drones

MDPI

Article

AMS-YOLO: Asymmetric Multi-Scale Fusion Network for Cannabis Detection in UAV Imagery

Xuelin Li [1,2], Huanyin Yue [1,2,*], Jianli Liu [3,4] and Aonan Cheng [3]

[1] State Key Laboratory of Resources and Environmental Information System, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China; lixuelin23@mails.ucas.ac.cn
[2] University of Chinese Academy of Sciences, Beijing 100101, China
[3] National Engineering Research Center of Surveying and Mapping, China TopRS Technology Company Limited, Beijing 100039, China; liujl.18b@igsnrr.ac.cn (J.L.); chengaonan@hhu.edu.cn (A.C.)
[4] College of Geography and Geomatics, Xuchang University, Xuchang 461000, China
* Correspondence: yuehy@lreis.ac.cn

Abstract

Cannabis is a strictly regulated plant in China, and its illegal cultivation presents significant challenges for social governance. Traditional manual patrol methods suffer from low coverage efficiency, while satellite imagery struggles to identify illicit plantations due to its limited spatial resolution, particularly for sparsely distributed and concealed cultivation. UAV remote sensing technology, with its high resolution and mobility, provides a promising solution for cannabis monitoring. However, existing detection methods still face challenges in terms of accuracy and robustness, particularly due to varying target scales, severe occlusion, and background interference. In this paper, we propose AMS-YOLO, a cannabis detection model tailored for UAV imagery. The model incorporates an asymmetric backbone network to improve texture perception by directing the model's focus towards directional information. Additionally, it features a multi-scale fusion neck structure, incorporating partial convolution mechanisms to effectively improve cannabis detection in small target and complex background scenarios. To evaluate the model's performance, we constructed a cannabis remote sensing dataset consisting of 1972 images. Experimental results show that AMS-YOLO achieves an mAP of 90.7% while maintaining efficient inference speed, outperforming existing state-of-the-art detection algorithms. This method demonstrates strong adaptability and practicality in complex environments, offering robust technical support for monitoring illegal cannabis cultivation.

Keywords: UAV imagery; cannabis detection; YOLO

check for updates

Academic Editor: Diego González-Aguilera

Received: 10 July 2025
Revised: 17 August 2025
Accepted: 3 September 2025
Published: 6 September 2025

Citation: Li, X.; Yue, H.; Liu, J.; Cheng, A. AMS-YOLO: Asymmetric Multi-Scale Fusion Network for Cannabis Detection in UAV Imagery. *Drones* **2025**, *9*, 629. https://doi.org/10.3390/drones9090629

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Cannabis (*Cannabis sativa* L.), a traditional medicinal and industrial crop, presents significant global social and health challenges due to the addictive and hallucinogenic properties of the tetrahydrocannabinol (THC) it contains [1,2]. In China, cannabis species are classified as strictly controlled drug source plants, and any cultivation without proper authorization is considered illegal. Despite this, illegal cultivation persists in certain regions, particularly in remote mountainous areas or where it is intercropped with other crops. These illicit operations are often highly covert and mobile, posing considerable challenges for law enforcement and supervision.

Traditional cannabis detection methods mainly rely on manual patrols, which are labor-intensive and have low coverage efficiency. In recent years, remote sensing monitoring technologies have gradually emerged as efficient alternatives. Some studies have attempted to use hyperspectral imagery and satellite data for cannabis cultivation identification. For example, Pereira et al. [3] employed near-infrared hyperspectral imaging combined with machine learning models for spectral classification; Sujud et al. [4] integrated Sentinel-1/2 imagery with random forest algorithms to identify cannabis cultivation areas in Lebanon; Bicakli et al. [5] used high-resolution PlanetScope satellite imagery to analyze spectral differences between cannabis and other plants in northern Turkey, finding that the spectral differences were most significant between May and June. These methods show certain effectiveness in mesoscale areas (large fields), but due to limitations in spatial resolution and issues such as vegetation confusion, they still struggle to identify low-density, covertly distributed illegal plantations. This results in high false detection rates and makes it difficult to meet the needs of refined governance [6,7].

In contrast, Unmanned Aerial Vehicles (UAVs) offer advantages such as low-altitude maneuverability, flexible deployment, and high-resolution imaging, making them particularly suitable for high-precision target detection in complex and heterogeneous environments [8,9]. In the domain of plant phenotyping and agricultural monitoring, UAV-acquired imagery enables the capture of fine-grained structural and textural information, providing detailed data support for accurate target identification [10,11]. However, certain plant species, such as low-growing or morphologically variable crops, present additional detection challenges. These include multi-scale target variation, background clutter, occlusion, and high inter-class similarity, all of which can negatively impact detection performance in natural field conditions.

With the rapid advancement of deep learning technologies, convolutional neural networks (CNNs) have become the dominant approach for image recognition tasks. In particular, CNNs have proven highly effective in target detection by extracting semantic information and spatial relationships through multi-layer feature learning, leading to significant improvements in detection accuracy [12,13]. To address challenges posed by complex backgrounds and multi-scale targets, researchers have recently proposed a series of innovative convolution techniques [14–16]. For example, dilated convolution [17] expands the receptive field of the convolution kernel, allowing the model to extract features more effectively from targets of varying scales. Depthwise separable convolution [18] enhances detection efficiency by reducing the computational load while maintaining high accuracy. Furthermore, the development of attention mechanisms has led to the integration of attention modules in many deep learning models, enabling the model to focus on the most relevant parts of the image, thus improving target localization accuracy [19–21]. In target detection, common attention mechanisms such as spatial attention and channel attention allow the network to adaptively prioritize important regions, reducing background noise interference and boosting detection performance in complex environments. These innovative convolution techniques, coupled with attention mechanisms, provide powerful support for tackling the challenges of target detection.

Object detection algorithms are mainly classified into two categories: two-stage and one-stage methods. Two-stage algorithms, dominated by the R-CNN series, include R-CNN [22], Fast R-CNN [23], Faster R-CNN [24], Cascade R-CNN [25], and Mask R-CNN [26], among others. These algorithms first generate region proposals through techniques like selective search, which are then classified and regressed. While these algorithms achieve high detection accuracy for small objects or complex scenes, their computational complexity is high, making them difficult to adapt to the stringent real-time requirements of UAV platforms. One-stage object detection algorithms, exemplified by the YOLO (You Only

Look Once) series, leverage an end-to-end network architecture to simultaneously predict object locations and categories, demonstrating significant advantages in inference speed. Since the introduction of YOLOv1 [27], the series has continuously evolved: YOLOv3 [28] introduced multi-scale detection structures, YOLOv5 [29] adopted a lightweight backbone network, and YOLOv8 [30] further optimized the decoupled head and anchor-free mechanism, significantly improving detection performance. Among the recent versions, YOLOv11 integrates the C3k2 module and C2PSA, exhibiting excellent detection accuracy and inference efficiency in small-object and occlusion scenarios [31]. More recently, YOLOv13 introduced a hypergraph-based correlation enhancement mechanism and a full-pipeline feature aggregation strategy, achieving additional accuracy improvements over YOLOv11 while maintaining computational efficiency [32]. However, YOLOv13 was released only shortly before the preparation of this study and has not yet undergone extensive validation in UAV-based small-object detection scenarios. Furthermore, its deployment performance and compatibility on lightweight UAV platforms remain to be comprehensively assessed.

In the field of illicit crop monitoring, several studies have investigated the detection of opium poppy using UAV-acquired imagery and deep learning techniques. For instance, Wang, C. et al. [33] employed YOLOv3-based object detection frameworks on UAV-acquired imagery to identify poppy cultivation areas. Zhou et al. [34] proposed a YOLOv3-based detection model enhanced with SPP and GIoU modules and optimized with a MobileNetv2 backbone, tailored for UAV-based poppy monitoring at low altitudes. Wang et al. [35] proposed a two-stage UAV-based detection framework combining YOLOv5s and DenseNet121 to reduce false positives and improve the efficiency of opium poppy image screening in illicit cultivation monitoring. Other approaches include hyperspectral imaging with spectral matching [36] and multi-resolution UAV monitoring for spatial localization of poppy fields [37]. While these efforts confirm the feasibility of UAV-based deep learning for illegal crop detection, they predominantly focus on poppies, use earlier YOLO versions, and are not specifically tailored to complex vegetative backgrounds. To date, and to the best of our knowledge, no published work has systematically applied YOLO or other deep learning–based object detection methods to cannabis identification. Given the increasing relevance of cannabis monitoring in law enforcement and the technical challenges posed by its covert growth patterns, this represents a critical research gap that the present study aims to address.

To tackle the aforementioned challenges, this study proposes an enhanced YOLO-based model, Asymmetric Multi-Scale YOLO (AMS-YOLO), specifically optimized for UAV-based cannabis detection tasks. A dedicated UAV imagery dataset was constructed to enable comprehensive evaluation of the model. The main contributions of the model are as follows:

1. This study proposes a UAV-based cannabis detection approach named AMS-YOLO, which achieves a high detection accuracy of 90.7% mAP@0.5 on a self-constructed dataset while maintaining a compact model size of only 6.4 MB. This provides an efficient and practical solution for the intelligent monitoring of illegal cannabis cultivation.
2. The backbone network incorporates an asymmetric padding mechanism, which enhances the model's ability to capture fine-grained visual features such as the serrated edges and palmate textures of cannabis leaves. Meanwhile, the neck network employs a multi-scale feature fusion mechanism, effectively mitigating the degradation of small-object features in deep networks and improving the model's ability to distinguish targets in complex backgrounds.
3. A high-resolution UAV cannabis imagery dataset has been constructed, covering various growth stages, lighting conditions (e.g., strong light, low light, and normal

illumination), and occlusion scenarios (e.g., power line blockage and shadow interference). This dataset provides a standardized and robust foundation for training and evaluating detection models.

2. Materials and Methods

2.1. Datasets

Compared to satellite imagery, UAV-acquired images offer higher resolution and greater flexibility, providing distinct advantages in detecting illegal cannabis cultivation. Our field flights have demonstrated that morphological features of cannabis plants, such as palmate compound leaves and stem textures, can only be effectively distinguished in images with a ground sampling distance (GSD) of 0.8 cm or less. In contrast, traditional satellite remote sensing is limited to identifying cultivation areas at the hectare scale and cannot meet the requirements for individual plant detection. In this study, a DJI Matrice 300 UAV platform (DJI, Shenzhen, Guangdong, China) was employed, using a Share6100 camera (Share UAV Technology Co., Ltd., Shenzhen, Guangdong, China) to capture images with a spatial resolution of 0.8 cm at a flying height of 120 m. The acquired imagery is shown in Figure 1, with the image metadata provided in Table 1. The original dataset collected in this study spans five different prefecture-level cities, covering a latitudinal range of over 3°, thereby capturing diverse geographic environments that influence target appearance. In terms of temporal distribution, data were acquired between August and September, which closely aligns with the growth cycle of cannabis—an annual plant typically sown in May–June and harvested in September–October. This period corresponds to the plant's mature stage, making it a critical window for field detection. Furthermore, image acquisition was conducted under various lighting conditions, including midday sunlight, overcast skies, and dusk, to simulate environmental variations at the same growth stage. The total volume of raw data exceeds 6 TB.

Figure 1. Low-altitude remote sensing imagery of villages captured by UAV.

Table 1. Image metadata.

Category	Attribute
UAV	DJI M300
Camera	Share6100
Relative flight altitude	120 m
Image resolution	9552 × 6368
Spatial resolution	0.8 cm

In this study, raw imagery data collected from a UAV platform underwent a multi-round, rigorous screening process, ultimately identifying 717 valid images containing illegal cannabis cultivation areas from a vast image database. The screening process is as follows: Initially, technical personnel with experience in cannabis identification conducted a preliminary screening of all raw images based on texture, tone, and spatial distribution characteristics. Subsequently, experts verified the candidate images using geographic information data and historical cultivation records to ensure that each selected image contained at least one clearly marked cannabis cultivation area.

To evade detection, illegal cannabis cultivation is often organized in covert ways, such as single-plant cultivation, patch cultivation, concealed cultivation, field cultivation, forest cultivation, and mixed cultivation, as shown in Figure 2. These diverse cultivation strategies cause the cannabis features in the images to be blurred and difficult to distinguish, presenting significant challenges for target detection.

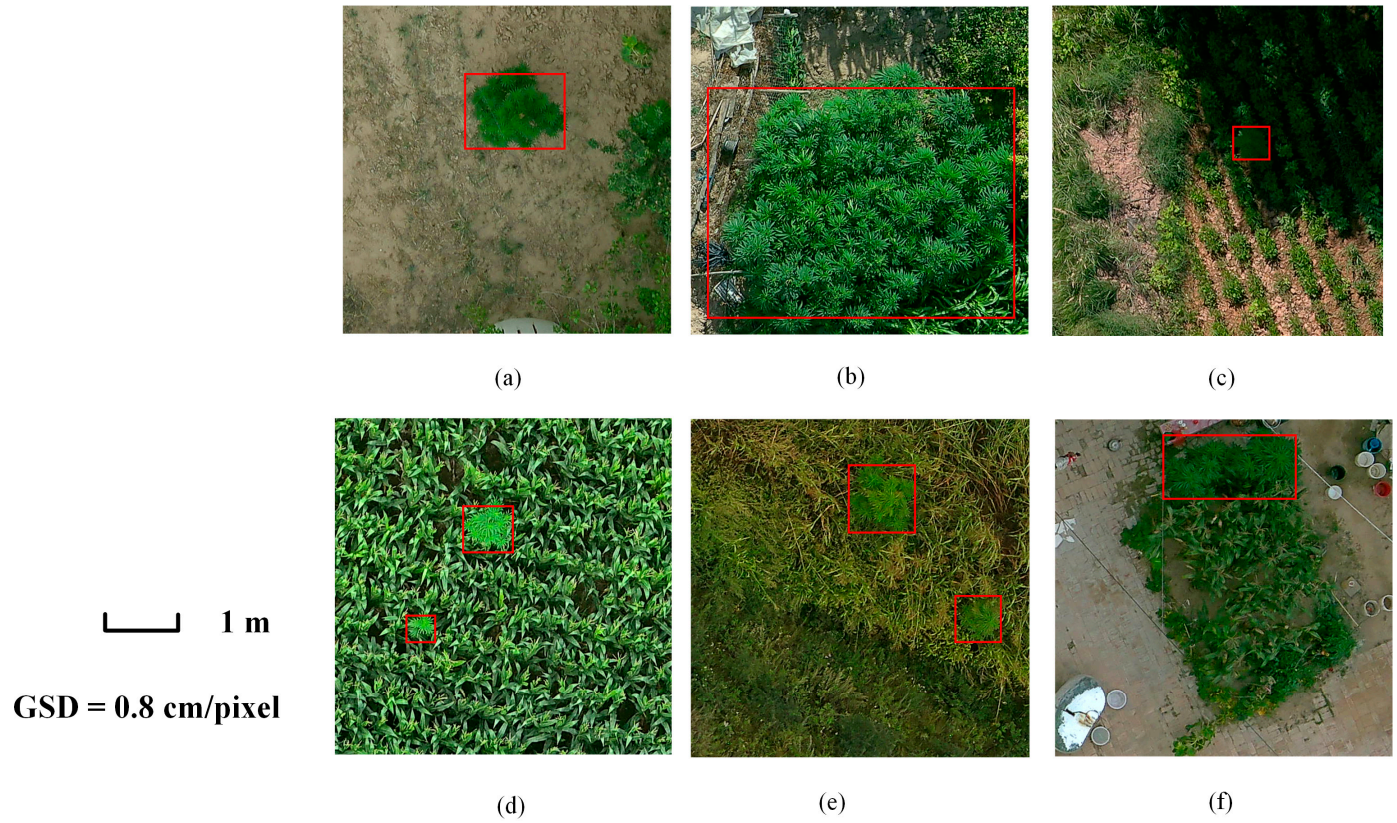

Figure 2. Examples of cannabis planting methods: (**a**) single-plant cultivation; (**b**) patch cultivation; (**c**) concealed cultivation; (**d**) field cultivation; (**e**) forest cultivation; (**f**) mixed cultivation. The red boxes highlight the target plant, cannabis.

The raw images captured by the UAV are high-resolution, with each image measuring 9522 × 6368 pixels and having a data size of approximately 18 MB. Cannabis targets typically occupy a very limited area of the original UAV image, making them visually inconspicuous and challenging to detect. Direct downsampling of the original image before feeding it into a deep learning model leads to substantial loss of pixel information in the already minute target regions, which in turn hinders effective detection.

To effectively preserve target details and meet the input requirements of the model, this study employs a sliding window technique to divide the original images into 640 × 640 pixel sub-images, with a 20% overlap between adjacent windows. Samples containing cannabis were selected to construct a dataset of 1972 samples. The dataset was

divided into 1572 images for training and 400 images for testing, following an 8:2 split, ensuring a solid foundation for model training and performance assessment.

As illustrated in Figure 3, the dataset covers a wide range of environmental conditions and interference factors commonly encountered in UAV-based cannabis detection. These include variations in illumination (normal daylight, direct high-intensity sunlight, and low-light conditions at dusk) as well as physical occlusions, such as overhead objects (e.g., power lines) and cast shadows from surrounding vegetation. This diversity ensures that the dataset more accurately reflects the challenges of real-world aerial monitoring and enhances the robustness evaluation of detection models.

Figure 3. Representative UAV imagery samples illustrating the diversity and complexity of the dataset: (**a**) normal illumination without occlusion; (**b**) high-intensity illumination under direct sunlight; (**c**) low-light condition at dusk; (**d**) occluded target partially covered by overhead objects such as power lines; and (**e**) shadowed target within the canopy shade of surrounding plants.

2.2. AMS-YOLO

YOLOv11 is a model series developed by Ultralytics that incorporates the C3K2 module into its backbone and neck structures to improve detection accuracy, inference speed, and overall efficiency. The series adopts an anchor-free design, features a decoupled detection head, and employs depthwise separable convolutions to minimize computational complexity without compromising performance. In this study, the lightweight variant YOLOv11n is selected as the baseline model due to its favorable balance between efficiency and detection accuracy, particularly for UAV-based small-object detection tasks. However, despite its advantages, YOLOv11n still encounters limitations in certain application scenarios. Specifically, it struggles to distinguish cannabis plants from visually similar background textures, which reduces its effectiveness in complex, vegetation-dense environments. Additionally, the model exhibits limited scale adaptability, resulting in suboptimal performance when detecting cannabis plants of varying sizes. While YOLOv11n offers a commendable trade-off between accuracy and computational cost, its streamlined architecture constrains its ability to extract rich visual details, which affects its effectiveness in cannabis identification.

In response to these challenges, this study introduces AMS-YOLO, an improved architecture built upon YOLOv11n. The optimized network structure is illustrated in Figure 4. AMS-YOLO introduces targeted refinements to the backbone and neck modules, improving both feature representation and multi-scale information integration. In the backbone, pinwheel-shaped convolution (PConv) replaces the shallowest convolutional block, and a four-direction asymmetric padding mechanism is applied to strengthen the model's ability to capture fine details, such as the serrated edges of leaves and palmate textures, thus providing a solid foundation for representing complex target features. Additionally, some C3K2 modules are replaced with Asymmetric Padding C2f (APC2f) modules, where the Asymmetric Padding Bottleneck (APBottleneck) in the module constructs anisotropic receptive fields through asymmetric padding, improving feature separation capabilities

while maintaining a lightweight design. In the neck network, hierarchical features extracted by the backbone are processed through the optimized multi-scale fusion (MSF) module to further enhance feature representation. This structure filters redundant channels at the input using partial convolutions and also applies partial convolutions at the output to strengthen feature reconstruction.

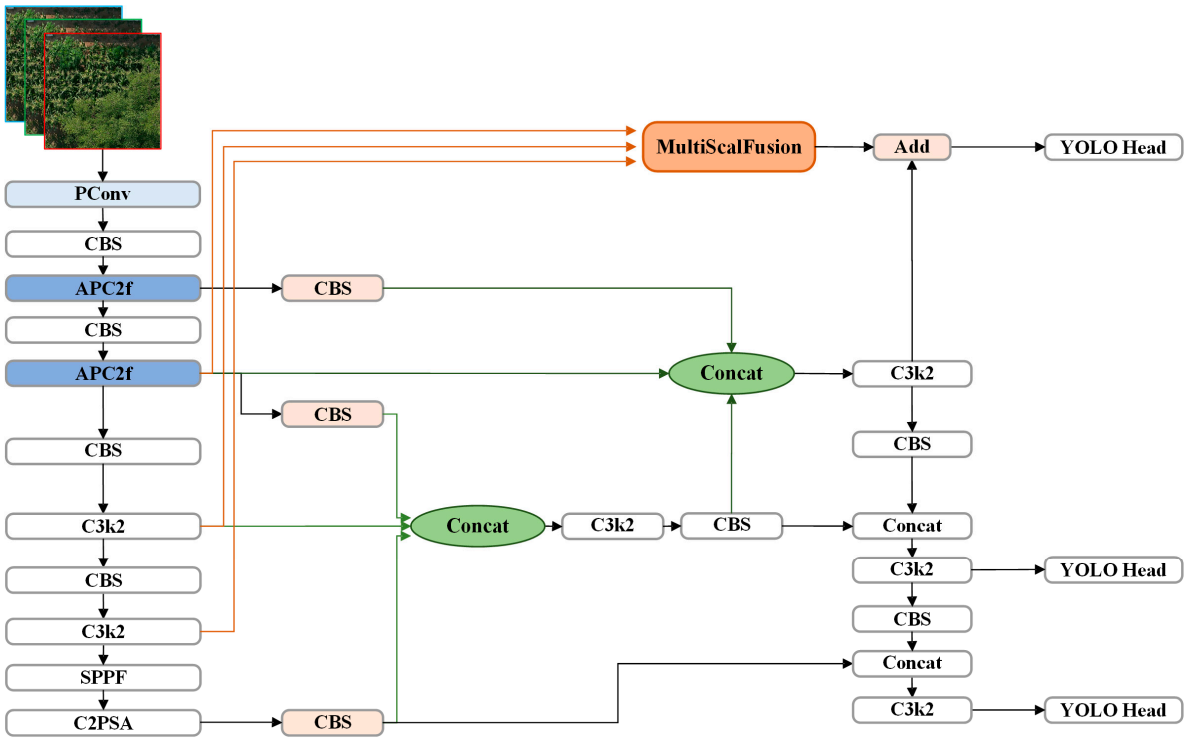

Figure 4. The AMS-YOLO architecture.

2.2.1. Efficient Asymmetric Backbone Network

In this study, the PConv and APC2f module structures are introduced into the backbone network, forming a new efficient asymmetric backbone network. This enhancement effectively improves the accuracy and robustness of YOLOv11n in cannabis detection tasks while maintaining model complexity. The modified network structure exhibits stronger perceptual capabilities for long-distance, low-contrast, and small-sized targets.

To enhance the detection capabilities of YOLOv11n for ground plant targets in complex environments, this study modifies the backbone structure by replacing the first Convolution-BatchNorm-SiLU module (CBS) with a PConv, which is directionally sensitive. Unlike traditional fixed-size convolutional kernels, PConv combines $1 \times k$ and $k \times 1$ directional convolutions, along with a four-direction asymmetric padding structure. This configuration provides better feedback on the texture features of the plant's central region, helping to improve the network's ability to distinguish plant leaf shape features, particularly in scenes with dense weeds or crowded vegetation. Furthermore, the introduction of PConv not only enhances the expression of lower-level features but also improves the model's ability to capture subtle differences between target plants and the surrounding weed background.

The structure of PConv is shown in Figure 5. The input feature map, with dimensions of $c \times h \times w$, undergoes asymmetric padding in four directions. The padding is applied as Padding (left, right, top, and bottom), indicating the number of pixels to be added on each side. The padded feature map is then processed in parallel through four convolution blocks, with output dimensions of c_m, using convolution kernels of size (1, k) or (k, 1) for

the PConv operation, as described in Equation (1). Finally, the resulting feature maps are concatenated, and a standard convolution is applied to adjust the dimensions.

$$Conv = SiLU(BN(Conv2d(c_{out}, k, s, p))) \quad (1)$$

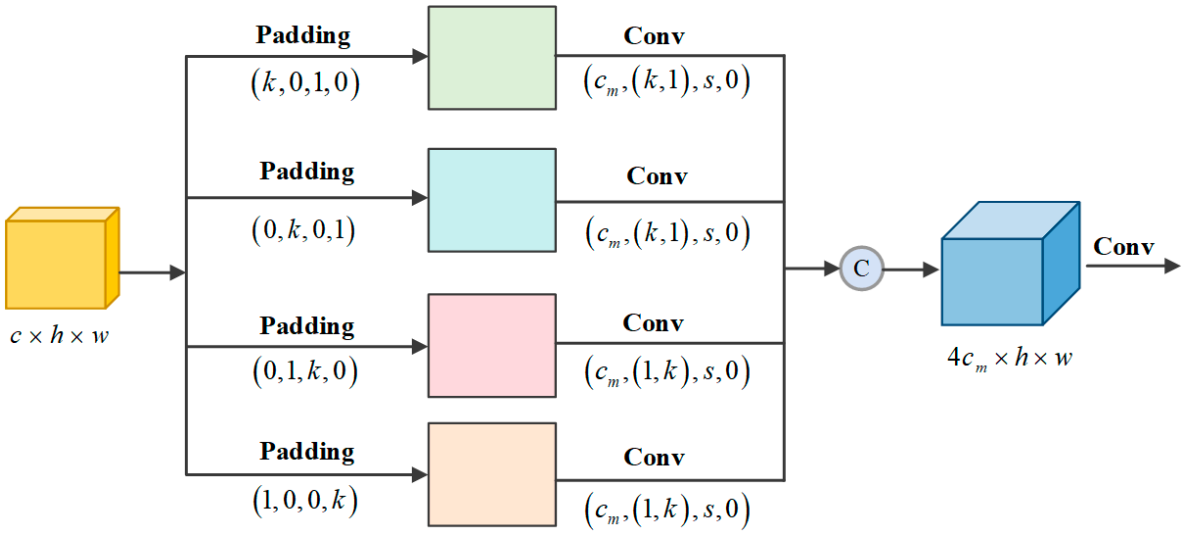

Figure 5. Structure of PConv.

In the YOLOv11n model, the C3K2 module enhances local feature fusion capabilities through grouped convolutions and residual connections, while using a bottleneck module for multi-scale feature extraction to process the input feature maps. However, traditional bottleneck structures typically include multiple convolutional layers, and their fixed receptive fields limit the detection performance of plant textures in complex backgrounds. The introduction of the APBottleneck structure effectively addresses this issue. As shown in Figure 6, compared to traditional bottleneck structures, APBottleneck offers a more efficient feature extraction approach. It is based on the PConv concept, incorporating asymmetric padding in four directions within the spatial dimensions to guide the network in extracting texture features of plant stems and leaves along different directions. This enables the model to capture edge and shape variations of the target. The four directional features are then concatenated and fused through parallel convolutions, resulting in a clearer and more focused representation of the plant target. The replacement of the standard bottleneck with the APBottleneck in the APC2f module enhances the extraction of fine plant texture features without significantly increasing the computational load, thus aiding the model's subsequent precise recognition.

To fully leverage the feature modeling capabilities of the APC2f module, this study replaces the first two C3K2 modules in the backbone network structure of YOLOv11n with APC2f modules. This modification enhances the network's directional selectivity and local responsiveness during multi-scale feature fusion. The decision to replace the first two C3K2 modules in the backbone network is based on their position in the shallow layers of the network, where they contain rich semantic information and are capable of effectively detecting small targets in large-size feature maps. The APC2f module helps mitigate challenges such as target occlusion, significantly improving the accuracy of target recognition, particularly for small targets under challenging conditions.

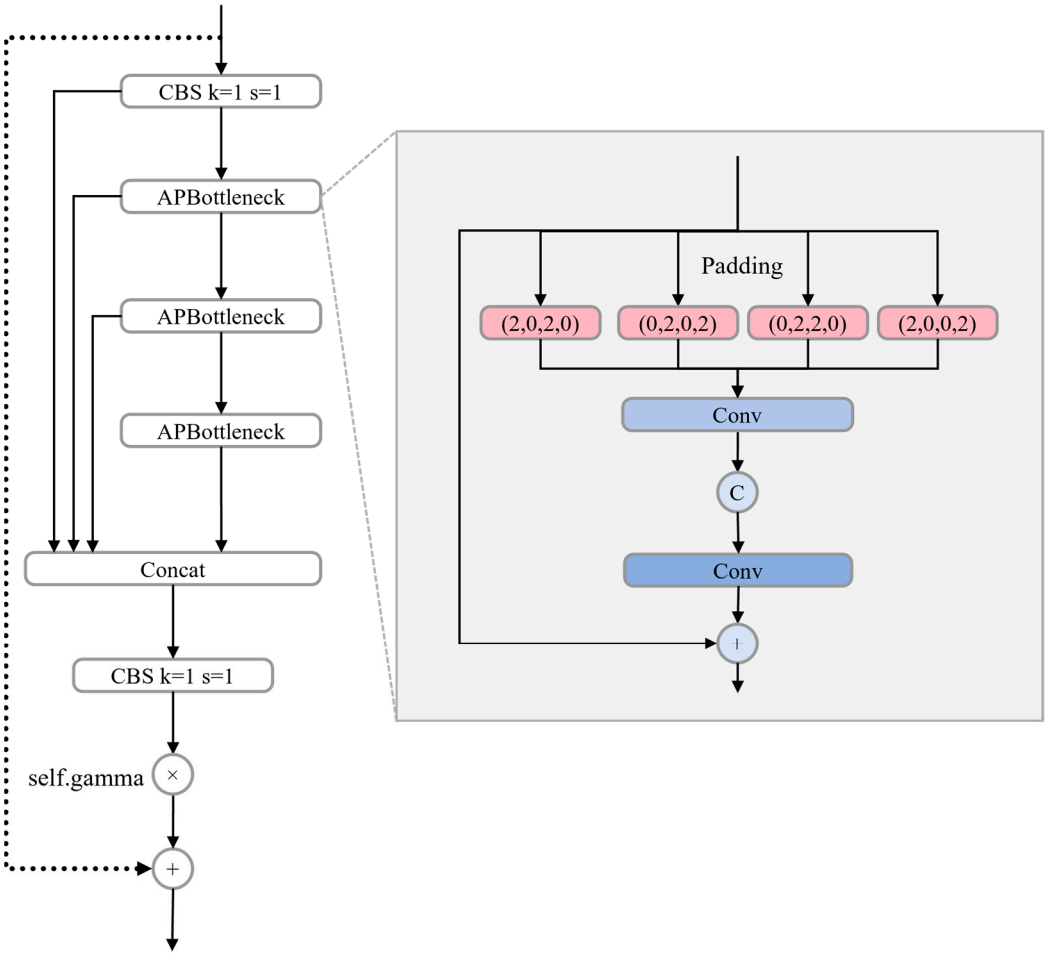

Figure 6. Structure of APC2f.

2.2.2. MSF-Type Neck Network

To address the issue of insufficient feature propagation for small and low-growing plant targets in traditional FPN-based plant detection, this study introduces MSF to construct a novel neck network architecture. The specific structure is illustrated in Figure 7. Specifically, we designed the MSF module to better combine multi-scale feature maps with detailed information from shallow feature maps of the same aspect ratio. This design is inspired by the SSFF module [38]. We first apply partial convolutions to the three feature maps, generated by the backbone, which contain different levels of detail and sizes. This approach reduces computational redundancy while maintaining efficient inference. Assuming the size of the input feature map is $h \times w \times c$, and c_p is the number of channels involved in the convolution operation, the size of the convolution kernel is $k \times k \times c_p$. The total computational load of partial convolution can be derived as $h \times w \times k^2 \times c_p^2$. Therefore, the ratio of the total computational load of partial convolution to that of standard convolution is $\frac{c_p^2}{c^2}$. Next, the feature maps are horizontally stacked and the nearest neighbor interpolation method is used to align all feature maps to the resolution size of the highest resolution feature map which contains most of the information necessary for more precise plant texture details and detection. Then, 3D convolutions are applied to extract the scale-sequence features of concatenated feature map. After 3D batch normalization and LeakyReLU activation, the features undergo 3D pooling and another partial convolution to complete scale-sequence feature extraction. This MSF structure is designed to address the challenge of inadequate multi-scale feature integration, particularly for small target detection.

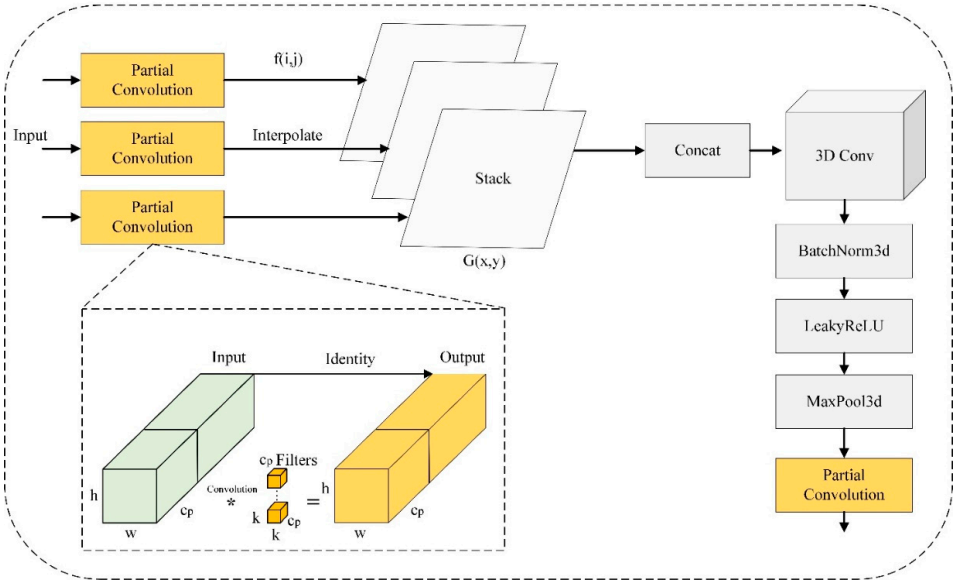

Figure 7. Structure of MSF.

2.3. Experimental Setting and Assessment Indicators

The configuration of the main experimental setup is provided in Table 2, while Table 3 presents the key hyperparameter settings.

Table 2. Configuration of experimental equipment environment.

Parameter	Parameter Value
Operating System	Ubuntu 22.04
Running Memory	120 GB
Graphics Card	GeForce RTX 4090
Video Memory	24 GB
Deep Learning Framework	Pytorch 2.1.0
CUDA Version	12.1
Python Version	3.10

Table 3. Hyperparameter settings.

Hyperparameter	Parameter Value
init learning	0.01
epoch	300
batchsize	16
Img size	640 × 640
Optimizer	SGD
Number of Multi-threads	16

To thoroughly assess the model's performance, this study utilizes several metrics: precision (P), recall (R), mean average precision (mAP), number of parameters (Params), and model size (Size).

Precision (P) is calculated as the ratio of true positive predictions to the total number of positive predictions, as shown by the following formula:

$$P = \frac{TP}{TP + FP} \tag{2}$$

Recall (R) reflects the model's ability to identify target samples, calculated as the ratio of correctly predicted positive samples to the total number of true positive samples, as expressed by the following formula:

$$R = \frac{TP}{TP + FN} \tag{3}$$

Here, TP, FP, and FN represent the numbers of true positives, false positives, and false negatives, respectively.

Mean average precision (mAP) is a crucial metric for assessing the model's overall detection performance. It is derived from the average precision (AP) for each category, where AP is the area under the precision–recall (P-R) curve, as shown in the following formula:

$$AP = \int_0^1 P(R)dR \tag{4}$$

$$mAP = \frac{1}{N}\sum_{i=0}^{N} AP_i \tag{5}$$

where AP_i represents the AP value of the i-th category and N is the number of categories in the training dataset. The mAP is commonly reported at different Intersection over Union (IoU) thresholds. Specifically, mAP@0.5 refers to the average of AP values calculated at an IoU threshold of 0.5, while mAP@0.5:0.95 represents the average AP across multiple IoU thresholds ranging from 0.5 to 0.95, in increments of 0.05. These metrics are used to assess how well a model can detect objects at different levels of overlap.

3. Results

3.1. Comparative Experiments of Different Models

To evaluate the effectiveness of the AMS-YOLO model, we compared it against several mainstream deep learning-based object detection algorithms, including Faster R-CNN, the YOLO series (v5, v6, v8n, v10n, and v11n), and RT-DETR-R18, which is the most lightweight variant within the RT-DETR family [39]. To ensure a fair comparison, all selected YOLO models have comparable parameter sizes and are suitable for practical deployment. Table 4 summarizes the performance of these models on the cannabis detection task, including precision (P), recall (R), mAP@0.5, mAP@0.5:0.95, parameter count, and model size.

Table 4. Experimental results of different models.

Model	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Parameters (M)	Size (MB)
Faster-RCNN	82.5	78.5	85.7	43.4	33.09	132.4
RT-DETR-R18	84.5	80.2	87.0	47.1	20.08	80.7
YOLOv5	84.0	80.4	87.4	47.7	2.50	5.3
YOLOv6	83.3	81.1	88.4	48.6	4.23	8.7
YOLOv8n	81.0	81.3	87.7	48.2	3.00	6.3
YOLOv10n	81.8	79.8	86.5	46.2	2.27	5.8
YOLOv11n	86.5	78.9	88.6	48.4	2.58	5.5
AMS-YOLO	87.6	79.8	90.7	49.2	3.02	6.4

The experimental results demonstrate that AMS-YOLO significantly improves detection performance. Its mAP@0.5 reaches 90.7%, representing a 2.1% improvement over the baseline model, YOLOv11n, and a 3.3–4.2% improvement over models such as YOLOv5 and YOLOv8n. In terms of balancing precision and recall, AMS-YOLO achieves a precision of 87.6% and a recall of 79.8%, showing improvements in both metrics compared

to YOLOv11n, which has a precision of 86.5% and a recall of 78.9%. This validates the effectiveness of the asymmetric backbone network, combined with the multi-scale fusion strategy, for better separation of target–background features. Furthermore, AMS-YOLO achieves the highest mAP@0.5:0.95 score of 49.2% among all compared models. This more stringent evaluation metric highlights the model's ability to maintain detection accuracy across a wider range of IoU thresholds. The 0.8% improvement over YOLOv11n, which achieves 48.4%, indicates that AMS-YOLO offers enhanced robustness in complex scenarios, particularly under partial occlusion and cluttered backgrounds.

From an efficiency perspective, AMS-YOLO has a parameter count of 3.02 million and a model size of 6.4 MB, which is slightly higher than YOLOv11n, with 2.58 million parameters and a model size of 5.5 MB. However, AMS-YOLO shows a significant improvement in parameter utilization. Each million parameters correspond to an mAP@0.5 of 30.0%, representing a 12.7% increase compared to YOLOv6, which has 4.23 million parameters and achieves an mAP of 88.4%. This improvement is attributed to the MSF-type neck network, which reduces computational redundancy through partial convolution, enhancing the model's ability to detect small targets while maintaining a lightweight design.

Compared to non-YOLO architectures, AMS-YOLO demonstrates superior overall performance and deployment efficiency. While Faster R-CNN achieves a recall of 78.5%, it falls short in both mAP@0.5, with a score of 85.7%, and mAP@0.5:0.95, at 43.4%. Additionally, it incurs significant computational costs due to its complex two-stage structure, with 33.09 million parameters and a model size of 132.4 MB. RT-DETR-R18, the most compact variant in the RT-DETR transformer series, offers higher accuracy than Faster R-CNN, with mAP@0.5 at 87.0% and mAP@0.5:0.95 at 47.1%. However, it still underperforms AMS-YOLO and is significantly heavier, with 20.08 million parameters and a model size of 80.7 MB. In contrast, AMS-YOLO achieves state-of-the-art accuracy, with mAP@0.5 at 90.7% and mAP@0.5:0.95 at 49.2%, all while maintaining a lightweight design with 3.02 million parameters and a model size of 6.4 MB. This is made possible by its directionally sensitive feature extraction and cross-scale fusion strategies. The optimal balance between precision and efficiency makes AMS-YOLO particularly well-suited for real-time, onboard deployment in UAV-based cannabis detection tasks, where resource constraints and inference speed are critical.

3.2. Visual Comparison of Test Results

To clearly highlight the AMS-YOLO model's superiority in cannabis detection, we visualized and compared its detection results. As shown in Figure 8, in a vegetation background with dense weeds, the original model produces significant false positives for weeds that have leaf shapes similar to cannabis. This phenomenon indicates that the original model lacks adequate directional sensitivity in plant texture, leading to a high overlap in the feature embedding space between weeds and cannabis. Moreover, the original model generates 3–4 overlapping detection boxes on individual cannabis plants, suggesting insufficient feature discrimination ability for similar targets or backgrounds. There are local similarities between the leaf textures and geometric shapes of weeds and cannabis. Due to the absence of a directional sensitivity mechanism in traditional convolution, the inter-class distance in the feature space becomes too small, causing the classifier to mistakenly identify different parts of the same plant as multiple targets, resulting in overlapping detection boxes. Additionally, the original model has a weak ability to represent detailed information from shallow features, leading to blurred edge features and exacerbating detection box localization errors.

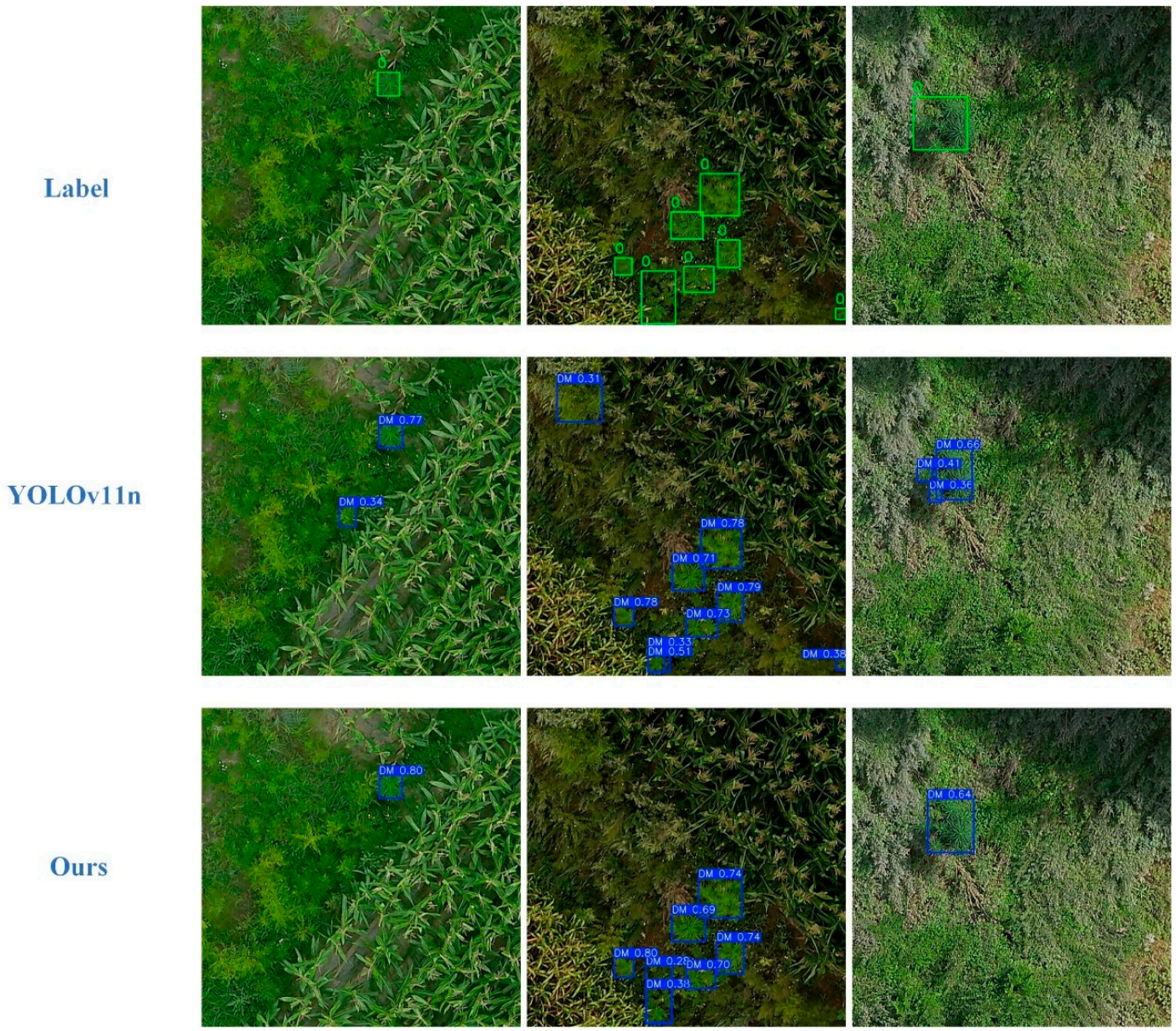

Figure 8. Detection effect comparison between YOLOv11n Model and AMS-YOLO.

To further quantify the model performance in complex scenarios, we manually reviewed the visualized predictions across all 339 images in the test set, comparing ground truth, YOLOv11n predictions, and AMS-YOLO outputs. A strict evaluation rule was adopted: if a single false positive or false negative occurred within an image, it was counted once toward the respective error type. Under this criterion, AMS-YOLO produced 10 false positive and 10 false negative samples, while YOLOv11n yielded 22 false positives and 17 false negatives. These results indicate that AMS-YOLO reduces both types of errors by over 40%, demonstrating greater robustness under challenging conditions such as occlusion, background interference, and intra-class variability. AMS-YOLO shows enhanced robustness and is more effective at identifying cannabis plants of different sizes, as well as distinguishing them from weed interference. These visualization results indicate that the proposed efficient asymmetric backbone and MSF-type neck network enhance the model's robustness in complex scenarios, thereby improving its suitability for practical cannabis detection tasks.

3.3. Ablation Experiments

Ablation experiments were performed on the cannabis dataset with the same experimental setup to evaluate the contribution of each improvement module to cannabis detection in UAV imagery. The goal was to assess the impact of the innovative modules on detecting cannabis plants.

3.3.1. Overall Ablation Experiment

To explore the contribution of different modules to the model's enhancement, the original YOLOv11n model was used as the baseline, and we gradually incorporated (1) APC2f; (2) PConv; and (3) the MSF-type neck network structure. The performance metrics of different models were then compared (a check mark (√) indicates that the corresponding improvement module was introduced, while a cross (×) indicates that it was not used).

As shown in Table 5, starting with the baseline model, YOLOv11n, which achieves a precision of 86.5%, recall of 78.9%, and mAP@0.5 of 88.6%, the first enhancement, APC2f, improves recall to 80.7% and mAP@0.5 to 89.4%, although precision decreases slightly. Further improvements by incorporating PConv in YOLOv11n-1-2 lead to a higher recall of 81.9% and mAP@0.5 of 89.8%, with a slight drop in precision to 84.0%. Finally, the full model, AMS-YOLO (YOLOv11n-1-2-3), which includes the MSF-type neck network structure, achieves the highest performance, with precision reaching 87.6%, recall at 79.8%, and mAP@0.5 at 90.7%. Although the parameter count and model size increase slightly, the enhanced model demonstrates superior detection accuracy, confirming the effectiveness of the combined modifications.

Table 5. Ablation experiment results.

Model	1	2	3	P (%)	R (%)	mAP@0.5 (%)	Parameters (M)	Size (MB)
YOLOv11n	×	×	×	86.5	78.9	88.6	2.58	5.5
YOLOv11n-1	√	×	×	86.0	80.7	89.4	2.64	5.6
YOLOv11n-1-2	√	√	×	84.0	81.9	89.8	2.65	5.6
YOLOv11n-1-2-3 (Ours)	√	√	√	87.6	79.8	90.7	3.02	6.4

3.3.2. Ablation of PConv Location

To systematically investigate the optimal deployment strategy of PConv in the backbone network, this study conducted a progressive replacement experiment based on a network architecture already integrated with the APC2f module. The experiment targeted the different hierarchical positions of the five CBS modules in the backbone network. The distribution of these five CBS modules is illustrated in Figure 9. In the implementation process, a univariate control approach was adopted: only one CBS module was replaced with a PConv unit at a time, while the structure, hyperparameters, and APC2f configuration of the other modules were kept unchanged. Independent training and evaluation were then performed on the cannabis detection dataset.

The experimental results demonstrate that the position of PConv in the backbone network significantly impacts the model's detection performance. As shown in Table 6, when the CBS module at position ① was replaced with PConv, the model's mAP@0.5 reached 89.8%, an improvement of 1.2% over the baseline model and 0.4% over the configuration with APC2f, making it the optimal deployment strategy. This position corresponds to the shallow feature extraction stage of the backbone network, where initial features rich in spatial details such as edges and textures are extracted. The four-direction asymmetric padding mechanism of PConv enhances the model's ability to capture fundamental mor-

phological features, such as the leaf extension direction and serrated edges of plants. This serves as a more discriminative foundation for the subsequent semantic feature fusion. Additionally, introducing PConv at this stage added just 0.01 M parameters (0.39% of the total), with the model size remaining unchanged at 5.6 MB.

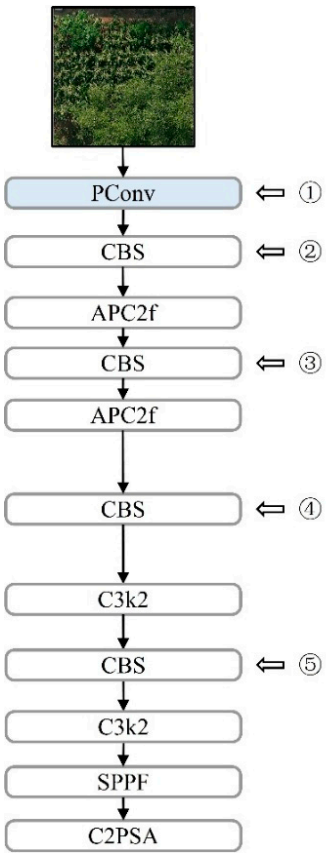

Figure 9. Position of PConv in backbone network structure.

Table 6. Ablation experiment results of PConv.

Model	P (%)	R (%)	mAP@0.5 (%)	Parameters (M)	Size (MB)
YOLOv11n (Baseline)	86.5	78.9	88.6	2.58	5.5
YOLOv11n + APC2f	86.0	80.7	89.4	2.64	5.6
YOLOv11n + APC2f + ①	84.0	81.9	89.8	2.65	5.6
YOLOv11n + APC2f + ②	84.9	81.0	88.2	2.64	5.6
YOLOv11n + APC2f + ③	83.1	83.0	89.4	2.63	5.6
YOLOv11n + APC2f + ④	83.9	79.6	88.5	2.59	5.5
YOLOv11n + APC2f + ⑤	85.1	80.9	88.5	2.66	5.7

Performance varied significantly at other positions: replacing the module at position ② resulted in a decrease in mAP@0.5 to 88.2%, as this stage had already entered the semantic fusion phase, and the multi-directional feature concatenation of PConv introduced redundancy in semantic information. When PConv was applied at position ④, mAP@0.5 dropped to 88.5%, below the baseline level, as the deep features were highly abstracted, and PConv's local directional sensitivity was less effective in enhancing high-level semantic representations, possibly introducing feature noise. Notably, when the module was replaced at position ③, the recall rate increased to 83.0%, indicating that PConv still has potential for enhancing small target feature capture. However, precision decreased to 83.1%, reflecting the need to further optimize the balance between directional feature enhancement and

background suppression. The experimental results indicate that replacing the original CBS convolution with PConv at position ① achieves the best detection performance for the model.

3.3.3. Ablation of MSF-Type Neck Network

To validate the effectiveness of the proposed MSF-type neck network, we conducted a comparative performance experiment on different MSF structures in the neck network, building upon the previously integrated PConv and APC2f models. In the original MSF structure (MSF_p0), standard convolution is used for feature aggregation, and after fusion of multi-scale features through 3D convolutions, the output is directly produced, as shown in Figure 10. Subsequently, partial convolution modules are introduced at different positions within the original MSF, with partial convolution applied on the input side as MSF_p1 and on the output side as MSF_p2, as shown in Figures 11 and 12, respectively. Both MSF_p1 and MSF_p2, along with our final MSF, replace the original ReLU function with LeakyReLU.

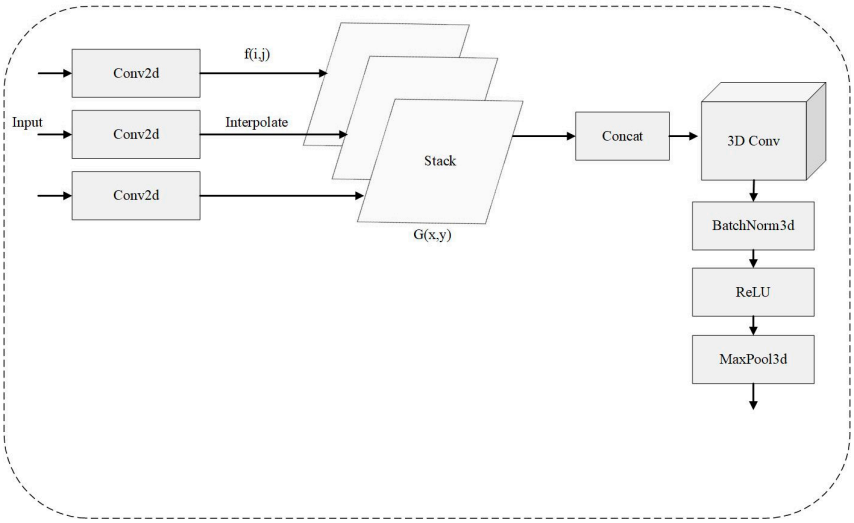

Figure 10. Original MSF_p0 structure.

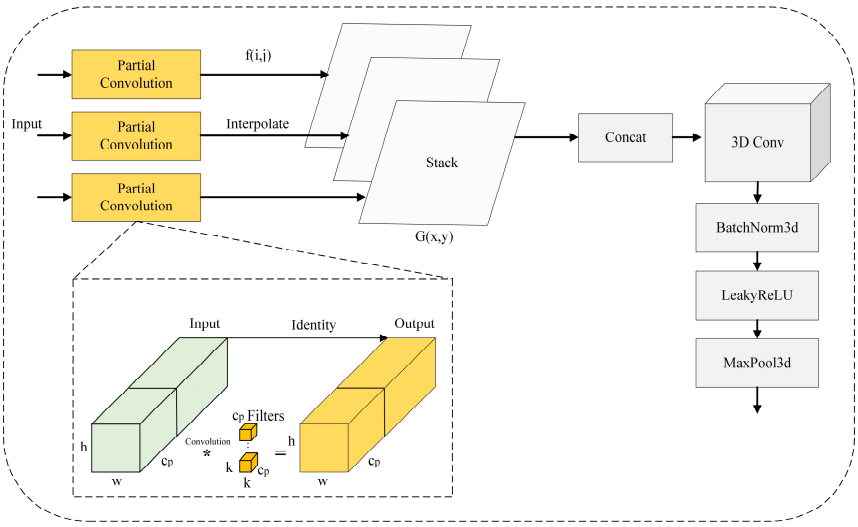

Figure 11. MSF_p1 structure.

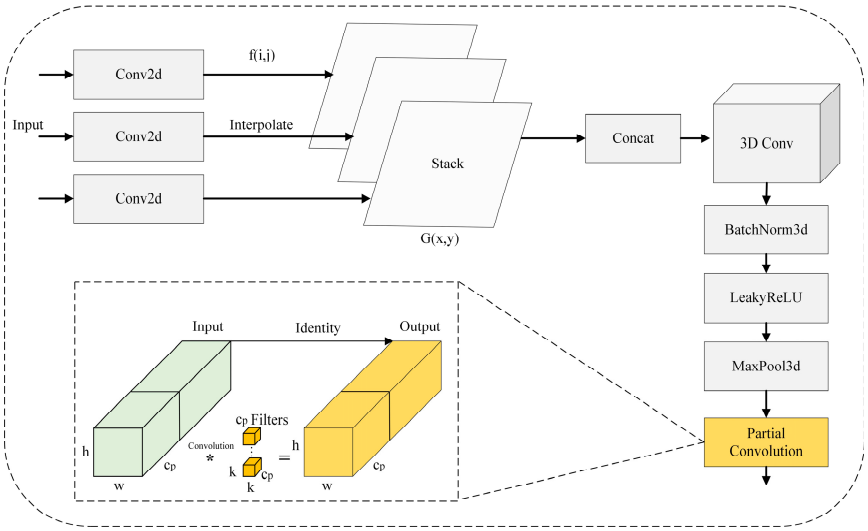

Figure 12. MSF_p2 structure.

As shown in Table 7, all the improved MSF structures outperform the original design: MSF_p1 achieves a 0.4% improvement in mAP@0.5 over MSF_p0, indicating that the input-side partial convolution effectively filters redundant features through channel-selective computations, enhancing the cross-scale transmission efficiency of small target features. MSF_p2 further increases mAP@0.5 to 89.8%, validating the enhanced feature reconstruction capability of partial convolution at the output side. The MSF configuration that introduces partial convolution at both ends achieves a mAP@0.5 of 90.7%, a 2.1% improvement over the baseline. This gain results from the bidirectional optimization mechanism: the input side reduces feature redundancy, while the output side enhances feature representation, collaboratively addressing the small target feature degradation issue in the original neck network. The synergistic effect of partial convolutions at both ends of feature fusion enables a dual improvement in detection accuracy and background suppression while maintaining computational efficiency.

Table 7. Ablation experimental results of different MSF neck network structures.

Model	P (%)	R (%)	mAP@0.5 (%)	Parameters (M)	Size (MB)
YOLOv11n (Baseline)	86.5	78.9	88.6	2.58	5.5
YOLOv11n + MSF_p0	84.3	82.1	89.1	3.02	6.4
YOLOv11n + MSF_p1	86.0	80.6	89.5	3.02	6.4
YOLOv11n + MSF_p2	85.9	81.1	89.8	3.02	6.5
YOLOv11n + MSF	87.6	79.8	90.7	3.02	6.4

4. Discussion

In this study, AMS-YOLO demonstrated superior performance in detecting illegal cannabis cultivation from UAV imagery, combining high detection accuracy with lightweight design suitable for real-world deployment. Compared with YOLOv11n, AMS-YOLO improved mAP@0.5 from 88.6% to 90.7% and outperformed other lightweight models such as YOLOv5 and YOLOv8n by 3.3–4.2%. These improvements highlight the contribution of the asymmetric backbone network and the multi-scale feature fusion strategy in enhancing target–background feature separation, thereby improving detection capability in complex vegetative environments.

In terms of accuracy metrics, AMS-YOLO achieved the highest precision (87.6%) among all compared models and a recall of 79.8%. High precision is particularly impor-

tant in cannabis detection, as it helps minimize false positives, which can cause significant disruption in practical applications. Furthermore, AMS-YOLO achieved the highest mAP@0.5:0.95 score of 49.2%, indicating improved localization consistency under more stringent IoU thresholds. From an efficiency standpoint, although AMS-YOLO has a slightly higher parameter count (3.02 M) and model size (6.4 MB) compared to YOLOv11n, it achieves a parameter utilization of 30.0% mAP@0.5 per million parameters. This represents a 12.7% improvement over YOLOv6, highlighting its superior efficiency for lightweight deployment.

Nevertheless, the model still exhibits limitations in certain real-world scenarios. Manual inspection of failure cases revealed three primary situations where AMS-YOLO struggled: (1) when weed species exhibit palmate compound leaves with serrated edges in the imagery, their visual resemblance to cannabis foliage may lead to occasional false positives; (2) when cannabis plants displayed atypical textures in imagery, for instance due to wind-induced leaf deformation; and (3) when targets appeared at the image boundaries and were partially truncated, leading to missed detections. These findings indicate that the model remains susceptible to background confusion, morphological variability, and boundary effects. Possible remedies include incorporating advanced attention mechanisms, boundary-aware detection modules, or multi-view image augmentation to enhance discrimination and context modeling.

Beyond YOLO variants, a broader comparison with alternative approaches could offer additional insights. Faster R-CNN, due to its two-stage detection structure, incurs substantial computational overhead and slower inference, limiting its suitability for real-time UAV-based applications despite reasonable accuracy. RT-DETR-R18, the most lightweight variant in the RT-DETR transformer series, benefits from enhanced context modeling but still lacks competitiveness in overall efficiency and compactness. These comparisons underscore the strengths of AMS-YOLO's end-to-end, lightweight architecture, emphasizing its suitability for practical deployment in UAV-based cannabis monitoring applications.

The present study did not directly evaluate performance under adverse weather or low-light conditions. Although the dataset includes variations such as overcast skies, strong sunlight, and dusk illumination, no data were collected during extreme weather events (e.g., rain, fog, or haze). Such adverse conditions can reduce image contrast, blur fine structural details, and increase background noise. This is particularly challenging for detecting cannabis plants, whose morphology and coloration closely resemble surrounding vegetation, thereby greatly increasing the difficulty of accurate identification. Future work may address this limitation by employing simulated weather augmentation, image enhancement techniques, or synthetic data generation to assess and improve model robustness under these challenging conditions.

Finally, this study has certain experimental limitations. The dataset, while diverse in geography and illumination, remains relatively small due to the difficulty of collecting high-quality, verified UAV cannabis imagery. This may affect the generalization ability of the model to new regions or unseen cultivation patterns. Future work will focus on expanding the dataset in collaboration with relevant authorities, integrating cross-regional imagery, and exploring domain adaptation techniques to further enhance robustness and applicability.

In conclusion, AMS-YOLO demonstrates that targeted architectural enhancements—specifically the asymmetric backbone and multi-scale feature fusion strategy—can effectively improve both accuracy and efficiency in UAV-based cannabis detection. By enhancing target–background feature separation, the model addresses core challenges such as complex vegetative environments, small-object characteristics, and variable illumination, outperforming several mainstream detection frameworks and offering practical value for

intelligent law enforcement monitoring. Future work will aim to expand the dataset in scale and geographic diversity, incorporate multi-regional imagery under varied acquisition conditions, and explore architectural refinements such as attention mechanisms and domain adaptation to enhance robustness against morphological variability, boundary truncation, and background confusion. Furthermore, simulating adverse weather and low-light conditions through augmentation or synthetic data generation may help assess and improve model resilience in real-world deployments, paving the way for AMS-YOLO to evolve into a reliable, field-ready solution for aerial surveillance applications.

5. Conclusions

Utilizing UAVs to patrol cannabis cultivation areas has become an important method for combating illegal cannabis farming. However, most current detection methods still rely on manual visual interpretation, which is inefficient, highly subjective, and has difficulty meeting the demands for large-scale, accurate monitoring. To address this issue, this study proposes an object detection model based on convolutional neural networks—AMS-YOLO—to automatically detect cannabis plants in UAV remote sensing imagery. The model is an improvement over the YOLOv11n framework, achieving a detection accuracy of 90.7% on the test dataset, significantly enhancing both detection efficiency and automation. Through comparative experiments with several classical object detection algorithms on the same dataset, this paper systematically analyzes the performance differences of various models in the cannabis detection task. The results demonstrate that AMS-YOLO, without significantly increasing computational load, achieves a 2.1% accuracy improvement over the existing best model, offering both high detection speed and strong robustness, suggesting its potential applicability in practical scenarios. Future research will focus on further optimizing the model structure, with particular attention to the impact of challenging factors such as complex backgrounds and occlusion on detection performance, to continuously improve the model's accuracy and stability in real-world applications.

Author Contributions: Conceptualization, X.L. and J.L.; methodology, X.L., J.L. and A.C.; investigation, J.L. and X.L.; resources, J.L. and A.C.; data curation, X.L. and A.C.; writing—original draft preparation, X.L.; writing—review and editing, J.L. and A.C.; project administration, H.Y. and J.L.; funding acquisition, H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Key Research and Development Program of China (No. 2023YFB3905705), the Strategic Priority Research Program of Chinese Academy of Sciences (No. XDB0740100), the Key Research and Development Program of Guangxi (GuikeAB25069501), and the Science, Technology and Innovation Bureau of Shenzhen Municipality under the program "Undertaking the National Key Research and Development Program of China" (Grant No. CJGJZD20240729141101003), Shenzhen, China.

Data Availability Statement: The original contributions presented in the manuscript are included in the article. The data and code used in this study are available at https://github.com/lx0106/AMS-YOLO. Any further inquiries can be directed to the corresponding author. Accessed on 4 August 2025.

Conflicts of Interest: Authors Jianli Liu and Aonan Cheng were employed by the company China TopRS Technology Company Limited. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Blanco, C.; Hasin, D.S.; Wall, M.M.; Flórez-Salamanca, L.; Hoertel, N.; Wang, S.; Kerridge, B.T.; Olfson, M. Cannabis Use and Risk of Psychiatric Disorders: Prospective Evidence From a US National Longitudinal Study. *JAMA Psychiatry* **2016**, *73*, 388–395. [CrossRef] [PubMed]
2. Ganesh, S.; Cortes-Briones, J.; Ranganathan, M.; Radhakrishnan, R.; Skosnik, P.D.; D'Souza, D.C. Psychosis-Relevant Effects of Intravenous Delta-9-Tetrahydrocannabinol: A Mega Analysis of Individual Participant-Data from Human Laboratory Studies. *Int. J. Neuropsychopharmacol.* **2020**, *23*, 559–570. [CrossRef]
3. Pereira, J.F.Q.; Pimentel, M.F.; Amigo, J.M.; Honorato, R.S. Detection and Identification of *Cannabis sativa* L. Using near Infrared Hyperspectral Imaging and Machine Learning Methods. A Feasibility Study. *Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* **2020**, *237*, 118385. [CrossRef]
4. Sujud, L.; Jaafar, H.; Haj Hassan, M.A.; Zurayk, R. Cannabis Detection from Optical and RADAR Data Fusion: A Comparative Analysis of the SMILE Machine Learning Algorithms in Google Earth Engine. *Remote Sens. Appl. Soc. Environ.* **2021**, *24*, 100639. [CrossRef]
5. Bicakli, F.; Kaplan, G.; Alqasemi, A.S. *Cannabis sativa* L. Spectral Discrimination and Classification Using Satellite Imagery and Machine Learning. *Agriculture* **2022**, *12*, 842. [CrossRef]
6. Wang, X.; Wang, A.; Yi, J.; Song, Y.; Chehri, A. Small Object Detection Based on Deep Learning for Remote Sensing: A Comprehensive Review. *Remote Sens.* **2023**, *15*, 3265. [CrossRef]
7. Wu, B.; Zhang, M.; Zeng, H.; Tian, F.; Potgieter, A.B.; Qin, X.; Yan, N.; Chang, S.; Zhao, Y.; Dong, Q.; et al. Challenges and Opportunities in Remote Sensing-Based Crop Monitoring: A Review. *Natl. Sci. Rev.* **2023**, *10*, nwac290. [CrossRef] [PubMed]
8. Deng, J.; Wang, R.; Yang, L.; Lv, X.; Yang, Z.; Zhang, K.; Zhou, C.; Pengju, L.; Wang, Z.; Abdullah, A.; et al. Quantitative Estimation of Wheat Stripe Rust Disease Index Using Unmanned Aerial Vehicle Hyperspectral Imagery and Innovative Vegetation Indices. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–11. [CrossRef]
9. Dong, Y.; Ma, Y.; Li, Y.; Li, Z. High-Precision Real-Time UAV Target Recognition Based on Improved YOLOv4. *Comput. Commun.* **2023**, *206*, 124–132. [CrossRef]
10. Tanaka, T.S.T.; Wang, S.; Jørgensen, J.R.; Gentili, M.; Vidal, A.Z.; Mortensen, A.K.; Acharya, B.S.; Beck, B.D.; Gislum, R. Review of Crop Phenotyping in Field Plot Experiments Using UAV-Mounted Sensors and Algorithms. *Drones* **2024**, *8*, 212. [CrossRef]
11. Aierken, N.; Yang, B.; Li, Y.; Jiang, P.; Pan, G.; Li, S. A Review of Unmanned Aerial Vehicle Based Remote Sensing and Machine Learning for Cotton Crop Growth Monitoring. *Comput. Electron. Agric.* **2024**, *227*, 109601. [CrossRef]
12. Xia, C.; Wang, X.; Lv, F.; Hao, X.; Shi, Y. ViT-CoMer: Vision Transformer with Convolutional Multi-Scale Feature Interaction for Dense Predictions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–21 June 2024; IEEE: New York, NY, USA, 2024; pp. 5493–5502.
13. Liu, K.; Fu, Z.; Jin, S.; Chen, Z.; Zhou, F.; Jiang, R.; Chen, Y.; Ye, J. ESOD: Efficient Small Object Detection on High-Resolution Images. *IEEE Trans. Image Process.* **2025**, *34*, 183–195. [CrossRef]
14. Yang, J.; Liu, S.; Wu, J.; Su, X.; Hai, N.; Huang, X. Pinwheel-Shaped Convolution and Scale-Based Dynamic Loss for Infrared Small Target Detection. In Proceedings of the AAAI Conference on Artificial Intelligence; AAAI Press: Palo Alto, CA, USA, 2025; Volume 39, pp. 9202–9210.
15. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.-H.; Chan, S.-H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; IEEE: New York, NY, USA, 2023; pp. 12021–12031.
16. Hou, Q.; Lu, C.-Z.; Cheng, M.-M.; Feng, J. Conv2Former: A Simple Transformer-Style ConvNet for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 8274–8283. [CrossRef] [PubMed]
17. Yu, F.; Koltun, V. Multi-Scale Context Aggregation by Dilated Convolutions 2016. In Proceedings of the International Conference on Learning Representations (ICLR), San Juan, Puerto Rico, 2–4 May 2016; pp. 1–13.
18. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861. [CrossRef]
19. Rao, Y.; Zhao, W.; Liu, B.; Lu, J.; Zhou, J.; Hsieh, C.-J. DynamicViT: Efficient Vision Transformers with Dynamic Token Sparsification. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual, 6–14 December 2021; pp. 13937–13949.
20. Qiao, S.; Chen, L.-C.; Yuille, A. DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; IEEE: New York, NY, USA, 2021; pp. 10208–10219.
21. Pu, Y.; Wang, Y.; Xia, Z.; Han, Y.; Wang, Y.; Gan, W.; Wang, Z.; Song, S.; Huang, G. Adaptive Rotated Convolution for Rotated Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; IEEE: New York, NY, USA, 2023; pp. 6566–6577.

22. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; IEEE: New York, NY, USA, 2014; pp. 580–587.
23. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE: New York, NY, USA, 2015; pp. 1440–1448.
24. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
25. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; IEEE: New York, NY, USA, 2018; pp. 6154–6162.
26. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
27. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 26 June–1 July 2016; IEEE: New York, NY, USA, 2016; pp. 779–788.
28. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
29. Khanam, R.; Hussain, M. What Is YOLOv5: A Deep Look into the Internal Features of the Popular Object Detector. *arXiv* **2024**, arXiv:2407.20892. [CrossRef]
30. Yaseen, M. What Is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. *arXiv* **2024**, arXiv:2408.15857.
31. Khanam, R.; Hussain, M. YOLOv11: An Overview of the Key Architectural Enhancements. *arXiv* **2024**, arXiv:2410.17725. [CrossRef]
32. Lei, M.; Li, S.; Wu, Y.; Hu, H.; Zhou, Y.; Zheng, X.; Ding, G.; Du, S.; Wu, Z.; Gao, Y. YOLOv13: Real-Time Object Detection with Hypergraph-Enhanced Adaptive Visual Perception. *arXiv* **2025**, arXiv:2506.17733.
33. Wang, C.; Wang, Q.; Wu, H.; Zhao, C.; Teng, G.; Li, J. Low-Altitude Remote Sensing Opium Poppy Image Detection Based on Modified YOLOv3. *Remote Sens.* **2021**, *13*, 2130. [CrossRef]
34. Zhou, J.; Tian, Y.; Yuan, C.; Yin, K.; Yang, G.; Wen, M. Improved UAV Opium Poppy Detection Using an Updated YOLOv3 Model. *Sensors* **2019**, *19*, 4851. [CrossRef]
35. Wang, Q.; Wang, C.; Wu, H.; Zhao, C.; Teng, G.; Yu, Y.; Zhu, H. A Two-Stage Low-Altitude Remote Sensing Papaver Somniferum Image Detection System Based on YOLOv5s+DenseNet121. *Remote Sens.* **2022**, *14*, 1834. [CrossRef]
36. He, Q.; Zhang, Y.; Liang, L. Identification of Poppy by Spectral Matching Classification. *Optik* **2020**, *200*, 163445. [CrossRef]
37. Rominger, K.; Meyer, S.E. Application of UAV-Based Methodology for Census of an Endangered Plant Species in a Fragile Habitat. *Remote Sens.* **2019**, *11*, 719. [CrossRef]
38. Kang, M.; Ting, C.-M.; Ting, F.F.; Phan, R.C.-W. ASF-YOLO: A Novel YOLO Model with Attentional Scale Sequence Fusion for Cell Instance Segmentation. *Image Vis. Comput.* **2024**, *147*, 105057. [CrossRef]
39. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-Time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 16–21 June 2024; IEEE: New York, NY, USA, 2024; pp. 16965–16974.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.