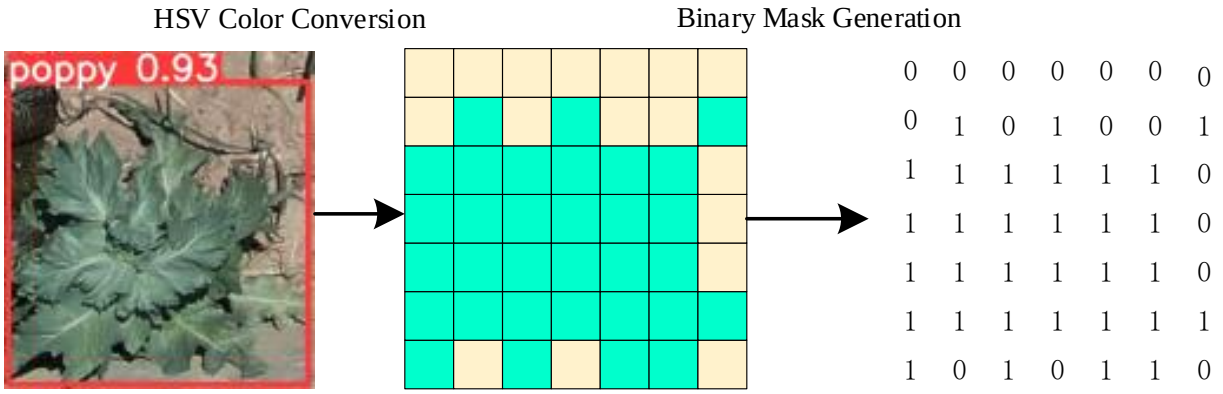


Figure 3 Schematic Diagram of Binary Mask Generation

In the specific application scenario of poppy detection, the binary mask assumes multiple critical functional roles. The apparatus functions as a spatial information filter, serving to efficiently and clearly segment all suspected target areas from complex backgrounds. This is achieved based on predefined poppy color characteristics. The result is a clear and pure foreground target, which can then be subjected to detailed analysis and quantitative assessment. The mask itself constitutes an effective noise-filtering mechanism. Through the implementation of rigorous pixel-level thresholding, the vast majority of background interference that does not correspond to the target colour characteristics is eliminated. This has been shown to significantly reduce data complexity and computational burden in subsequent processing stages (Smith, 2020), enhancing the overall system's processing efficiency and robustness (Jones et al., 2019). Of particular note is the binary mask's capacity for integration as an attention-guiding mechanism. When combined with deep learning models, the prior spatial information embedded in the mask can be utilised to weight feature maps or constrain regions of interest. This approach enables the network model to concentrate its limited computational resources and attention on image subregions with a high probability of containing poppies. This approach has been demonstrated to suppress background activation, enhance feature discriminative power, and ultimately improve the model's detection accuracy and generalization performance.

2.2.3. Feature Fusion Model

The proposed model integrates YOLOv9 with colour filtering techniques. The model utilises a dual-path parallel processing approach and incorporates a multi-level feature interaction design philosophy. The integrated model's structural diagram is presented in Figure 4.

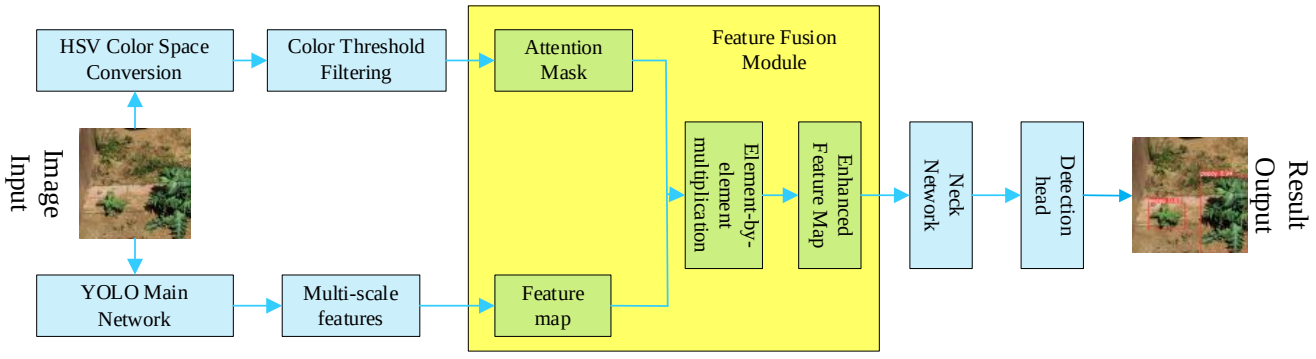


Figure 4 Architecture Diagram of the Fusion Model

During the fusion of HSV color filtering and YOLOv9, the feature fusion operation serves as the core of the entire design. This operation requires performing a pointwise multiplication between the color weight matrix and the feature matrix, thereby guiding the network to focus on regions of specific colors. The working principle of the feature fusion operation is illustrated in Figure 5.

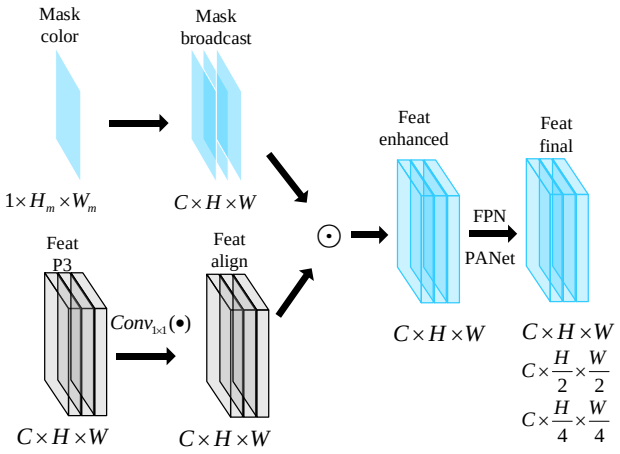


Figure 5 Architecture Diagram of the Fusion Model

As demonstrated in the accompanying figure, Maskcolor signifies the initial colour attention mask derived from the poppy image. The initial dimension of the image is $1\times H_m\times W_m$, and it is a single-channel binary image. Following the process of channel broadcasting, the result is Maskbroadcast.

$$Mask_{broadcast}=Broadcast(Mask_{resize},channel(Feat_{align})) \tag{1}$$

$$Mask_{resize}=\mathrm{Re}\,size(Mask_{color},size(Feat_{align})) \tag{2}$$

In the formula, $Feat_{align}$ denotes the feature map after channel alignment, $Mask_{resize}$ represents the attention mask after size matching, $channel(\bullet)$ is the number of target channels obtained, $\mathrm{Re}\,size$ is the size scaling operation that adjusts the original color attention mask size from $1\times H_m\times W_m$ to $1\times H\times W$, and $size(\bullet)$ is the target size obtained.

To fuse with the color attention mask, the feature map also requires a matching operation to compute the matched feature map $Feat_{align}$. The specific calculation formula is as follows.

$$Feat_{align}=Conv_{1\times1}(Feat_{P3}) \tag{3}$$

In the formula, $Feat_{P3}$ represents the feature map output from the P3-scale backbone network, with dimension $C\times H\times W$. $Conv_{1\times1}(\bullet)$ denotes a 1×1 convolution operation, which aims to align channels without altering the feature space dimensions. After obtaining $Feat_{align}$ and $Mask_{broadcast}$, a pointwise multiplication operation is performed to fuse the features. The calculation formula is as follows:

$$Feat_{enhanced}=Feat_{align}\odot Mask_{broadcast} \tag{4}$$

In the formula, $Feat_{enhanced}$ represents the fused feature map, and $\odot$ denotes the element-wise multiplication operation. By performing upsampling and downsampling operations on the enhanced feature map, top-down aggregation is achieved, enabling multi-scale object recognition. The calculation formula is as follows:

$$Feat_{FPN}=UpSample(Feat_{high})+Feat_{low} \tag{5}$$

$$Feat_{PAN}=DownSample(Feat_{low})+Feat_{high} \tag{6}$$

$$Feat_{final}=\left[Feat_{PAN(P2)},Feat_{PAN(P3)},Feat_{PAN(P4)},Feat_{PAN(P5)}\right] \tag{7}$$

In the equation, $UpSample(\bullet)$ denotes the upsampling operation, which is used to propagate high-level feature maps downward to match the dimensions of lower-level features. $DownSample(\bullet)$ denotes the downsampling operation, used to propagate lower-level feature maps upward to match the dimensions of higher-level features. $Feat_{high}$ represents high-level semantic features, characterized by low resolution and rich semantic information. $Feat_{low}$ represents low-level detail feature maps, characterized by high resolution and precise positional information; $Feat_{final}$ denotes the final feature map after multi-scale feature fusion.

3. Experiments and Analysis

3.1. Experimental Design

In order to validate the performance of the enhanced algorithm outlined in this paper, a bespoke dataset was constructed. The utilisation of drones for the purpose of aerial photography in the designated survey area was conducted under the oversight of public security authorities. The dataset is composed primarily of raw photographs captured under diverse environmental conditions, backgrounds, and lighting scenarios. The original image resolution is 3000×4000 pixels. Furthermore, in order to enhance the efficiency of the model training process, a script was utilised to resize the images to

640×640 pixels.

During the experiment, it is necessary to select evaluation metrics for object detection based on specific task requirements, seeking a balance between accuracy, speed, and efficiency to ensure that the metrics comprehensively and accurately reflect the model's performance. The algorithm evaluation metrics for detection accuracy are calculated as shown in Equations (8) and (9).

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \frac{\sum AP}{N(Class)} \tag{9}$$

In the formula, $N(Class)$ denotes the category of the detection target, AP represents the average precision of detection, and mAP is the most critical evaluation parameter, representing the average value of AP across different categories. The calculation method for mAP varies under different precision-recall curves. mAP_0.5 indicates the numerical value when the precision-recall curve is greater than or equal to 0.5.

All experiments in this paper were conducted on a central processing unit (CPU) with 14 vCPUs of Intel® Xeon® Gold 6330 CPU @ 2.00GHz and 90GB of RAM. The graphics processing unit (GPU) used for accelerated training was an RTX 3090 with 24GB of VRAM and CUDA version 11.3. The deep learning framework employed was PyTorch 1.10.0. The training cycle was set to 250 epochs, with each batch size processing 16 images.

To validate the effectiveness of the proposed fusion module, a comparative experiment was designed. Using the YOLOv9 model as the baseline, the fusion model incorporating the HSV color filtering method was compared against it. Through quantitative performance evaluation and visual analysis of results, the fusion module's improvement effects on key metrics such as detection accuracy and false positive rate in the poppy identification task were comprehensively examined.

3.2. Results and Analysis

To systematically evaluate the effectiveness of the HSV color filtering method integrated with the YOLOv9 model, this study designed a comparative experiment. On a self-built drone aerial poppy dataset, the proposed fusion model was systematically compared with the original YOLOv9 model as a baseline. The specific experimental results are shown in the table below.

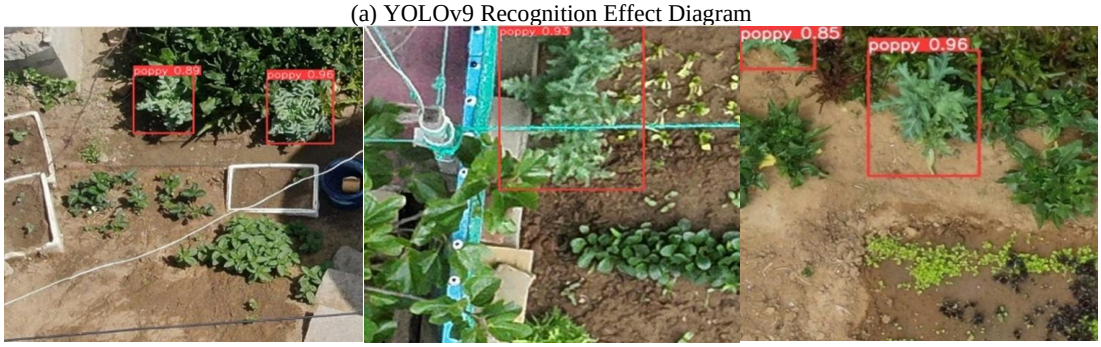
Table 1 Comparative Experiments of HSV Fusion Models

Network Model	mAP_0.5(%)	Parameter quantity	Floating-point operations	False positive rate(%)
YOLOv9	69.7	60495910	263.9	18.5
Hybrid Model	70.2	62317112	268.5	3.2

The experiments detailed in Table 3-1 demonstrate that the fusion model attained a mere 0.5% enhancement in mAP_0.5. This is due to the fact that the colour filtering method corrects false positives for targets with similar shapes and colours, but its impact on detection accuracy across the entire dataset remains limited. When deeply integrated with the improved YOLOv9 model, the advantages of the colour filtering method are more fully realised. In comparison with the YOLOv9 model, the fusion model reduces the false positive rate from 18.5% to 3.2%, representing a decrease of approximately 15.3%. This substantial enhancement serves to underscore the pivotal function of prior knowledge in enhancing model specificity. With regard to computational efficiency, the introduction of the colour filtering algorithm, which exhibits extremely low computational complexity, results in an increase in parameters of merely 3%, a figure that is significantly outweighed by the substantial performance gains it delivers.



(a) YOLOv9 Recognition Effect Diagram



(b) Fusion Model Recognition Rendering

Figure 6 Comparison of Recognition Performance

As demonstrated in Figure 6, this study elected to utilise representative detection results for the purpose of visual comparative analysis. (a) The figure displays the recognition results of the original YOLOv9 model, revealing that the model misidentifies interference crops with similar morphological features, incorrectly classifying non-poppy targets as poppies. (b) presents the detection outcomes of the fused model following the incorporation of the HSV colour space filtering mechanism. This model employs the unique spectral characteristics of poppies to perform a secondary screening of candidate regions based on colour attributes. The system has been demonstrated to successfully eliminate misclassified vegetable targets that share similar shapes but lack matching colour features.

A thorough experimental analysis has been conducted, revealing that the colour filtering method provides robust colour space prior knowledge, thus complementing and enhancing deep learning-based shape feature recognition. In lieu of compromising the original model's detection capabilities, a marked enhancement in accuracy and reliability is observed in real-world complex scenarios through feature-level attention modulation. This provides essential technical assistance for the advancement of drone poppy detection technology towards its practical application.

4. Conclusion

This study targets the key challenges in drone-based detection of illicit poppy cultivation: small target size, high morphological similarity to other crops, and complex, concealed growing environments. To mitigate these issues, we propose a fusion model, MCI-YOLOv9, which integrates YOLOv9 with HSV-based color prior knowledge. The model employs a dual-path architecture to fuse poppy-specific color features with general shape semantics at the feature level. Experimental results show our strategy drastically reduces the false positive rate to 3.2%, a 15.3 percentage point improvement over the baseline YOLOv9, while maintaining high detection efficiency and average precision. Visual analysis confirms the model's enhanced robustness in discarding morphologically similar but color-inconsistent vegetation in complex farmland scenes.

The developed model offers law enforcement a high-precision, low false-alarm tool suitable for real-time UAV deployment. Future work will explore integrating multispectral and temporal growth features to improve model generalization across different phenological stages and lighting conditions.

Acknowledgements

This research was funded by the Tianjin Education Commission Research Program Project [Project No. 2024ZD017], the Tianjin.

References

[1] Rominger K, Meyer S E. Application of UAV-based methodology for census of an endangered plant species in a fragile habitat[J]. Remote Sensing, 2019, 11(6): 719.

[2] Zhang X, Luo D, Zhou L. Opium Poppy Image Detection Based on Improved YOLOV5[C]//2024 International Symposium on Digital Home (ISDH). IEEE, 2024: 127-132.

[3] Wang Q, Wang C, Wu H, et al. A two-stage low-altitude remote sensing papaver somniferum image detection system based on YOLOv5s+ DenseNet121[J]. Remote Sensing, 2022, 14(8): 1834.

[4] Radovic M, Adarkwa O, Wang Q. Object recognition in aerial images using convolutional neural networks[J]. Journal of Imaging, 2017, 3(2): 21.

[5] Zhang Z, Xia W, Xie G, et al. Fast opium poppy detection in unmanned aerial vehicle (UAV) imagery based on deep neural network[J]. Drones, 2023, 7(9): 559.

[6] Mieites V, Gutiérrez-Gutiérrez J A et al. Single-Image Multi-Parametric Representation of Optical Properties through Encodings to the HSV Color Space[J]. Applied Sciences, 2023, 14(1): 155.

[7] Zhang Z, Xia W, Xie G, et al. Fast opium poppy detection in Unmanned Aerial Vehicle (UAV) imagery based on deep neural network[J]. Drones, 2023, 7(9): 559.

[8] Lahoti A, Karp S, Winston E, et al. Role of Locality and Weight Sharing in Image-Based Tasks: A Sample Complexity Separation between CNNs, LCNs, and FCNs[J]. arXiv preprint arXiv:2403.15707, 2024.

[9] Rahman A, He L, Wang H. Activation function optimization scheme for image classification[J]. Knowledge-Based Systems, 2024, 305: 112502.

[10] Galanis N I, Vafiadis P, Mirzaev K G, et al. Convolutional neural networks: A roundup and benchmark of their pooling layer variants[J]. Algorithms, 2022, 15(11): 391.

[11] Babaiee Z, Kiasari P M, Rus D, et al. Unveiling the unseen: Identifiable clusters in trained depthwise convolutional kernels[J]. arXiv preprint arXiv:2401.14469, 2024.

[12] Pham M T, Courtrai L, Friguet C, et al. YOLO-Fine: One-stage detector of small objects under various backgrounds in remote sensing images[J]. Remote Sensing, 2020, 12(15): 2501.

[13] Hu J, Shi C J R, Zhang J. Saliency-based YOLO for single target detection[J]. Knowledge and Information Systems, 2021, 63(3): 717-732.

[14] Lei L, Fang Z, Ren J, et al. Two-click based Fast Small Object Annotation in Remote Sensing Images[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024.

[15] Dendorfer P, Ošep A, Milan A, et al. Motchallenge: A benchmark for single-camera multiple target tracking[J]. arXiv preprint arXiv:2010.07548, 2020.