

Article

Precision Greenhouse Rose Phenotyping from UAV Imagery Using a Multi-Source Dataset and Lightweight BloomRoseNet

Yingchao Wang ¹ , Jun Hao ² , Peng Zhou ¹ , Wei Chen ¹, Shan Sun ¹, Na Li ^{1,*} , Feng Xue ³ , Zixiang Qin ^{3,4}, Hao Wu ⁵ and Fan Zhao ⁶ 

¹ Department of Computer Science and Technology, Xinjiang College of Science & Technology, Korla 841000, China

² School of Artificial Intelligence and Computer Science, Hubei Normal University, Huangshi 435002, China

³ Graduate School of Information, Production and Systems, Waseda University, Kitakyushu 808-0135, Japan

⁴ College of Oceanography and Space Informatics, China University of Petroleum (East China), Qingdao 266580, China

⁵ Graduate School of Information Science and Technology, The University of Tokyo, Tokyo 113-8656, Japan

⁶ Graduate School of Frontier Sciences, The University of Tokyo, Kashiwa 277-8564, Japan; zhaofan25ut@163.com

* Correspondence: lina_xjust@163.com

Highlights

What are the main findings?

- A multi-source GreenHouse Rose dataset was constructed by integrating UAV overhead rose images, screened RoseTracker images, and supplementary multi-view rose images to improve data diversity for greenhouse rose detection.
- The proposed lightweight BloomRoseNet achieved improved rose detection performance, reaching 86.1% precision, 85.3% recall, and 90.5% mAP@50, while maintaining only 2.08 M parameters and a 4.5 MB model size.

What are the implications of the main findings?

- The proposed framework supports efficient UAV-based greenhouse rose phenotyping and can assist flowering assessment, harvest planning, and precision cultivation management.
- Its robustness under blur, occlusion, and illumination variation indicates practical potential for real-time greenhouse monitoring under non-ideal imaging conditions.

Abstract

Accurate detection of blooming roses and flower buds is essential for greenhouse phenotyping, cultivation scheduling, harvest planning, and yield management. However, UAV-derived greenhouse imagery presents major challenges because rose targets are often small, densely distributed, partially occluded, and visually similar to complex backgrounds. This study proposes a lightweight rose detection framework that combines multi-source dataset construction with an improved YOLOv12n-based detector, termed BloomRoseNet. A *GreenHouse Rose dataset* was constructed by integrating self-collected UAV overhead images, screened RoseTracker images, and supplementary multi-view rose images to increase diversity in scale, growth stage, viewpoint, and background complexity. BloomRoseNet introduces task-oriented improvements for fine-grained feature extraction, adaptive feature fusion, and attention-enhanced detection. The supplementary multi-view data improved precision, recall, and mAP@50 from 85.2%, 82.8%, and 89.2% to 86.1%, 85.3%, and 90.5%, respectively. Compared with the baseline YOLOv12n, BloomRoseNet increased precision, recall, mAP@50, and mAP@50:95 by 3.2, 3.3, 3.6, and 1.6 percentage points, respectively, while reducing parameters from 2.55 M to 2.08 M and model size from 5.5 MB to 4.5 MB.



Academic Editor: Qiang Li

Received: 26 June 2026

Revised: 1 August 2026

Accepted: 3 August 2026

Published: 11 August 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article

distributed under the terms and

conditions of the [Creative Commons](https://creativecommons.org/licenses/by/4.0/)

[Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

The model also maintained real-time inference capability and stronger robustness under blur, occlusion, and illumination disturbances. The proposed framework provides an effective and practical solution for UAV-based greenhouse rose monitoring and supports precision cultivation management.

Keywords: greenhouse rose detection; UAV imagery; multi-source dataset; BloomRoseNet; small object detection

1. Introduction

1.1. Background and Environmental Relevance

The floriculture industry is an important part of modern horticulture and contributes substantially to agricultural economy, rural employment, and the production of high-value crops worldwide [1,2]. Among ornamental crops, roses are one of the most commercially important cut flowers because of their high market demand, aesthetic value, and extensive role in the global flower supply chain [3,4]. In greenhouse rose production, accurate knowledge of flowering dynamics, including the spatial distribution and developmental status of buds and blooming flowers, is essential for cultivation scheduling, harvest planning, yield estimation, and labor allocation [5,6]. Reliable monitoring of these phenological traits can therefore directly improve production efficiency and product quality.

From an environmental and management perspective, intelligent rose monitoring also has broader significance for sustainable protected agriculture. Conventional greenhouse management still relies heavily on manual scouting and visual counting. These practices are labor-intensive, time-consuming, and susceptible to subjective bias [7,8]. They are also increasingly inadequate under the growing demand for precision horticulture, in which growers need accurate crop-status information to optimize irrigation, fertilization, temperature regulation, and harvest timing [5]. Automated phenotyping methods can reduce unnecessary labor, improve the timing of management interventions, and support more efficient use of agricultural resources. In this sense, the development of robust image-based monitoring systems is not only a technical issue, but also an important step toward data-driven and environmentally sustainable greenhouse production.

In recent years, unmanned aerial vehicles (UAVs) have emerged as promising platforms for agricultural monitoring because they enable rapid, flexible, and relatively low-cost data acquisition over large cultivation areas [9–13]. Compared with fixed cameras or manual close-up imaging [14], UAVs provide a more comprehensive view of the spatial distribution of crop canopies and reproductive organs, especially when the captured images are further processed into orthomosaics. In greenhouse rose production, UAV-based orthomosaic mapping enables whole-area assessment rather than fragmented local observation. This is highly valuable for identifying flowering hotspots, comparing cultivation beds, and quantifying flower abundance at different developmental stages. Recent international studies have further demonstrated the potential of UAV imagery and deep learning for flower mapping across different vegetation and cultivation environments [12,13]. However, their results also indicate that dense vegetation, environmental variation, and small floral targets can substantially reduce detection reliability. Moreover, Zhao et al. [15] and Yang et al. [16] showed that UAV-based crop imaging can face substantial challenges, including motion blur, geometric distortion, target-scale variation, and reduced visibility of small floral structures in dense and cluttered scenes. These limitations highlight the need for improved visual detection methods tailored to rose phenotyping.

1.2. Advances in Flower Detection and Rose Monitoring

The application of computer vision to flower detection and developmental stage recognition has progressed rapidly in precision agriculture [17–21]. Early studies mainly focused on close-range images and relatively simple scenes. More recent research has shifted toward more complex field or greenhouse environments with dense targets, occlusion, and strong background interference [22,23]. Deep learning, especially convolutional neural network (CNN)-based object detection [24,25], has greatly accelerated this progress by enabling end-to-end learning of discriminative features from agricultural imagery [26].

Recent international studies have also explored new data and learning strategies for flower detection. Vierbergen et al. [17] used synthetic images generated from a simulated environment to improve *Calendula* flower detection, while Bhattarai et al. [26] investigated weakly supervised flower and fruit localization in complex orchard canopies. These studies show that data diversity and effective representation of dense floral targets are important for reliable detection.

A large body of related work has demonstrated the feasibility of flower detection in multiple crops. Estrada et al. [27] proposed a deep-learning-based method to detect and count peach flowers in orchard environments, where targets are numerous, small, and highly overlapped. However, orchard-based peach flower detection is still different from greenhouse rose monitoring, where viewpoint constraints, canopy density, and the need for stage-specific discrimination introduce additional complexity. Similar efforts have also been reported by Rahim et al. [28] for grapevine inflorescence quantification, in which lightweight detectors achieved promising performance and cross-variety adaptability. Nevertheless, that task focuses on inflorescence quantification rather than the fine-grained distinction between buds and blooming flowers. Xu et al. [29] further explored deep-learning-based analysis of chrysanthemum inflorescence morphology, showing the potential of data-driven methods for flowering-crop analysis. But the study focused more on morphology feature extraction than on robust detection of small and densely distributed floral targets. Although these studies focus on different species, they provide important methodological support for rose monitoring. In particular, they address common visual challenges such as small-object detection, partial occlusion, illumination variability, and the coexistence of multiple phenological stages.

Nevertheless, chrysanthemum detection and rose detection involve different visual characteristics. Chrysanthemum studies generally focus on visible inflorescences with relatively concentrated floral structures or on the classification of overall flowering degree [19,20,29]. In contrast, rose appearance changes continuously from compact buds to partially and fully opened flowers. Roses also have more irregular petal boundaries and are frequently tilted, overlapped, or hidden by leaves and branches. Small rose buds may resemble young leaves or folded leaf tips. Therefore, greenhouse rose monitoring combines small-object detection with fine-grained discrimination between closely related developmental stages.

Compared with general flower detection, rose detection in greenhouse environments is more challenging. In many cases, the task requires not only locating floral targets but also distinguishing buds from blooming flowers. This distinction is critical for practical rose cultivation because cut-flower management and yield planning depend strongly on developmental stage [30,31]. However, buds and blooming flowers often show high intra-class variation and low inter-class separability, especially under side views, partial occlusion, or nonuniform illumination. Small buds may also resemble surrounding leaves, stems, or immature floral structures, which makes reliable detection particularly difficult [32].

Several studies have laid the foundation for automated rose monitoring. RoseTracker proposed a stage-specific rose monitoring framework and released the RoseBlooming

dataset. It demonstrated that overhead imagery combined with object detection and tracking can improve rose counting and reduce omission errors compared with single-image approaches [2]. That work emphasized the importance of growth-stage monitoring in cut-flower cultivation and highlighted the practical value of overhead rose observation. At the same time, it also pointed out that publicly available rose datasets for stage-specific detection remain limited, especially under practical farm conditions. In addition, the study showed that small rose buds are frequently hidden by overgrown leaves and that multi-view or video-based analysis can help recover targets missed in individual frames. However, RoseTracker was mainly developed from overhead images collected under a specific greenhouse setting, and its temporal tracking strategy relies on consecutive image frames [2]. The study did not fully examine cross-source image diversity or independent detection in UAV-derived orthomosaic patches. Its applicability to roses captured by different platforms, viewpoints, and cultivation environments therefore remains insufficiently evaluated. These findings are highly relevant to greenhouse rose orthomosaic analysis, where dense foliage and overlapping branches also hinder the reliable identification of small targets.

More recent studies have extended rose monitoring from local overhead imagery to UAV-assisted large-area observation. For example, Xu et al. [33] used UAV images to detect female cucumber flowers in greenhouse environments, showing the potential of aerial observation for flower monitoring at a broader spatial scale. However, the target species, canopy structure, and phenological requirements differ considerably from those of greenhouse roses, which limits direct transfer of the method. In rose-specific studies, Zhao et al. [34] developed a UAV-assisted monitoring framework by combining improved deep learning detection with image restoration techniques. Their results showed that image-quality enhancement can partially compensate for the loss of small-scale details in aerial imagery and improve the detection of tiny or distant rose targets. However, that study mainly focused on recovering image details through super-resolution restoration, and its detector was developed using a single UAV-derived rose dataset [34]. The effects of cross-source dataset integration and viewpoint diversity were not systematically examined. Moreover, the reliable detection of dense and very small rose buds remained dependent on the small-object [35–37] representation ability of the detector.

Taken together, RoseTracker and the UAV-assisted framework developed by Zhao et al. provide important foundations for automated rose monitoring [2,34]. However, existing studies have not fully addressed two related limitations. First, current rose datasets are commonly collected from limited acquisition systems, viewpoints, cultivars, or greenhouse environments, which restricts model generalization across different image sources. Second, wide-view UAV imagery contains many small, dense, and partially occluded rose buds with weak texture and boundary information. Lightweight detectors may lose these features during deep feature extraction and multi-scale fusion. Accordingly, this study combines multi-source dataset construction with task-oriented lightweight feature enhancement to improve dataset diversity and small-target detection while maintaining low model complexity.

1.3. Evolution of the YOLO Framework

Among deep-learning-based object detectors, the YOLO (You Only Look Once) family is one of the most influential frameworks because it offers a strong balance between speed, accuracy, and deployment efficiency [38–40]. Compared with two-stage detectors such as Faster R-CNN, which first generate candidate regions and then classify them, YOLO performs localization and classification in a unified single-stage pipeline [41,42]. This design

makes YOLO especially suitable for real-time or near-real-time agricultural applications, where large numbers of images must be processed efficiently.

The original YOLO model formulated object detection as a regression problem, which greatly simplified the detection pipeline and improved inference speed [38]. Later versions addressed its early limitations in localization accuracy and small-object detection. YOLOv2 improved the original framework through architectural refinements and anchor-based detection, while YOLOv3 further strengthened performance by introducing multi-scale prediction and a stronger backbone network [43–45]. These advances made YOLO more effective for scenes containing objects of different sizes, which is highly relevant to agricultural imagery.

Subsequent developments continued to improve the framework. YOLOv4 enhanced both efficiency and accuracy through optimized training strategies and feature aggregation design [46]. YOLOX introduced an anchor-free formulation, a decoupled head, and improved label assignment [47]. YOLOv6 and YOLOv7 further emphasized industrial deployment and real-time performance, while maintaining strong detection accuracy [48,49]. More recent versions such as YOLOv9 and YOLOv10 extended the framework with new gradient optimization strategies and end-to-end real-time design, further improving the balance between performance and latency [50,51]. These advances have also supported the extension of newer YOLO variants to task-specific applications, particularly in challenging small-object detection scenarios [52]. This trend has continued in even newer variants. YOLOv13 further extends the YOLO series by improving detection robustness and adaptability in complex scenes, providing additional potential for real-time vision tasks that involve small targets and background interference [53]. Recent application studies have also shown that YOLO-based frameworks remain highly effective for small-object detection in challenging environments. For example, enhanced YOLOv13 has been used for real-time detection of floating plastic debris under water-surface interference and complex aquatic backgrounds [54].

In agricultural computer vision, deep-learning-based detection methods have been widely applied to fruits, flowers, leaves, and other crop organs [12,55–57] because they offer strong feature representation ability, robust performance, and practical deployment potential. Among these methods, the YOLO family has become especially popular due to its favorable balance between accuracy and efficiency. This is particularly relevant to greenhouse flower monitoring, where detectors must handle tiny buds, dense target distributions, and complex backgrounds. Modern YOLO variants are advantageous in such settings because they provide stronger multi-scale representation, improved feature aggregation, and greater flexibility for incorporating task-specific modules [58,59]. Therefore, the YOLO family provides a strong and practical foundation for customized rose detection in greenhouse environments [60].

1.4. YOLOv12n: Architectural Improvements

Recent work has continued to push the YOLO family toward stronger feature modeling without sacrificing real-time efficiency. In this context, YOLOv12 was introduced as an attention-centric real-time detector, aiming to combine the speed advantage of the YOLO series with the stronger representation capability of attention-based design [61]. This shift is meaningful because modern detection tasks increasingly require the model to preserve fine-grained cues under cluttered backgrounds and weak object visibility.

Among the lightweight YOLOv12 variants, YOLOv12n is particularly suitable for practical deployment because it provides a favorable balance between computational efficiency and feature representation. This characteristic is important for greenhouse rose monitoring, where small buds must be detected under dense foliage, target overlap, and substantial appearance variation. Recent studies on small-object detection have emphasized the importance of multi-scale representation, efficient feature aggregation, and suppression of background interference [62–64]. YOLOv12 has recently been evaluated for UAV-based marigold flower detection, demonstrating its applicability to flower recognition [55]. However, that study used YOLOv12 mainly as a comparative detector and did not develop a task-specific YOLOv12n architecture for distinguishing rose buds from blooming flowers. These findings suggest that YOLOv12n is not only a compact real-time detector, but also a suitable foundation for task-oriented improvement in greenhouse rose detection.

However, directly applying the original YOLOv12n to greenhouse rose detection still faces several challenges:

- The lightweight backbone may miss subtle features of small and partially occluded rose buds.
- The original feature fusion strategy may be insufficient for complex multi-scale greenhouse backgrounds.
- The default detection head may remain sensitive to dense targets, overlap, and background interference.

Therefore, although YOLOv12n provides a suitable lightweight foundation, its direct application still requires task-oriented refinement for complex greenhouse rose detection.

1.5. Objectives and Contributions of This Study

This study aims to develop an accurate, lightweight, and practical framework for rose detection in UAV-derived vertical aerial imagery, with particular emphasis on distinguishing blooming roses from flower buds under complex cultivation conditions. To this end, a multi-source dataset construction strategy and an improved YOLOv12n-based detector are jointly introduced to enhance detection robustness against small targets, occlusion, dense canopies, and heterogeneous greenhouse backgrounds. An overview of the proposed research framework is presented in Figure 1.

The main contributions of this study are summarized as follows:

- A multi-source rose detection dataset, termed the *GreenHouse Rose dataset*, is constructed by integrating self-collected UAV rose images, screened RoseTracker overhead images, and supplementary multi-view rose images. This dataset enriches the diversity of rose targets in scale, growth stage, viewpoint, and background complexity, and provides a more representative benchmark for greenhouse rose detection.
- A lightweight improved detector, *BloomRoseNet*, is developed for complex greenhouse rose scenes based on YOLOv12n. The model improves the detection of dense, small, and visually similar rose targets, while maintaining a favorable balance between accuracy, efficiency, and deployment cost.
- Systematic validation of data effectiveness and model robustness. We perform comprehensive experiments, including ablation study, comparison with mainstream detectors, supplementary-data validation, and robustness analysis under blur, occlusion, and illumination disturbances. The results demonstrate that the proposed framework not only improves overall detection performance, but also provides stable and practical support for large-area greenhouse phenotyping and intelligent cultivation management.

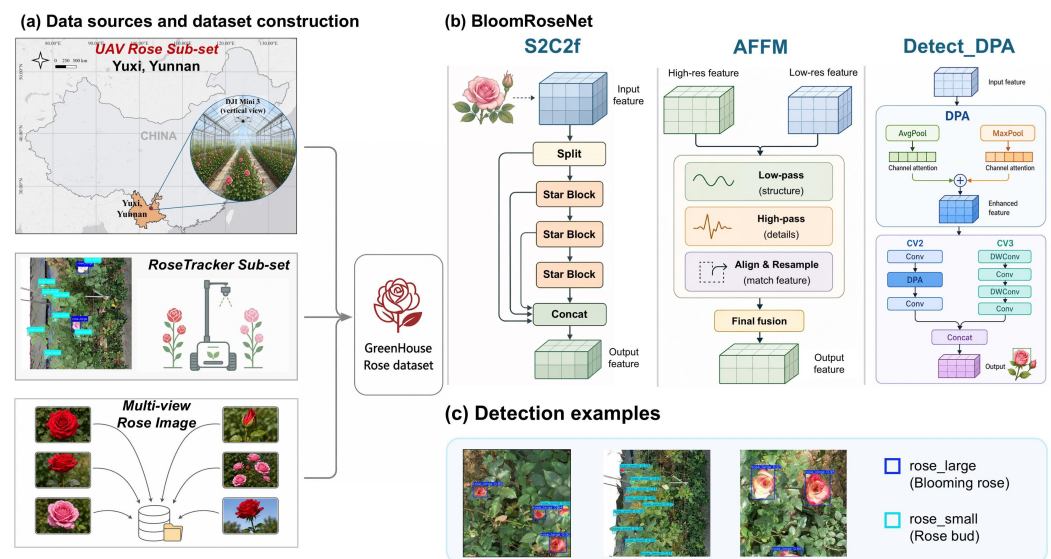


Figure 1. Overview of the GreenHouse Rose dataset construction and the proposed BloomRoseNet framework. (a) Multi-source dataset construction from UAV greenhouse images, the RoseTracker subset, and supplementary multi-view rose images. (b) Main improved modules of BloomRoseNet. (c) Representative detection results for rose_large and rose_small targets.

2. Materials and Methods

2.1. Dataset

To develop a robust rose detector for UAV-based greenhouse monitoring, a multi-source dataset construction strategy was adopted in this study. First, two overhead-view datasets, namely a self-collected UAV rose subset and the RoseTracker dataset, were combined to establish the primary overhead benchmark, referred to as the Overhead Rose dataset. Subsequently, additional multi-view rose images collected from several plant vision datasets were introduced as a supplementary source to enrich viewpoint and scene diversity. This use of heterogeneous visual sources is also consistent with recent multimodal learning studies that emphasize interaction-aware modeling of complementary information [65]. By integrating these three subsets, the final multi-source dataset, named the GreenHouse Rose dataset, was constructed.

2.1.1. Self-Collected UAV Rose Subset

The primary dataset used in this study was collected in a large-scale greenhouse rose cultivation area located in Yuxi City, Yunnan Province, China, which is a representative fresh-cut flower production region. The survey was conducted in a spacious greenhouse environment that allowed stable low-altitude UAV flight for overhead image acquisition. A consumer-grade DJI Mini 3 platform was used to capture the images. During data collection, the UAV was operated at an altitude of approximately 3 m with a camera field of view of about 90°. To ensure continuous coverage of the planting rows and sufficient overlap between adjacent images, the UAV followed a serpentine flight path throughout the greenhouse. The original aerial images were recorded at a resolution of 3840×2160 pixels.

The raw UAV images acquired during the survey were first screened to remove marginal scenes without rose distribution, as well as other images unsuitable for dataset construction due to environmental interference. The retained large overhead images were then cropped into image patches of 640×640 pixels, yielding 1134 image samples. These cropped samples were manually annotated and subsequently reviewed by experts to ensure the accuracy and consistency of the labels. All rose instances were manually annotated with bounding boxes using Labelme version 5.10.0. Two target categories were

defined in this dataset, namely '*rose_small*', representing buds, and '*rose_large*', representing blooming roses. These terms were used only as class labels rather than quantitative size categories. Specifically, *rose_small* represented unopened or visually closed flower buds with compact shapes and limited petal exposure, whereas *rose_large* represented blooming roses with expanded floral structures and clearly visible petals. The categories were therefore distinguished according to phenological stage and observable morphological and textural characteristics rather than bounding-box size.

The annotations were completed by one expert familiar with rose growth stages and subsequently reviewed by a second expert. Incorrect category labels, omitted targets, and inaccurate bounding boxes were corrected during the review. Ambiguous instances near the bud-to-bloom transition were jointly examined based on petal opening and overall floral morphology. This self-collected dataset served as the core component of the overhead rose detection benchmark established in this study.

2.1.2. RoseTracker Overhead Rose Dataset

To complement the self-collected UAV rose subset, the RoseTracker dataset was further incorporated as an additional overhead rose data source. To maintain consistency with our task setting, the target categories in RoseTracker were unified with those of the self-collected dataset, namely '*rose_small*' for buds and '*rose_large*' for blooming roses.

Considering that the original RoseTracker collection contains image samples acquired under different conditions, a screening process was conducted before integration. Only images with relatively similar illumination conditions, imaging characteristics, and shooting settings to our target greenhouse monitoring scenario were retained. After screening, 1038 images were selected from RoseTracker and used in this study. Compared with the self-collected UAV rose data, these images provide additional overhead rose samples from a different acquisition system and cultivation setup, thereby increasing inter-source diversity while preserving the consistency of the overhead-view detection task. The selected RoseTracker images were maintained at their original size of 640×640 pixels.

2.1.3. Construction of the Overhead Rose Dataset

After label unification and data screening, the self-collected UAV rose subset and the RoseTracker subset were merged to form the Overhead Rose dataset, which contained 2172 images in total. Since both subsets consist primarily of overhead-view rose imagery, the resulting dataset preserves the spatial and visual characteristics of UAV-style top-view monitoring while improving sample diversity across acquisition devices and cultivation conditions. From this dataset, to ensure an unbiased evaluation protocol, 250 overhead-view images were randomly selected and fixed as an independent test set throughout all experiments. The remaining 1922 images were used for training and validation under the overhead-only setting.

2.1.4. Supplementary Multi-View Rose Images and the Final GreenHouse Rose Dataset

Although the Overhead Rose dataset provides a dedicated benchmark for greenhouse rose detection from a top-view perspective, both dataset construction and preliminary experiments revealed an important limitation. In practical greenhouse scenes, due to the height of rose plants, the weight and inclination of flowers, and the lateral angular difference between the plant structure and the camera viewpoint, some buds are not always clearly visible from directly overhead observations. As a result, *rose_small* instances are more prone to inaccurate recognition and missed detection. This limitation is closely related to the fact that both the self-collected dataset and the RoseTracker subset mainly contain images captured from a single overhead viewpoint.

To address this issue, supplementary rose images were further screened from several public Kaggle datasets, including *Flowers Recognition* and *Flower Classification*. Although these datasets were originally developed for image classification, unlike the overhead datasets, these images cover multiple viewpoints and more diverse acquisition environments. Only images with clear rose targets, limited blur and occlusion, and reliable object boundaries were retained. Images dominated by a single close-up flower or excessively dense rose clusters were excluded. Priority was given to images in which rose instances were relatively evenly distributed and the number of visible targets was similar to that in the self-collected greenhouse dataset.

A total of 328 multi-view rose images were selected, manually annotated, and subsequently reviewed by experts to ensure label reliability. The image size was unified to 640×640 pixels through cropping and padding operations. These images were introduced as an auxiliary data source to expand the visual diversity of the training data. Such diversity cannot be effectively simulated by conventional data augmentation alone, because augmentation can modify geometric appearance or color distribution, but cannot truly introduce novel viewpoint-dependent structural features of rose targets.

By integrating the self-collected UAV rose subset, the screened RoseTracker subset, and the supplementary multi-view rose images, the final dataset, termed the GreenHouse Rose dataset, was established. It contained 2500 images in total, including 2172 overhead-view images and 328 supplementary multi-view images. To ensure a consistent and fair evaluation of the contribution of the supplementary multi-view data, the same 250 overhead-view images were fixed as the independent test set for all experiments. The remaining images were used to construct the training and validation sets, resulting in 2000 training images, 250 validation images, and 250 test images.

The GreenHouse Rose dataset therefore serves as the final benchmark in this study, enabling the proposed model to learn both the structured characteristics of overhead greenhouse scenes and the broader appearance variability of rose targets under different viewpoints and imaging environments. A statistical visualization of the dataset is presented in Figure 2. As shown in the figure, the annotated objects are distributed across two categories, namely *rose_large* and *rose_small*, with *rose_small* accounting for a larger proportion. The ratio between the two categories is approximately 1.4:1, indicating only a mild class imbalance. Therefore, no class-specific loss weighting or resampling strategy was introduced, and the standard YOLO loss functions were retained. In addition, most images contain only a few annotated targets, with an average of 3.0 objects per image and a median of 2. The object size distribution is substantially more imbalanced, with small objects dominating the dataset, which highlights the small-target detection characteristics and practical challenges of greenhouse rose monitoring.

2.2. Model Architecture

2.2.1. Proposed Method: BloomRoseNet

In this study, an improved YOLOv12n-based detector, termed BloomRoseNet, is proposed for rose detection in complex greenhouse environments.

YOLOv12 is adopted as the baseline framework in this study because it offers a favorable balance between detection accuracy and computational efficiency, which is highly desirable for practical greenhouse inspection and UAV-based flower monitoring. As an attention centric real time object detector, YOLOv12 retains the standard one stage detection paradigm composed of a backbone, a neck, and a detection head, while further strengthening feature modeling through attention-based architectural design. In particular, YOLOv12 introduces an attention-oriented framework that improves representation capability while

maintaining competitive inference speed, and its nano variant is especially suitable for lightweight deployment under resource constrained conditions.

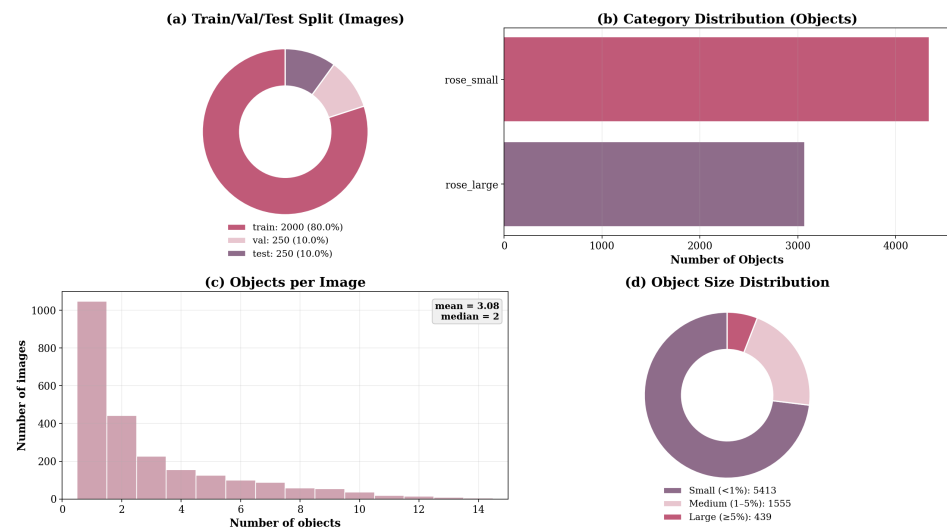


Figure 2. Statistical visualization of the GreenHouse Rose dataset: (a) train/validation/test split of images, (b) category distribution of annotated rose objects, (c) distribution of the number of objects per image, and (d) distribution of object sizes based on the ratio between bounding-box area and image area, indicating that the dataset is dominated by small objects.

Given an input image, the backbone of YOLOv12 performs hierarchical feature extraction through a sequence of convolutional and residual style modules, progressively generating multi-level feature maps with different spatial resolutions and semantic depths. Such a hierarchical representation is essential for object detection tasks involving scale variation, because shallow layers preserve high-resolution texture and edge information, whereas deeper layers encode stronger semantic abstraction. In addition, YOLOv12 integrates the A2C2f design with area attention-related mechanisms in its architecture, enabling more effective long-range dependency modeling than conventional purely convolution-based YOLO variants. This design improves the expressive power of the network while preserving the real time characteristics required for practical applications.

These properties make YOLOv12n an attractive starting point for rose detection in greenhouse environments. However, when it is directly applied to our task, several limitations become evident. First, the targets in this study, including fully opened roses and flower buds, often exhibit substantial scale variation, dense spatial distribution, and frequent mutual occlusion. Small buds are particularly vulnerable to being overwhelmed by surrounding leaves, branches, and overlapping flowers. Under such circumstances, repeated feature aggregation in the neck may still dilute fine-grained local cues, thereby reducing the recall of small rose targets. Second, greenhouse imagery usually contains complex background clutter, uneven illumination, and strong local appearance fluctuations caused by shadows, overlapping foliage, and reflective surfaces. Although YOLOv12 has enhanced representation ability, the original feature extraction and fusion pipeline still lacks sufficient task-oriented selectivity for distinguishing roses from visually similar background patterns. Third, accurate localization remains challenging because the boundaries of roses are often partially visible or ambiguous, especially in crowded scenes where petals and buds are intertwined with branches and leaves. In these cases, the original detection head may not focus adequately on the most informative target regions, leading to unstable classification and bounding box regression. These task-specific challenges indicate that a lightweight YOLO baseline, although efficient, still requires task-oriented

refinement for reliable greenhouse rose detection under small-object, cluttered-background, and occlusion conditions.

To address the above limitations, this study proposes an improved YOLOv12-based detector specifically tailored for rose detection in complex greenhouse scenes. As shown in Figure 3, the proposed model preserves the overall backbone, neck, and detection head pipeline of YOLOv12n, but introduces three task-oriented modifications at carefully selected locations. First, the original A2C2f modules are replaced by the proposed Star-Based Cross-Stage Partial C2f (S2C2f) modules, which incorporate Star_Block-based nonlinear feature transformation to enhance local fine-grained feature extraction and to better discriminate roses from cluttered backgrounds. Second, the conventional feature fusion operation in the neck is replaced by the adaptive frequency feature fusion module (AFFM) which strengthens intra-class consistency and boundary clarity through adaptive low-frequency and high-frequency refinement. Third, the original detection head is replaced by Detect Dual Pool Attention (Detect_DPA), a lightweight dual pool attention-based head that enhances channel response to partially occluded rose targets while reducing redundant parameters. Through these targeted improvements, the proposed network forms a rose-oriented detection framework that is better suited to the challenges of multi-scale targets, background interference, boundary ambiguity, and dense occlusion, while preserving the efficiency advantage of the YOLOv12n baseline.

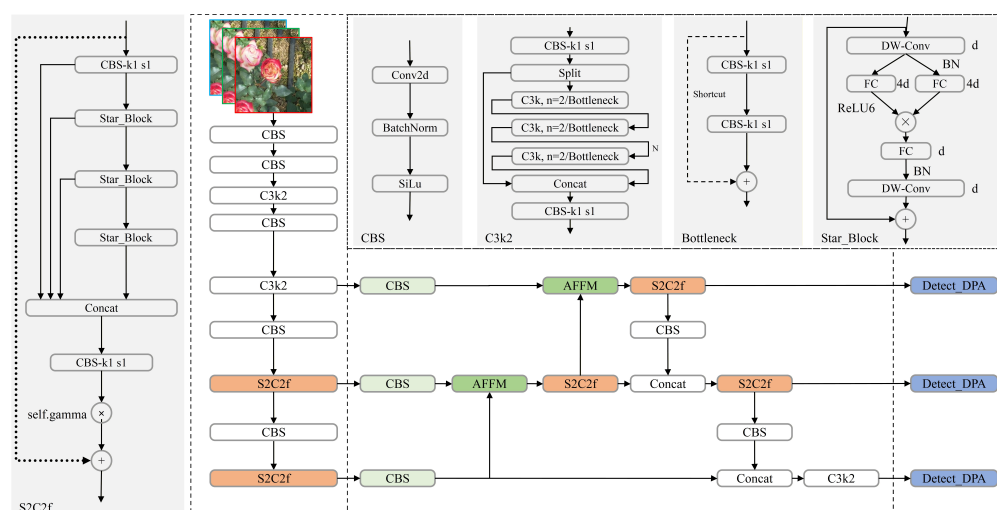


Figure 3. Overall framework of the proposed BloomRoseNet for greenhouse rose detection.

2.2.2. S2C2f, a Star-Based Cross-Stage Partial Feature Extraction Module

In complex greenhouse scenes, rose targets, including fully opened roses and flower buds, are frequently affected by severe occlusion, uneven illumination, and interference from surrounding branches and leaves. Although the area attention mechanism in the original A2C2f module of YOLOv12 [61] is effective for modeling long-range dependencies, its task-specific benefit may be limited in scenes dominated by small and densely occluded rose targets. Broad spatial aggregation may introduce responses from surrounding leaves and branches, thereby weakening the local texture and boundary cues required for distinguishing small buds. In addition, standard convolution operations may not sufficiently capture subtle local patterns and fine-grained texture differences, causing critical discriminative information to be weakened during deep feature extraction. Inspired by StarOperation [66], this study therefore proposes a star-based cross-stage partial C2f module, termed S2C2f. Through element-wise multiplicative interaction, the star operation enables efficient high-order combinations of local features and implicitly projects them into a high-dimensional nonlinear space, thereby enhancing fine-grained feature representation

without introducing substantial computational burden. As shown by the heatmap comparison in Figure 7, S2C2f produces more concentrated responses around true rose regions and reduces activation in confusing background areas. The architecture of the proposed S2C2f module is shown in Figure 4.

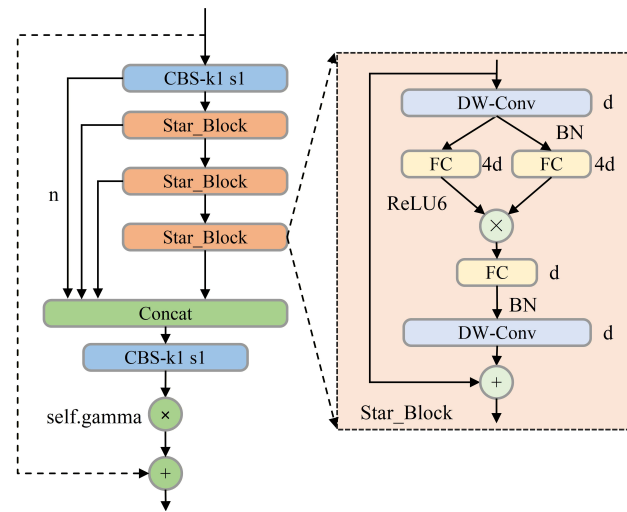


Figure 4. The structure of improvement-S2C2f.

The Star_Block serves as the core feature extraction unit of S2C2f. Its key characteristic lies in the star operation, in which element-wise multiplication enables the input features to be mapped into an implicit high-dimensional nonlinear space with an approximate dimensionality of $\left(\frac{d}{\sqrt{2}}\right)^2$, in a manner analogous to the principle of polynomial kernel functions.

First, the input feature map $X \in \mathbb{R}^{C \times H \times W}$ is passed through a depthwise convolution layer to extract local spatial features:

$$X' = \text{DWConv}_{7 \times 7}(X), \quad (1)$$

where the kernel size of the depthwise convolution is 7×7 , and the group number is equal to the number of input channels C . This operation effectively captures local structural information, such as rose edges and texture patterns, without introducing excessive computational complexity.

Second, the preprocessed feature X' is fed into two independent 1×1 convolution branches, which expand the channel dimension to α times the original size, where α denotes the expansion ratio and is set to 3 in this study:

$$X_1 = f_1(X') = W_1^T X' + b_1, \quad (2a)$$

$$X_2 = f_2(X') = W_2^T X' + b_2 \quad (2b)$$

where $W_1, W_2 \in \mathbb{R}^{C \times \alpha C}$ denote learnable weight matrices, and b_1 and b_2 are bias terms. These two branches learn complementary feature transformations and provide diverse representations for the subsequent star operation.

Third, the outputs of the two branches are fused by element-wise multiplication, followed by nonlinear activation:

$$X_{\text{star}} = \text{ReLU}(X_1) \odot X_2, \quad (3)$$

where $\odot$ denotes the Hadamard product. According to the theoretical analysis in [66], this operation can be expanded as:

$$X_{\text{star}} = \left(\sum_{i=1}^{d+1} W_1^i X^i \right) \odot \left(\sum_{j=1}^{d+1} W_2^j X^j \right) = \sum_{i=1}^{d+1} \sum_{j=1}^{d+1} \alpha_{(i,j)} X^i X^j, \quad (4)$$

where d denotes the number of input channels and $\alpha_{(i,j)} = W_1^i W_2^j + W_1^j W_2^i (i \neq j)$. This formulation indicates that the star operation requires only relatively low computational complexity ($O(d)$), while implicitly constructing a high-dimensional feature space containing $\frac{(d+2)(d+1)}{2} \approx \left(\frac{d}{\sqrt{2}}\right)^2$ linearly independent terms. This property is highly analogous to polynomial kernels. For rose detection, features such as petals of different sizes and branches with different orientations can therefore be more effectively distinguished through such implicit nonlinear mapping, which helps alleviate feature confusion between large roses and small buds under leaf and branch occlusion.

Finally, the features produced by the star operation are compressed back to the original channel dimension through a 1×1 convolution, followed by a second depthwise convolution for spatial refinement:

$$X_{\text{out}} = \text{DWConv}_{27 \times 7}(g(X_{\text{star}})), \quad (5)$$

where $g(\cdot)$ denotes the channel compression convolution. A residual connection is then introduced:

$$Y = X + X_{\text{out}}. \quad (6)$$

This residual design stabilizes gradient propagation and allows multiple Star_Blocks to be stacked without causing performance degradation.

Based on the above design, the proposed S2C2f removes the redundant ABlock structure in the original A2C2f, including the area attention and multilayer perceptron components, and replaces the original ABlock with the proposed Star_Block. In this way, S2C2f forms a more efficient feature extraction module. In the final network, all A2C2f modules in YOLOv12 are replaced by S2C2f modules.

2.2.3. AFFM, an Adaptive Frequency Feature Fusion Module

Conventional feature fusion strategies, such as direct concatenation after upsampling, often suffer from intra-class inconsistency and blurred boundaries, which adversely affect the detection accuracy of both large roses and small buds. Inspired by the Adaptive Spectral Detail Refinement strategy reported in [67], this study proposes an adaptive frequency feature fusion module, AFFM, which enhances feature consistency and boundary sharpness through frequency domain filtering and offset-based resampling. The structure of AFFM is shown in Figure 5.

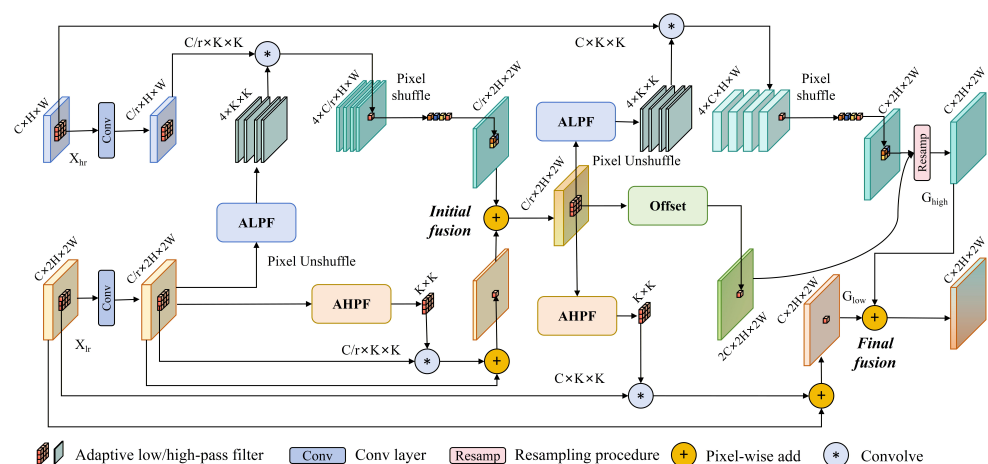


Figure 5. The structure of improvement-AFFM.

From the perspective of frequency decomposition, AFFM addresses two major issues in feature fusion. Intra-class inconsistency can be interpreted as high-frequency disturbance within feature representations, whereas boundary displacement is often associated with the loss of high-frequency details. To address these issues, AFFM is composed of three key components, namely an adaptive low-pass filter generator, an adaptive high-pass filter generator, and an offset generator. In addition, a frequency adaptive gating mechanism is introduced to dynamically regulate the contributions of the low-frequency and high-frequency branches according to the characteristics of features at different scales.

Given a high-resolution feature map, X_{hr} , and a low-resolution feature map, X_{lr} , AFFM first performs channel compression and preliminary fusion, after which the fused features are processed in parallel by three branches.

In the adaptive low-pass filtering branch, the adaptive low-pass filter generator predicts spatially varying low-pass kernels to filter and upsample the low-resolution features:

$$\tilde{X}_{lr} = \text{PixelShuffle} \left\{ \sum_{p,q \in \Omega_{LP}} W_{LP}^{p,q} \cdot X_{lr}(i+p, j+q) \right\}, \quad (7)$$

where the filter kernel W_{LP} is normalized by Softmax to ensure that its weights are non negative and sum to 1. This operation suppresses high-frequency disturbances caused by petal and pistil texture variations inside rose regions, thereby improving intra-class feature consistency.

In the adaptive high-pass filtering branch, the adaptive high-pass filter generator constructs high-pass kernels to recover and enhance missing high-frequency boundary information:

$$\tilde{X}_{hr} = X_{hr} + \left(X_{hr} - \sum_{p,q \in \Omega_{HP}} W_{HP}^{p,q} \cdot X_{hr}(i+p, j+q) \right). \quad (8)$$

This branch effectively sharpens petal edges and sepal contours, making the boundaries of roses with different sizes clearer and more distinguishable.

In the offset resampling branch, the offset generator first computes the local cosine similarity map S , which captures the feature correlation between each spatial location and its neighboring regions:

$$S_{i,j}^{p,q} = \frac{Z_{i,j} \cdot Z_{i+p,j+q}}{\|Z_{i,j}\|_2 \cdot \|Z_{i+p,j+q}\|_2}. \quad (9)$$

Based on the similarity map, the branch then predicts the offset direction and magnitude, which is $O = D \cdot A + O_{init}$, and uses grid sampling to replace low-similarity features with more coherent high-similarity features. This process further corrects large inconsistent regions and thin ambiguous boundaries.

Moreover, a frequency adaptive gating mechanism is introduced to dynamically adjust the contributions of low-frequency and high-frequency branches according to the global context of the high-resolution features:

$$G_{low}, G_{high} = \text{Sigmoid}(\phi(\text{GAP}(X_{hr}))), \quad (10a)$$

$$Y = \alpha_{low} G_{low} \odot \tilde{X}_{lr} + \alpha_{high} G_{high} \odot \tilde{X}_{hr} + \tilde{X}_{lr}^{align}, \quad (10b)$$

where $\phi(\cdot)$ denotes the gating subnetwork, and α_{low} and α_{high} denote the learnable frequency base weights. This design enables the network to adaptively regulate its frequency response across different feature scales. In particular, shallow features tend to emphasize high-frequency boundary details, whereas deep features focus more on low-frequency semantic consistency.

In the neck of YOLOv12, AFFM is introduced at the P3 and P4 levels, corresponding to Layers 12 and 14, to replace the original concatenation operations. P3 and P4 retain relatively high spatial resolution and rich high-frequency boundary information, and they mainly support the detection of small- and medium-scale rose targets, which account for the majority of instances in this study. Therefore, adaptive low- and high-frequency refinement at these levels can improve feature consistency, preserve petal and bud boundaries, and reduce misalignment during multi-scale fusion.

AFFM was not extended to P5 because P5 has the lowest spatial resolution and primarily contains high-level semantic and low-frequency information. At this stage, much of the fine boundary and texture information required by the high-frequency branch has already been attenuated. Applying AFFM at P5 would therefore provide limited additional benefit while introducing unnecessary resampling and computational operations. Accordingly, AFFM was deployed only at P3 and P4 to match the dominant target scales and frequency characteristics of greenhouse rose features.

2.2.4. Detect_DPA, a Dual Pool Attention-Based Detection Head

The original detection head of YOLOv12 adopts standard convolution stacking for feature extraction, which is often insufficient for focusing on the most informative target regions in densely occluded scenes. To overcome this limitation, inspired by the dual pool attention mechanism in DPA [68], this study proposes a lightweight detection head termed Detect_DPA. By introducing a parallel dual pooling attention mechanism, Detect_DPA enhances channel feature representation while reducing parameter redundancy. The structure of Detect_DPA is shown in Figure 6.

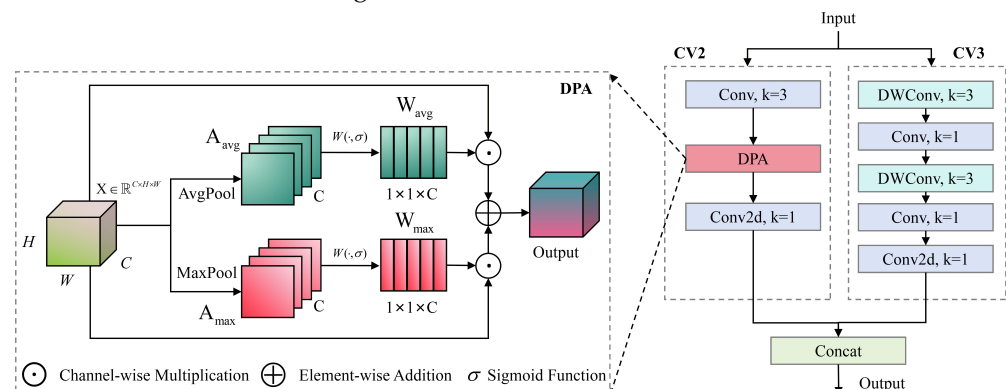


Figure 6. The structure of improvement—Detect_DPA.

Detect_DPA replaces the redundant convolution layers in the original detection head with a lightweight dual pool attention module. In the original detection head, the CV2 branch contains two consecutive standard convolution layers, resulting in relatively large parameter consumption and feature redundancy. In the proposed design, this branch is simplified into a sequential structure composed of Conv and DPA, which preserves feature extraction capability while significantly reducing the parameter count. The DPA module captures global statistical information and local salient responses through two parallel pooling branches, followed by channel-wise recalibration for feature enhancement.

Given an input feature map $X \in \mathbb{R}^{C \times H \times W}$, the DPA module first extracts channel descriptors through two parallel branches:

$$A_{\text{avg}} = \text{AvgPool}(X) = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W X_{i,j}, \quad (11a)$$

$$A_{\text{max}} = \text{MaxPool}(X) = \max_{i,j} X_{i,j}, \quad (11b)$$

where AvgPool denotes global average pooling, which captures global statistical information along the channel dimension, and MaxPool denotes global maximum pooling, which extracts the most salient response from each channel. The two descriptors are then passed through fully connected layers with shared parameters for channel interaction:

$$W_{\text{avg}} = \sigma(W_2 \cdot \delta(W_1 \cdot A_{\text{avg}})), \quad (12a)$$

$$W_{\text{max}} = \sigma(W_2 \cdot \delta(W_1 \cdot A_{\text{max}})), \quad (12b)$$

where $W_1 \in \mathbb{R}^{\frac{C}{r} \times C}$ and $W_2 \in \mathbb{R}^{\frac{C}{r} \times C}$ are the weight matrices of the fully connected layers, $r = 16$ is the reduction ratio, $\delta(\cdot)$ denotes the ReLU activation function, and $\sigma(\cdot)$ denotes the Sigmoid activation function. The two attention weights are then fused as:

$$Y = (X \odot W_{\text{avg}}) \oplus (X \odot W_{\text{max}}) \quad (13)$$

where $\oplus$ represents feature fusion. By exploiting the complementary characteristics of average pooling and maximum pooling, the model can simultaneously focus on the global distribution of features and the local peak responses, thereby strengthening the channel representation of rose targets.

In the improved YOLOv12 architecture, Detect_DPA replaces the original Detect module as the detection head. Specifically, in the CV2 branch, which is responsible for bounding box regression, the original two consecutive convolution layers are replaced by a Conv and DPA sequence, whereas the CV3 branch, which is used for classification, retains the original depthwise convolution structure. Through this modification, Detect_DPA reduces the parameter count relative to the original detection head while effectively enhancing the response to occluded rose targets by means of dual pooling attention, thereby improving both the detection accuracy and localization precision for large roses and small buds.

2.2.5. Training Setup and Evaluation Metrics

The experiments were conducted on a workstation equipped with Ubuntu 22.04 as the operating system, 120 GB of system memory, an Intel(R) Xeon(R) Platinum 8352V CPU, and an NVIDIA GeForce RTX 4090 GPU with 24 GB of video memory. The deep learning framework used in this study was PyTorch 2.1.2, with CUDA 11.8 and Python 3.10.8. The detailed experimental environment is summarized in Table 1.

Table 1. Experimental environment settings.

Parameter	Value
Operating system	Ubuntu 22.04
Memory	120 GB
GPU	GeForce RTX 4090
GPU memory	24 GB
CPU	Intel(R) Xeon(R) Platinum 8352V
Deep learning framework	PyTorch 2.1.2
CUDA version	11.8
Python version	3.10.8

During training, all input images were resized to 640×640 pixels. The models were trained for 200 epochs with a batch size of 32. Stochastic Gradient Descent, SGD, was adopted as the optimizer, with an initial learning rate of 0.001, a momentum of 0.937, and a weight decay of 0.0005. In addition, the number of data loading workers was set to 25. The detailed training hyperparameters are listed in Table 2.

Table 2. Training hyperparameter settings.

Parameter	Value
Input image size	640 × 640
Epochs	200
Batch size	32
Optimizer	SGD
Initial learning rate	0.001
Momentum	0.937
Weight decay	0.0005
Workers	25

To comprehensively evaluate the performance of the proposed model, both detection accuracy and model efficiency were considered. For detection performance, standard object detection metrics, including Precision, Recall, Average Precision, AP, and Mean Average Precision, mAP, were adopted. For computational efficiency and lightweight deployment capability, the number of parameters, model size, and inference speed were further reported, measured by Params (M), Size (MB), and FPS, respectively.

For each predicted bounding box, the detection outcome was categorized as a True Positive, TP, False Positive, FP, or False Negative, FN, according to its matching relationship with the ground truth. A prediction was regarded as a TP when it correctly identified a rose target with the correct category and sufficient overlap with the corresponding ground truth box. An FP referred to an incorrect prediction, such as background regions or non target structures being mistakenly detected as roses, whereas an FN denoted a missed target. Based on these quantities, Precision and Recall were calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP}, \quad (14a)$$

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (14b)$$

Precision measures the reliability of the predicted results, whereas Recall reflects the completeness of target detection.

The localization quality of each prediction was measured by Intersection over Union, IoU, which quantifies the overlap between the predicted bounding box and its corresponding ground truth box. Based on the Precision–Recall curve, the Average Precision for the i th category was computed as:

$$AP_i = \text{area}(PR_i), \quad (15)$$

where $\text{area}(PR_i)$ denotes the area under the Precision–Recall curve of the i th category. The Mean Average Precision was then obtained by averaging AP over all N categories:

$$\text{mAP} = \frac{1}{N} \sum_{i=1}^N AP_i. \quad (16)$$

In this study, both mAP@50 and mAP@50:95 were reported. Specifically, mAP@50 evaluates detection performance at an IoU threshold of 0.50, whereas mAP@50:95 averages the AP values over IoU thresholds ranging from 0.50 to 0.95 with an interval of 0.05. Compared with mAP@50, mAP@50:95 provides a more stringent assessment of localization accuracy and therefore offers a more comprehensive evaluation of overall detection performance.

In addition to detection accuracy, model complexity and deployment efficiency were also taken into account. Params (M) denotes the total number of learnable parameters in the model, reflecting its structural complexity. Size (MB) represents the storage size of the

trained model, which is directly related to memory requirements during deployment. FPS, frames per second, measures the inference speed of the model and indicates its real time processing capability. These three indicators were used to assess whether the proposed model achieves a favorable balance between accuracy and lightweight design.

To reduce the influence of stochastic model training and to quantify the uncertainty associated with the limited test-set size, all compared detectors were independently trained three times using random seeds 0, 42, and 3407. The training, validation, and test partitions were fixed across all runs, and the random seeds controlled model initialization, data shuffling, and stochastic data augmentation. Detection metrics are therefore reported as the mean $\pm$ standard deviation over the three independent runs. Model parameters, model size, and inference speed were reported once because they were unaffected by the random seed.

Statistical significance was evaluated using a paired image-level bootstrap procedure. In each of 5000 bootstrap iterations, test images were sampled with replacement, and the same sampled image indices were used for all compared models. Predictions and ground-truth annotations from the sampled images were aggregated to recalculate mAP@50. For each bootstrap sample, the metric was first averaged across the three random-seed runs, after which the paired difference between BloomRoseNet and each baseline detector was calculated. The percentile method was used to obtain the 95% confidence interval of each performance difference. Two-sided bootstrap p values were calculated and adjusted for the 14 pairwise comparisons using the Holm–Bonferroni procedure. An adjusted value of $p_{\text{adj}} < 0.05$ was considered statistically significant.

For the density-wise analysis, bootstrap resampling was independently performed within each image-density group. For the scale-wise analysis, images were resampled before retaining ground-truth instances belonging to the selected scale group, thereby preserving the dependence among multiple targets originating from the same image.

3. Experimental Results

This section presents the experimental results from five aspects: ablation analysis, comparison with mainstream methods, evaluation of supplementary multi-view data, robustness analysis, and scene-wise performance analysis under different density and scale conditions. First, the contribution of each proposed module is examined through ablation experiments. Then, the overall performance of the final model is compared with several representative object detection methods. Furthermore, the effectiveness of the supplementary data is validated on Overhead Rose dataset. Finally, robustness experiments are conducted to assess the model stability under challenging conditions.

3.1. Ablation Study

To evaluate the effectiveness of the proposed modules, an ablation study was conducted on our test set by progressively introducing S2C2f, AFFM, and Detect_DPA into the baseline YOLOv12n. Specifically, M0 represents the original YOLOv12n baseline, M1 represents the model with S2C2f only, M12 represents the model with both S2C2f and AFFM, and M123 represents the final BloomRoseNet integrating all three proposed modules. The quantitative results are presented in Table 3. Overall, the progressive integration of the three modules consistently improved the detection performance. Compared with the baseline YOLOv12n, the final model achieved gains of 3.2 percentage points in precision, 3.3 percentage points in recall, 3.6 percentage points in mAP@50, and 1.6 percentage points in mAP@50:95. Meanwhile, the parameter count, model size, and inference speed remained at a lightweight and real-time level, indicating that the proposed improvements enhanced detection performance without introducing excessive computational burden. Qualitative

detection and heatmap visualizations further support the quantitative ablation results, showing that the progressive integration of the proposed modules enabled the model to focus more accurately on rose targets under both cluttered backgrounds and dense small-object scenes (Figure 7).

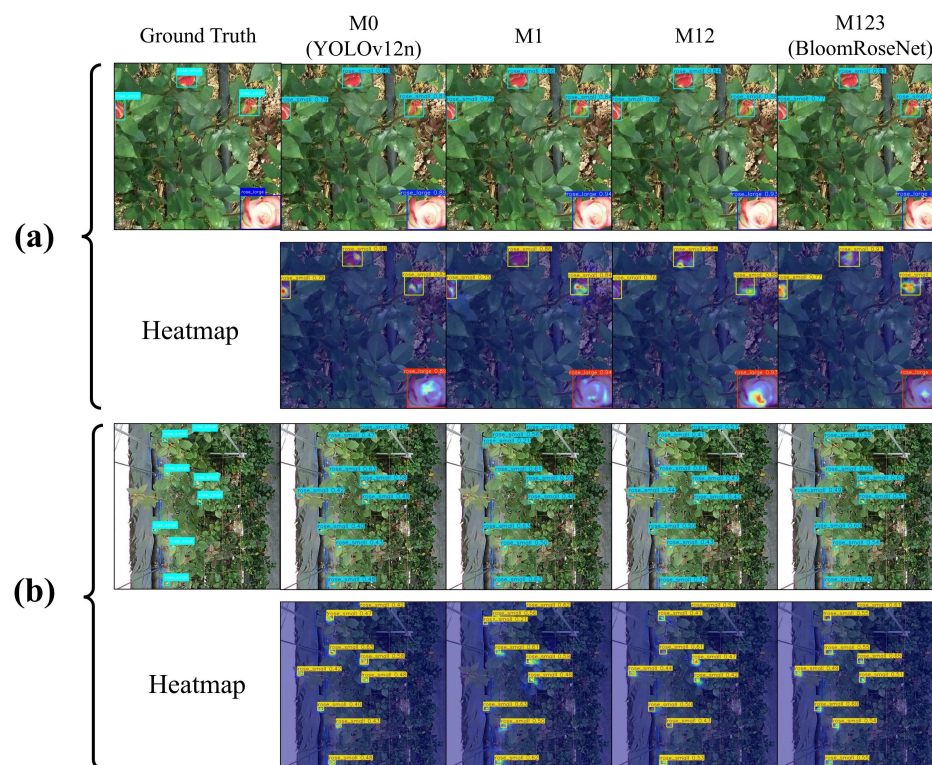


Figure 7. (a,b) Qualitative comparison of the ablation models on representative greenhouse rose images. For each case, the first row shows the detection results and the second row shows the corresponding heatmaps. Compared with the baseline, the progressively improved models exhibit more accurate localization and more concentrated responses on true rose regions, particularly in cluttered backgrounds and dense small-target scenes.

Table 3. Ablation study results of the proposed modules on test set.

Model	S2C2f	AFFM	Detect DPA	P (%)	R (%)	mAP@50 (%)	mAP@50:95 (%)	Params (M)	Size (MB)	FPS
M0	–	–	–	82.9	82.0	86.9	52.5	2.55	5.5	131
M1	✓	–	–	82.3	85.6	88.2	53.1	2.26	4.8	187
M12	✓	✓	–	81.9	86.6	89.0	53.5	2.18	4.7	147
M123	✓	✓	✓	86.1	85.3	90.5	54.1	2.08	4.5	134

Note: “✓” indicates that the corresponding module is included, while “–” indicates that the module is not used.

3.1.1. Contribution of S2C2f

Among the three proposed modules, S2C2f produced the most direct and comprehensive improvement in the early ablation stage. After introducing S2C2f into the baseline YOLOv12n, the recall increased markedly from 82.0% to 85.6%, representing the largest recall gain among the intermediate ablation steps. Meanwhile, mAP@50 improved from 86.9% to 88.2%, and mAP@50:95 increased from 52.5% to 53.1%. Although the precision decreased slightly from 82.9% to 82.3%, the overall detection performance showed a clear positive trend.

More importantly, the introduction of S2C2f also brought notable improvements in model efficiency. The number of parameters decreased from 2.55 M to 2.26 M, the model size was reduced from 5.5 MB to 4.8 MB, and the inference speed increased substantially

from 131 FPS to 187 FPS. This means that S2C2f not only enhanced the model's ability to detect rose targets, but also improved its computational efficiency at the same time.

From the perspective of the full ablation process, S2C2f can be regarded as the module that first established the performance advantage of the improved framework. Its introduction produced simultaneous gains in recall, mAP@50, lightweight design, and inference speed, which makes it the most prominent single-step improvement in the ablation study. Therefore, the results demonstrate that S2C2f played a foundational role in improving the overall detector performance.

3.1.2. Contribution of AFFM

On the basis of M1 model, the further introduction of AFFM led to a continued improvement in the overall detection results. Specifically, recall increased from 85.6% to 86.6%, which was the highest recall value among all ablation variants. At the same time, mAP@50 increased from 88.2% to 89.0%, and mAP@50:95 improved from 53.1% to 53.5%. Although precision showed a slight decrease from 82.3% to 81.9%, the overall trend remained positive in terms of comprehensive detection performance.

In addition, AFFM further reduced the model complexity. After introducing this module, the parameter count decreased from 2.26 M to 2.18 M, and the model size was reduced from 4.8 MB to 4.7 MB. Although the inference speed dropped from 187 FPS to 147 FPS compared with M1 model, it still remained within the real-time range. Compared with the baseline YOLOv12n, the model with S2C2f and AFFM achieved better recall, higher mAP values, and lower complexity, showing that AFFM further strengthened the improved detector on top of the gain already brought by S2C2f.

Overall, the contribution of AFFM is reflected more in the continued optimization of the intermediate feature representation, which helped the model maintain an upward trend in detection accuracy after the substantial gain obtained by S2C2f.

3.1.3. Contribution of Detect_DPA

After further incorporating Detect_DPA, the model achieved the best overall performance in the ablation study. The most notable improvement at this stage was the substantial increase in precision, which rose from 81.9% to 86.1%. Meanwhile, mAP@50 further improved from 89.0% to 90.5%, and mAP@50:95 increased from 53.5% to 54.1%, both reaching the highest values among all variants. Although recall decreased slightly from 86.6% to 85.3%, it still remained higher than the baseline result of 82.0%.

In terms of model complexity, Detect_DPA also further reduced the parameter count from 2.18 M to 2.08 M and the model size from 4.7 MB to 4.5 MB. The inference speed was 134 FPS, which was slightly lower than that of M1 and M12, but still sufficient for real-time deployment. Compared with the baseline YOLOv12n, the final model achieved gains of 3.2 percentage points in precision, 3.3 percentage points in recall, 3.6 percentage points in mAP@50, and 1.6 percentage points in mAP@50:95, while also becoming more lightweight.

These results indicate that Detect_DPA was the key module for improving the final detection quality, especially in terms of precision and overall localization performance. In other words, while S2C2f provided the most obvious early-stage gain and AFFM further enhanced the intermediate performance, Detect_DPA played the decisive role in pushing the final model to its best comprehensive result.

In summary, the ablation study verifies that the three proposed modules contributed positively to the overall performance of the final detector, although some individual metrics fluctuated during the progressive integration. S2C2f provided the most prominent and comprehensive single-step improvement, especially in recall, efficiency, and lightweight design. AFFM further improved the overall detection results and helped the model reach

the highest recall. Detect_DPA finally delivered the best precision and the highest mAP@50 and mAP@50:95 values. The combination of these three modules enabled the proposed model to achieve the best balance between detection accuracy, model complexity, and inference speed.

3.2. Comparison with Mainstream Methods

To further evaluate the superiority of the proposed model, comparative experiments were conducted on the test set against several mainstream object detection methods, including Faster R-CNN, SSD, a GhostNet-based detector, RT-DETR-L, Mamba-YOLO-T, YOLO-Worldv2, YOLOv8n, YOLOv9t, YOLOv10n, YOLOv11n, YOLOv12n, YOLOv13n, YOLO26n, and Rose-Mamba-YOLO. The quantitative results are summarized in Table 4.

Table 4. Comparison of the proposed method with mainstream object detection algorithms on the fixed test set. Detection metrics are reported as mean $\pm$ standard deviation over three independent runs.

Model	P (%)	R (%)	mAP@50 (%)	mAP@50:95 (%)	Params (M)	Size (MB)	FPS
Faster R-CNN	65.2 $\pm$ 0.3	68.8 $\pm$ 0.2	73.3 $\pm$ 0.3	27.7 $\pm$ 0.2	28.28	108.2	50
SSD	79.9 $\pm$ 0.3	80.4 $\pm$ 0.2	81.7 $\pm$ 0.3	46.3 $\pm$ 0.4	3.67	14.8	112
GhostNet	76.9 $\pm$ 0.4	80.9 $\pm$ 0.2	83.1 $\pm$ 0.2	50.3 $\pm$ 0.3	1.71	3.8	165
RT-DETR-L	83.8 $\pm$ 0.2	82.7 $\pm$ 0.5	87.8 $\pm$ 0.4	52.9 $\pm$ 0.4	25.67	53.6	58
Mamba-YOLO-T	84.6 $\pm$ 0.2	83.2 $\pm$ 0.5	88.1 $\pm$ 0.2	53.2 $\pm$ 0.3	5.99	12.3	121
YOLO-Worldv2	82.2 $\pm$ 0.4	83.4 $\pm$ 0.1	85.6 $\pm$ 0.3	51.9 $\pm$ 0.1	3.53	7.4	175
YOLOv8n	81.6 $\pm$ 0.2	81.6 $\pm$ 0.3	86.4 $\pm$ 0.3	52.3 $\pm$ 0.1	3.00	6.3	223
YOLOv9t	80.2 $\pm$ 0.4	81.4 $\pm$ 0.1	84.2 $\pm$ 0.3	51.7 $\pm$ 0.4	1.97	4.7	99
YOLOv10n	81.8 $\pm$ 0.6	81.7 $\pm$ 0.4	86.7 $\pm$ 0.5	52.1 $\pm$ 0.4	2.26	4.9	202
YOLOv11n	81.2 $\pm$ 0.3	81.1 $\pm$ 0.3	85.4 $\pm$ 0.2	51.5 $\pm$ 0.3	2.58	5.5	184
YOLOv12n	82.9 $\pm$ 0.1	82.0 $\pm$ 0.2	86.9 $\pm$ 0.1	52.5 $\pm$ 0.2	2.55	5.5	131
YOLOv13n	81.4 $\pm$ 0.3	79.7 $\pm$ 0.1	84.2 $\pm$ 0.1	51.0 $\pm$ 0.3	2.46	5.4	89
YOLO26n	82.5 $\pm$ 0.5	79.5 $\pm$ 0.2	84.0 $\pm$ 0.1	50.4 $\pm$ 0.3	2.37	5.1	144
Rose-Mamba-YOLO	85.2 $\pm$ 0.2	82.4 $\pm$ 0.4	89.0 $\pm$ 0.1	53.2 $\pm$ 0.3	8.28	16.6	118
BloomRoseNet	86.1 $\pm$ 0.1	85.3 $\pm$ 0.2	90.5 $\pm$ 0.2	54.1 $\pm$ 0.1	2.08	4.5	134

Note: Bold values indicate the best performance in each column.

Across the three independent runs, BloomRoseNet achieved mean precision, recall, mAP@50, and mAP@50:95 values of (86.1 $\pm$ 0.1)%, (85.3 $\pm$ 0.2)%, (90.5 $\pm$ 0.2)%, and (54.1 $\pm$ 0.1)%, respectively. The consistently small standard deviations indicate that the performance of the proposed model remained stable across different random initializations and was not dependent on a particular training run.

The paired image-level bootstrap analysis further showed that BloomRoseNet significantly outperformed all 14 baseline detectors in terms of mAP@50 after Holm–Bonferroni correction. Compared with Rose-Mamba-YOLO, which was the strongest baseline, BloomRoseNet improved the mean mAP@50 by 1.5 percentage points, with a 95% confidence interval of [0.2, 2.8] percentage points and an adjusted value of $p_{\text{adj}} = 0.041$. Compared with Mamba-YOLO-T, RT-DETR-L, and the baseline YOLOv12n, the corresponding improvements were 2.4, 2.7, and 3.6 percentage points, respectively. These results indicate that the performance advantage of BloomRoseNet exceeded the variability associated with random training initialization and finite test-set sampling.

Compared with the baseline YOLOv12n, BloomRoseNet also improved mAP@50:95 by 1.6 percentage points while reducing the parameter count from 2.55 M to 2.08 M and the model size from 5.5 MB to 4.5 MB. Compared with other lightweight YOLO variants, including YOLOv8n, YOLOv10n, and YOLOv11n, the proposed model consistently achieved higher mean precision, recall, and overall detection accuracy. Qualitative inference results further support these quantitative findings, showing that BloomRoseNet produced more

complete and accurate detections in complex greenhouse scenes containing dense foliage and small rose targets (Figure 8).

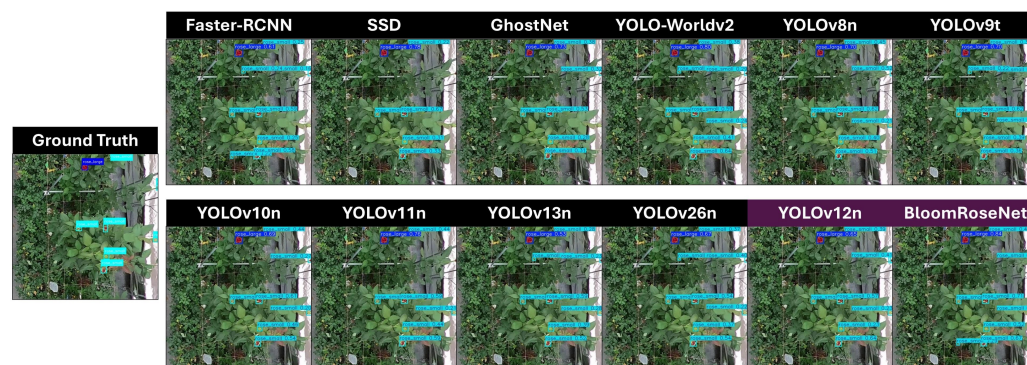


Figure 8. Qualitative comparison of inference results of different object detection methods on a representative greenhouse rose image.

Notably, two recent competitive architectures, RT-DETR-L and Mamba-YOLO-T, were also included for comparison. Although these models achieved relatively strong detection accuracy, neither surpassed BloomRoseNet on the test set. RT-DETR-L in particular relied on a much larger model scale, with 25.67 M parameters and a model size of 53.6 MB, while Mamba-YOLO-T also required clearly higher complexity than the proposed model. This result suggests that BloomRoseNet provides a more favorable trade-off between accuracy and model compactness for greenhouse rose detection.

In addition to the accuracy gains, the proposed model maintained a relatively compact architecture. Its parameter count and model size were 2.08 M and 4.5 MB, respectively, which were lower than those of most compared mainstream detectors while remaining close to the lightweight end among the YOLO-based models. Although its inference speed (134 FPS) was not the highest among all methods, it still satisfied real-time detection requirements. These results indicate that BloomRoseNet achieved a favorable balance between detection accuracy, model complexity, and inference efficiency. As further illustrated in Figure 9, BloomRoseNet also maintained stable inference efficiency across different hardware platforms, confirming its practical applicability for real-time greenhouse rose detection. To provide a more direct assessment of deployment latency, the average inference time per image was also tested under the same inference settings. The BloomRoseNet required approximately 7.4 ms per image on the RTX 4090, 18.9 ms on the RTX 3060, and 93.5 ms on the CPU i7-11800H. The RTX 3060 result demonstrates that the proposed model retains real-time inference capability on a relatively accessible GPU platform, whereas CPU-only inference remains feasible but exhibits substantially higher latency.

Compared with the two-stage detector Faster R-CNN, the proposed method showed substantial advantages in both efficiency and detection accuracy. Compared with one-stage detectors such as SSD and other YOLO-based methods, the proposed model demonstrated stronger competitiveness for UAV-based rose detection in greenhouse environments. These results confirm that the proposed improvements effectively enhance the baseline model and make it more suitable for practical deployment in resource-constrained agricultural scenarios.

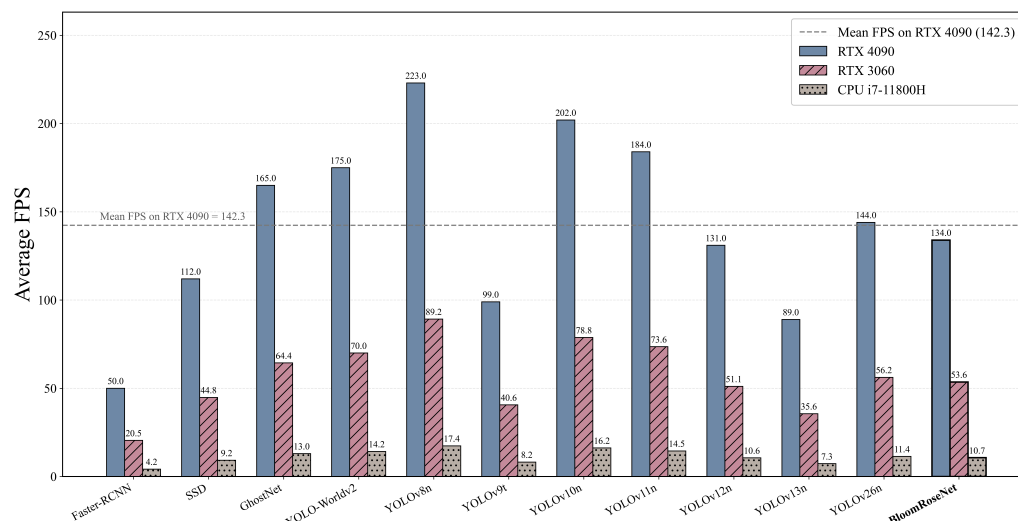


Figure 9. Cross-platform inference speed comparison of BloomRoseNet and other mainstream object detection methods on three hardware platforms, including RTX 4090, RTX 3060, and CPU i7-11800H.

3.3. Effect of Supplementary Multi-View Data on Detection Performance

To further verify the effectiveness of the supplementary multi-view rose data, the final model, BloomRoseNet, was trained under two different dataset settings: the complete GreenHouse Rose dataset and the Overhead Rose dataset. The GreenHouse Rose dataset consists of self-collected overhead images, RoseTracker data, and additionally screened and re-annotated multi-view rose images, whereas the Overhead Rose dataset contains only the self-collected overhead images and RoseTracker data. The quantitative comparison is summarized in Table 5.

Table 5. Effect of supplementary multi-view data on model performance under different training datasets.

Training Dataset	P (%)	R (%)	mAP@50 (%)	mAP@50:95 (%)	Params (M)	Size (MB)	FPS
GreenHouse Rose dataset	86.1	85.3	90.5	54.1	2.08	4.5	134
Overhead Rose dataset	85.2	82.8	89.2	54.1	2.08	4.5	134

Note: The GreenHouse Rose dataset includes self-collected overhead images, RoseTracker data, and additionally screened and re-annotated multi-view rose images. The Overhead Rose dataset includes only self-collected overhead images and RoseTracker data.

As shown in Table 5, BloomRoseNet trained on the GreenHouse Rose dataset achieved a precision of 86.1%, a recall of 85.3%, an mAP@50 of 90.5%, and an mAP@50:95 of 54.1%. In comparison, training on the Overhead Rose dataset resulted in a precision of 85.2%, a recall of 82.8%, an mAP@50 of 89.2%, and an mAP@50:95 of 54.1%. Under the same model configuration, the introduction of supplementary multi-view data improved precision by 0.9 percentage points, recall by 2.5 percentage points, and mAP@50 by 1.3 percentage points, while maintaining the same mAP@50:95 value. These results indicate that the supplementary data mainly enhanced the model's ability to detect more rose targets without reducing localization quality.

Figure 10 further provides a qualitative comparison of BloomRoseNet trained on the two dataset settings. Overall, the model trained on the GreenHouse Rose dataset produced more accurate predictions. In Sample (a), both models generated several false positive predictions for *rose_small*; however, the model trained on the final dataset assigned lower confidence scores to these incorrect detections, indicating better suppression of ambiguous background interference. In Sample (b), the model trained on the GreenHouse Rose dataset

also showed more reliable recognition of rose targets. These visual results are consistent with the quantitative comparison.

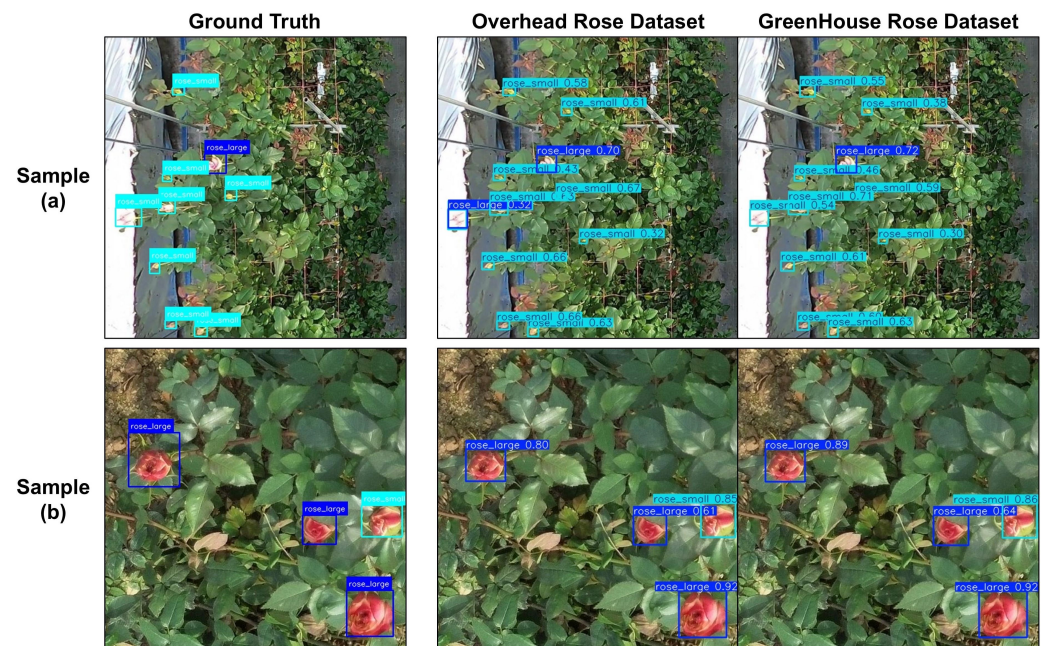


Figure 10. (a,b) Qualitative comparison of BloomRoseNet trained on different datasets.

Overall, the comparison confirms that the supplementary multi-view rose data effectively improved the detection performance of BloomRoseNet. This finding also supports the practical value of constructing the GreenHouse Rose dataset as a more diverse and representative training set for greenhouse rose detection.

3.4. Robustness Analysis

To further evaluate the reliability of the proposed method under practical greenhouse imaging disturbances, a robustness analysis was conducted by introducing four representative perturbation types into the test set, namely Gaussian blur, motion blur, random occlusion, and brightness variation. The corresponding results are shown in Figure 11.

These four perturbation settings were selected because they closely reflect the major sources of performance degradation in UAV-based rose monitoring. First, Gaussian blur was used to simulate image quality degradation caused by slight defocus, lens limitations, or imperfect reconstruction details in overhead images. Since the proposed task involves a large number of small rose buds, the loss of fine texture and boundary information may directly affect detection accuracy. Second, motion blur was introduced to mimic image smear caused by UAV motion, platform vibration, or unstable acquisition conditions during image capture. Compared with Gaussian blur, motion blur usually produces stronger directional distortion and is therefore more destructive to target localization. Third, random occlusion was considered to represent partial visibility loss in realistic greenhouse scenes, where rose flowers may be partially covered by leaves, branches, neighboring blooms, or crop structures. This factor is especially relevant to the detection of small and densely distributed targets. Fourth, brightness variation was applied to simulate illumination inconsistency caused by sunlight angle, greenhouse film transmission, local shadows, and exposure fluctuations. Such intensity changes are common in greenhouse environments and may alter the visual contrast between flowers and the background.

As illustrated in Figure 11, BloomRoseNet and other baseline detectors exhibited performance degradation as the perturbation intensity increased, but the degradation

trends were different across methods. Under Gaussian blur, the detection accuracy of all models gradually decreased with increasing blur strength, indicating that the loss of local texture and contour information negatively affected rose recognition. Under motion blur, the decline became more pronounced, especially at higher blur levels, confirming that directional image degradation is more harmful to accurate localization. Under the occlusion setting, all methods showed a clear drop in mAP@50 as the occlusion ratio increased, which suggests that partial disappearance of targets remains a major challenge in dense greenhouse scenes. For brightness variation, the performance of most models reached the highest level around the original illumination condition and decreased when the brightness offset became excessively positive or negative, forming an approximately inverted-U trend. This observation is consistent with the fact that both overexposure and underexposure reduce discriminative visual cues.

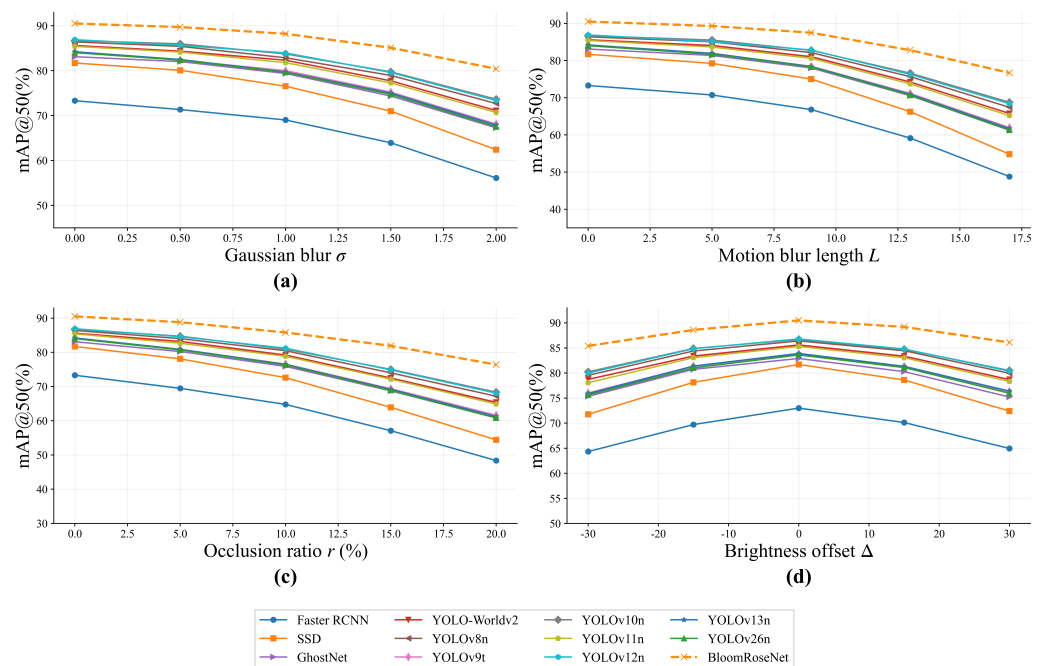


Figure 11. Robustness comparison of BloomRoseNet and other mainstream object detection methods under four perturbation settings: (a) Gaussian blur, (b) motion blur, (c) random occlusion, and (d) brightness variation.

Among all methods, BloomRoseNet consistently achieved the best performance across all four perturbation settings in Figure 12. In addition to obtaining the highest baseline accuracy under the clean test condition, the proposed model maintained a comparatively slower decline as the perturbation severity increased. This result indicates that the proposed improvements not only enhanced the detection accuracy on standard test images, but also improved the stability of feature extraction and target localization under degraded imaging conditions. In particular, the stronger robustness under blur, occlusion, and illumination changes is important for practical rose monitoring, where small flower buds are often more sensitive to visual quality loss and environmental interference.

Overall, the robustness analysis demonstrates that BloomRoseNet has better adaptability to common disturbances encountered in real greenhouse applications, which further supports its effectiveness and practical value for UAV-based rose detection.

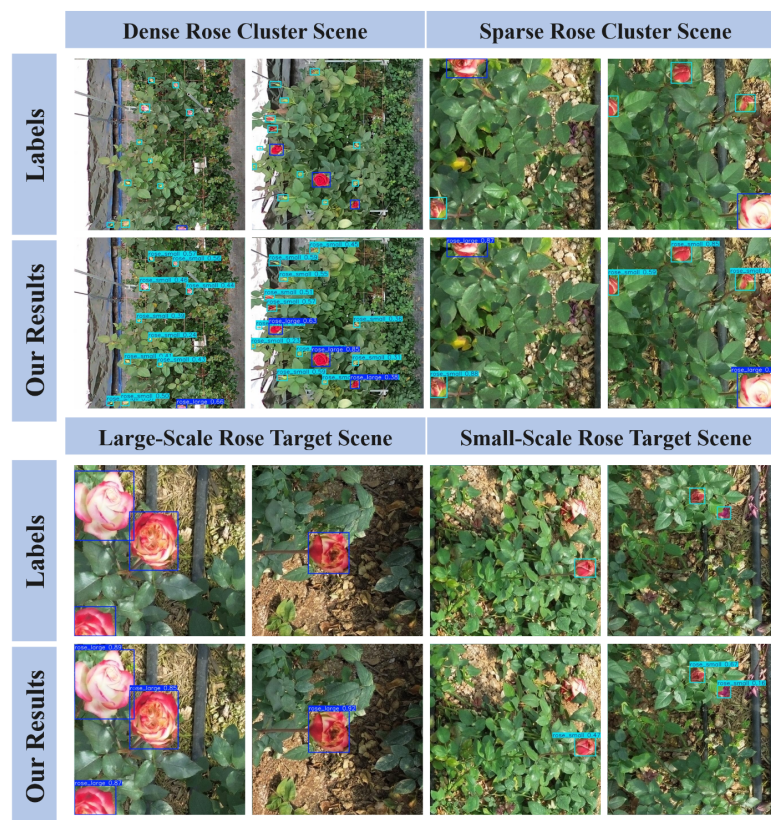


Figure 12. Representative qualitative examples for density- and scale-wise rose detection. The density examples compare rose scenes with different numbers of annotated instances in each image, and the scale examples compare rose targets with different bounding-box sizes. Ground-truth labels and BloomRoseNet predictions are shown for visual comparison.

3.5. Scene-Wise Performance Analysis

To further evaluate the detection performance of BloomRoseNet under different rose-density and object-scale conditions, the test set was reorganized into density-wise and scale-wise evaluation subsets. The density-wise evaluation groups images according to the number of annotated rose instances in each image, whereas the scale-wise evaluation groups individual rose instances according to the relative bounding-box area. Both *rose_large* and *rose_small* instances were included in the grouping process.

This analysis provides a more fine-grained evaluation than the overall test-set comparison. In practical greenhouse UAV imagery, the difficulty of rose detection is not only determined by the model architecture, but also by the number of visible rose targets and the visual size of each target. Dense images usually contain more occlusion and adjacent targets, while small-scale roses often provide fewer texture and boundary cues. Therefore, the density-wise and scale-wise evaluations are used to examine whether the proposed model remains reliable under these challenging conditions.

As summarized in Table 6, the density-wise evaluation contains 149 sparse images, 68 medium-density images, and 33 dense images. Although the dense group has the fewest images, it contains 300 rose instances, indicating a much higher target concentration. For the scale-wise evaluation, small targets account for the majority of the test instances, with 562 objects, while medium and large targets contain 161 and 46 objects, respectively. This distribution confirms that small-target detection is a central challenge in the proposed benchmark.

Table 6. Group definitions and sample statistics for scene-wise evaluation.

Setting	Group	Criterion	Images	Instances
Rose density	Sparse	1–2 instances per image	149	209
Rose density	Medium	3–5 instances per image	68	260
Rose density	Dense	6 or more instances per image	33	300
Rose scale	Small	Box area ratio below 1%	–	562
Rose scale	Medium	Box area ratio from 1% to 5%	–	161
Rose scale	Large	Box area ratio above 5%	–	46

3.5.1. Image-Level Rose Density Analysis

For image-level density analysis, the test images were divided into sparse, medium-density, and dense groups according to the number of annotated rose instances in each image. This evaluation examines whether a detector can maintain stable performance when the number of rose instances increases within a single image. Dense scenes are more challenging because adjacent roses, overlapping foliage, and small buds can easily lead to missed detections, duplicated predictions, or inaccurate localization.

As shown in Table 7, BloomRoseNet achieves the best mAP@50 in all three density groups. In sparse scenes, the proposed model obtains 91.2%, outperforming YOLOv12n by 2.8 percentage points. In medium-density scenes, BloomRoseNet reaches 90.7%, which is 3.7 percentage points higher than YOLOv12n. In dense scenes, where target overlap and background interference become more severe, BloomRoseNet achieves 89.0%, improving over YOLOv12n by 5.4 percentage points.

Table 7. Comparison with mainstream detectors under image-level rose-density groups. Group-wise detection performance is reported using mAP@50 (%), together with Recall (%) for the dense group. “Density mean” indicates the average mAP@50 across the sparse, medium-density, and dense groups, while “Dense gain” denotes the mAP@50 improvement over YOLOv12n in dense scenes.

Model	Overall	Sparse	Medium	Dense	Dense Recall	Density Mean	Dense Gain
Faster R-CNN	73.3	75.8	73.6	67.9	64.8	72.4	−15.7
SSD	81.7	83.1	81.6	78.1	75.6	80.9	−5.5
GhostNet	83.1	84.7	83.0	79.3	77.1	82.3	−4.3
RT-DETR-L	87.8	89.1	88.0	85.2	82.7	87.4	+1.6
Mamba-YOLO-T	88.1	89.4	88.4	85.7	83.4	87.8	+2.1
YOLO-Worldv2	85.6	87.2	85.5	81.7	79.1	84.8	−1.9
YOLOv8n	86.4	87.8	86.2	82.8	80.5	85.6	−0.8
YOLOv9t	84.2	85.7	84.0	80.4	77.9	83.4	−3.2
YOLOv10n	86.7	88.1	86.5	83.2	80.8	85.9	−0.4
YOLOv11n	85.4	86.8	85.3	81.9	79.3	84.7	−1.7
YOLOv12n	86.9	88.4	87.0	83.6	81.2	86.3	0.0
YOLOv13n	84.2	85.5	84.1	80.1	77.5	83.2	−3.5
YOLO26n	84.0	85.3	83.8	79.8	77.1	83.0	−3.8
BloomRoseNet	90.5	91.2	90.7	89.0	87.3	90.3	+5.4

Note: Bold values indicate the best performance in each column.

The dense-scene results are particularly important because this group contains only 33 images but 300 rose instances, indicating that each image includes a much larger number of targets. Most detectors show a clear performance decline from sparse to dense scenes. For example, YOLOv12n decreases from 88.4% to 83.6%, while Mamba-YOLO-T decreases from 89.4% to 85.7%. In contrast, BloomRoseNet decreases only from 91.2% to 89.0%, showing stronger stability under crowded greenhouse conditions.

Compared with other competitive detectors, BloomRoseNet also maintains a clear advantage in the dense group. Its dense-scene mAP@50 is 3.3 percentage points higher than Mamba-YOLO-T and 3.8 percentage points higher than RT-DETR-L. This indicates that the proposed model is more effective in separating adjacent rose targets and retaining small buds under high-density conditions. The result is consistent with the design motivation of BloomRoseNet, which aims to improve dense, small, and visually similar rose detection. The Recall results provide further evidence of the model's ability to reduce missed detections in crowded scenes. BloomRoseNet achieves a dense-scene Recall of 87.3%, exceeding YOLOv12n, Mamba-YOLO-T, and RT-DETR-L by 6.1, 3.9, and 4.6 percentage points, respectively, indicating that more rose instances are successfully retained under severe overlap and background interference.

The density mean further reflects the balanced performance of each detector across sparse, medium-density, and dense scenes. BloomRoseNet achieves the highest density mean of 90.3%, exceeding YOLOv12n by 4.0 percentage points. This result shows that its advantage is not limited to a single density condition. Instead, the proposed model maintains strong performance across different image-level target distributions, which is important for practical greenhouse rose monitoring where target density may vary substantially across planting rows and growth stages.

3.5.2. Instance-Level Rose Scale Analysis

For instance-level scale analysis, rose targets were divided into small, medium, and large groups according to the relative bounding-box area. This grouping reflects the physical size of each target in the image and is independent of the semantic categories *rose_large* and *rose_small*. During scale-wise evaluation, ground-truth instances outside the selected scale range were ignored during matching to avoid penalizing detections from other scale groups.

As shown in Table 8, all detectors generally achieve higher accuracy on large rose targets than on small rose targets. This trend indicates that target size has a clear influence on detection performance. Small roses contain fewer texture details and weaker boundary information, and they are more likely to be confused with leaves, stems, and background structures. Therefore, the small-scale group represents the most challenging part of the scale-wise evaluation.

Table 8. Comparison with mainstream detectors under instance-level rose-scale groups. Group-wise detection performance is reported using mAP@50 (%), together with Recall (%) for the small-target group. “Scale mean” indicates the average mAP@50 across the small, medium, and large target groups, while “Small gain” denotes the mAP@50 improvement over YOLOv12n for small rose targets.

Model	Overall	Small	Small Recall	Medium	Large	Scale Mean	Small Gain
Faster R-CNN	73.3	69.8	67.0	75.4	80.3	75.2	−15.2
SSD	81.7	79.1	76.5	82.7	86.5	82.8	−5.9
GhostNet	83.1	80.4	78.4	84.0	87.3	83.9	−4.6
RT-DETR-L	87.8	85.7	83.5	88.8	91.1	88.5	+0.7
Mamba-YOLO-T	88.1	86.2	84.3	89.0	91.6	88.9	+1.2
YOLO-Worldv2	85.6	83.2	80.5	86.3	89.7	86.4	−1.8
YOLOv8n	86.4	83.9	81.6	87.1	90.0	87.0	−1.1
YOLOv9t	84.2	81.7	79.4	85.1	88.4	85.1	−3.3
YOLOv10n	86.7	84.3	82.1	87.4	90.3	87.3	−0.7
YOLOv11n	85.4	82.8	80.2	86.2	89.2	86.1	−2.2
YOLOv12n	86.9	85.0	82.8	88.5	91.8	88.4	0.0
YOLOv13n	84.2	81.5	79.0	84.9	88.1	84.8	−3.5
YOLO26n	84.0	81.1	78.7	84.7	87.6	84.5	−3.9
BloomRoseNet	90.5	89.2	87.5	90.8	92.4	90.8	+4.2

Note: Bold values indicate the best performance in each column.

BloomRoseNet achieves the best performance in all three scale groups, with mAP@50 values of 89.2%, 90.8%, and 92.4% for small, medium, and large targets, respectively. Compared with YOLOv12n, the proposed model improves small-target mAP@50 by 4.2 percentage points, medium-target mAP@50 by 2.3 percentage points, and large-target mAP@50 by 0.6 percentage points. The improvement is most prominent on small targets, which directly supports the effectiveness of the proposed model for small rose buds and distant rose targets.

Compared with other strong baselines, BloomRoseNet also shows clear advantages in small-target detection. It exceeds Mamba-YOLO-T by 3.0 percentage points and RT-DETR-L by 3.5 percentage points on the small-scale group. This result is meaningful because both Mamba-YOLO-T and RT-DETR-L achieve strong overall performance in the mainstream comparison. The stronger small-target result suggests that BloomRoseNet provides more suitable feature representation for greenhouse rose detection, especially when target details are limited. A similar advantage is observed in the Recall results for small targets. BloomRoseNet obtains a small-target Recall of 87.5%, outperforming YOLOv12n, Mamba-YOLO-T, and RT-DETR-L by 4.7, 3.2, and 4.0 percentage points, respectively, which directly demonstrates its stronger ability to reduce missed detections of small rose buds and distant rose instances.

The scale mean further demonstrates the balanced detection ability of BloomRoseNet across different target sizes. The proposed model achieves a scale mean of 90.8%, which is 2.4 percentage points higher than YOLOv12n and 1.9 percentage points higher than Mamba-YOLO-T. This indicates that BloomRoseNet improves the detection of small targets without sacrificing performance on medium and large targets. Such balanced scale-wise performance is important for UAV greenhouse imagery, where roses may appear at different distances and under different canopy conditions.

Overall, the density-wise and scale-wise analyses provide complementary evidence beyond the overall test-set metrics. The density-wise evaluation reflects image-level complexity caused by different numbers of rose instances, whereas the scale-wise evaluation reflects instance-level difficulty caused by different bounding-box sizes. The results show that BloomRoseNet is not only superior in overall mAP@50, but also more reliable in dense scenes and small-target detection scenarios. These findings further demonstrate its practical value for greenhouse rose monitoring under diverse field conditions.

4. Discussion

4.1. Significance of the Multi-Source Dataset Construction

One of the main contributions of this study is the construction of the GreenHouse Rose dataset by integrating self-collected UAV overhead images, screened RoseTracker images, and supplementary multi-view rose images. The purpose of this strategy was not merely to increase the number of training samples, but to broaden the visual distribution of rose targets in terms of viewpoint, scale, flowering stage, local occlusion, and background composition. Such diversity is important because rose appearance may vary substantially even within relatively controlled greenhouse environments.

The experimental results support the effectiveness of this data-integration strategy. When BloomRoseNet was trained on the complete GreenHouse Rose dataset, precision, recall, and mAP@50 reached 86.1%, 85.3%, and 90.5%, respectively. In comparison, training on the Overhead Rose dataset alone resulted in precision of 85.2%, recall of 82.8%, and mAP@50 of 89.2%. The most evident gain was the 2.5 percentage-point increase in recall, indicating that the supplementary multi-view data helped the detector recognize a wider range of target appearances and reduced missed detections.

This improvement is particularly relevant to greenhouse phenotyping, where small or partially occluded roses may be difficult to distinguish from leaves, stems, and neighboring flowers. The simultaneous improvements in precision and recall indicate that the supplementary data did not merely increase model sensitivity, but also improved target discrimination. In contrast, mAP@50:95 remained unchanged at 54.1%, suggesting that the main benefit of the additional data was improved target recognition and appearance coverage rather than more accurate high-*IoU* boundary localization. For flower counting and developmental-stage monitoring, this result remains practically meaningful because reliable target discovery is generally more important than marginally tighter bounding-box alignment.

Nevertheless, the supplementary multi-view images cannot fully represent all variations encountered in production greenhouses. Differences in greenhouse covering materials, illumination systems, planting density, canopy structure, camera settings, and UAV imaging height may alter both target appearance and background composition. In addition, although the supplementary data include roses with different colors and viewpoints, the primary UAV data remain dominated by healthy red roses. The current dataset therefore provides broader visual diversity than a single-source dataset, but it should not be interpreted as covering all greenhouse conditions or rose cultivars.

4.2. Effectiveness of the Proposed Model Improvements

The ablation experiments demonstrate the complementary contributions of S2C2f, AFFM, and Detect_DPA to feature extraction, multi-scale fusion, and prediction refinement. S2C2f produced the most evident initial improvement: recall increased from 82.0% to 85.6%, mAP@50 from 86.9% to 88.2%, and mAP@50:95 from 52.5% to 53.1%. Meanwhile, the parameter count decreased from 2.55 M to 2.26 M and the inference speed increased from 131 FPS to 187 FPS. These results indicate that S2C2f strengthened fine-grained feature representation while reducing computational complexity, which is particularly beneficial for small rose buds with limited texture and boundary information.

The introduction of AFFM further increased recall to 86.6%, the highest among all ablation variants, suggesting that adaptive fusion of shallow spatial details and deeper semantic information improved the recovery of small, overlapping, and partially occluded targets. Detect_DPA subsequently increased precision from 81.9% to 86.1%, while mAP@50 and mAP@50:95 reached 90.5% and 54.1%, respectively. Although recall decreased slightly from 86.6% to 85.3%, it remained 3.3 percentage points above the baseline, indicating that Detect_DPA reduced false positives and achieved a better precision–recall balance.

Overall, the complete BloomRoseNet improved precision, recall, mAP@50, and mAP@50:95 over YOLOv12n by 3.2, 3.3, 3.6, and 1.6 percentage points, respectively. At the same time, the parameter count decreased from 2.55 M to 2.08 M and the model size from 5.5 MB to 4.5 MB. Thus, the proposed modules improve detection accuracy and efficiency through task-oriented feature enhancement rather than increased model capacity.

4.3. Implications for Small-Object Detection in Complex Greenhouse Scenes

Greenhouse UAV imagery presents substantial small-object detection challenges because rose buds occupy few pixels and are frequently affected by overlap, background interference, blur, and illumination variation. BloomRoseNet achieved the best overall performance among the evaluated detectors, with precision, recall, mAP@50, and mAP@50:95 of 86.1%, 85.3%, 90.5%, and 54.1%, respectively. The simultaneous improvement in precision and recall indicates that the model achieved a favorable balance between recovering difficult targets and suppressing false-positive responses, which is important for greenhouse phenotyping and flower counting.

BloomRoseNet also maintained a lightweight structure and real-time inference capability. The model contained only 2.08 M parameters, occupied 4.5 MB, and reached 134 FPS. Its stable performance across RTX 4090, RTX 3060, and Intel i7-11800H platforms further indicates its potential for greenhouse applications with different computing resources.

The same-dataset comparison with Rose-Mamba-YOLO provides the most direct task-specific benchmark. Rose-Mamba-YOLO introduces Mamba-inspired state-space modeling and the C3k2_RFCBAM, MDFPC, and CMFPN modules to enhance long-range contextual representation and multi-scale feature fusion [10]. In contrast, BloomRoseNet uses S2C2f and AFFM to strengthen lightweight fine-grained feature extraction and the adaptive fusion of high-resolution P3 and P4 features. Under the unified benchmark, BloomRoseNet achieved higher detection accuracy with fewer parameters and less model storage, indicating that targeted feature enhancement provides a favorable accuracy–efficiency trade-off for small rose targets.

Other greenhouse flower-detection systems address related but different operational requirements. VM-YOLO uses a VMamba-based architecture for clustered strawberry-flower detection but contains approximately 30 M parameters [69], whereas Wang et al. [18] combined YOLOv4-Tiny, a ZED 2 stereo camera, and Jetson TX2 for close-range three-dimensional flower localization. These published results cannot be directly ranked because they were obtained using different flower species, datasets, hardware platforms, and evaluation protocols. Nevertheless, they show that BloomRoseNet provides a comparatively compact solution for wide-view UAV-based phenotyping rather than close-range robotic localization.

The current framework remains limited to two-dimensional bounding-box detection, and its practical performance on embedded platforms such as the Jetson series has not yet been validated. Future deployment could further integrate visual-inertial localization, robust flight control, and auto-landing methods to improve image acquisition stability in repetitive greenhouse environments [70–73]. Overall, the results demonstrate that multi-source data construction combined with targeted lightweight architectural improvements is an effective strategy for small-object detection in complex greenhouse scenes.

4.4. Robustness and Practical Applicability

The robustness experiments further strengthen the practical value of the proposed method. In real greenhouse operation, the image quality is not always stable. Blur may be caused by slight defocus, motion, or reconstruction imperfection. Occlusion may come from leaves, stems, neighboring flowers, or cultivation structures. Brightness changes are also common due to sunlight angle, local shadow, and greenhouse covering materials. These factors directly affect the visibility of small rose targets.

Under all four perturbation settings, namely Gaussian blur, motion blur, random occlusion, and brightness variation, all compared models showed performance decline as the disturbance became stronger. However, BloomRoseNet consistently achieved the best overall performance and showed a slower degradation trend. This result is meaningful because it indicates that the proposed method is not only accurate on clean test images, but also more stable when visual conditions become unfavorable.

This robustness is closely related to both the dataset design and the model design. On the data side, the supplementary multi-view rose images increased the diversity of visual patterns seen during training. On the model side, the three proposed modules improved feature extraction, feature fusion, and final prediction quality. Their joint effect likely helped the detector preserve more useful information under degraded conditions. For practical greenhouse monitoring, such stability is highly important, because real data collection rarely follows ideal imaging conditions.

4.5. Cross-Source Adaptation on External Rose-Related Flower Datasets

To evaluate the adaptability of BloomRoseNet beyond the GreenHouse Rose dataset, four public flower image datasets were screened and re-annotated for rose-related object detection: *Flowers Recognition*, *Rose Detection Dataset*, *Flower Classification 10 Classes*, and *flower-datasets*. Images with severe blur, extremely small or ambiguous targets, or duplicate content were excluded, and only rose-related images were retained. After screening, the four datasets contained 300, 565, 280, and 275 images, respectively.

To maintain consistency with the main task, all retained images were annotated using two developmental-stage categories: *flower_large* for fully opened flowers and *flower_small* for buds or early-stage flowers. Each dataset was independently divided into training, validation, and test subsets at a ratio of 8:1:1. YOLOv12n, YOLOv13n, and BloomRoseNet were trained separately on the same fixed split of each dataset using an input size of 640×640 pixels and the same optimization, augmentation, and training settings as the main experiments. Therefore, this experiment evaluates dataset-specific adaptation after retraining rather than direct cross-domain inference.

As shown in Table 9, BloomRoseNet achieved the highest mAP@50 on all four external datasets, outperforming the best baseline by 2.8–3.4 percentage points. On *Flowers Recognition* and *Rose Detection Dataset*, mAP@50 increased from 87.4% to 90.8% and from 88.1% to 91.2%, respectively. These datasets generally contain clearer flower regions and relatively simple foreground structures, indicating that the proposed modules can improve localization even under less complex visual conditions.

Table 9. Adaptability evaluation on external rose-related flower datasets under a unified detection protocol.

Dataset	YOLOv12n	YOLOv13n	Best Baseline	BloomRoseNet	Δ
Flowers Recognition	87.4	86.9	87.4	90.8	+3.4
Rose Detection	88.1	87.6	88.1	91.2	+3.1
Flower Classification 10 Classes	84.9	85.5	85.5	88.6	+3.1
flower-datasets	86.7	86.1	86.7	89.5	+2.8

Note: The reported values are mAP@50 (%). Δ denotes the improvement of BloomRoseNet over the best baseline on each external dataset. The number of retained images varied among datasets because only images suitable for rose-related flower detection were included. Bold values indicate the best performance in each column.

The retained images from *Flower Classification 10 Classes* and *flower-datasets* showed greater variation in flower color, morphology, viewpoint, scale, and background composition. Under these more heterogeneous conditions, BloomRoseNet achieved mAP@50 values of 88.6% and 89.5%, exceeding the best baselines by 3.1 and 2.8 percentage points, respectively. The consistent improvements across all four datasets indicate that the proposed architecture retains its advantage after adaptation to rose-related images from different sources.

However, these results demonstrate architectural adaptability after target-domain retraining rather than zero-shot generalization to previously unseen datasets. Changes in greenhouse structure, illumination, imaging platform, cultivar morphology, and annotation criteria may still cause substantial domain shifts. Future work should therefore include direct cross-domain or leave-one-dataset-out evaluation without retraining to assess generalization to unseen greenhouses and rose cultivars more rigorously.

4.6. Failure Case Analysis

Although BloomRoseNet achieved strong overall performance, several challenging cases remained, as shown in Figure 13. All erroneous detections on the complete test set were manually reviewed and assigned to one dominant cause. Among the 82 identified failures, 10 were attributed to petal boundary ambiguity (12.2%), 14 to bud–leaf morphological similarity (17.1%), 25 to greenhouse background interference (30.5%), and

33 to bud–bloom category ambiguity (40.2%). Thus, category ambiguity and background interference together accounted for 70.7% of all failures, indicating that fine-grained developmental-stage discrimination and suppression of confusing background responses remain the principal bottlenecks.

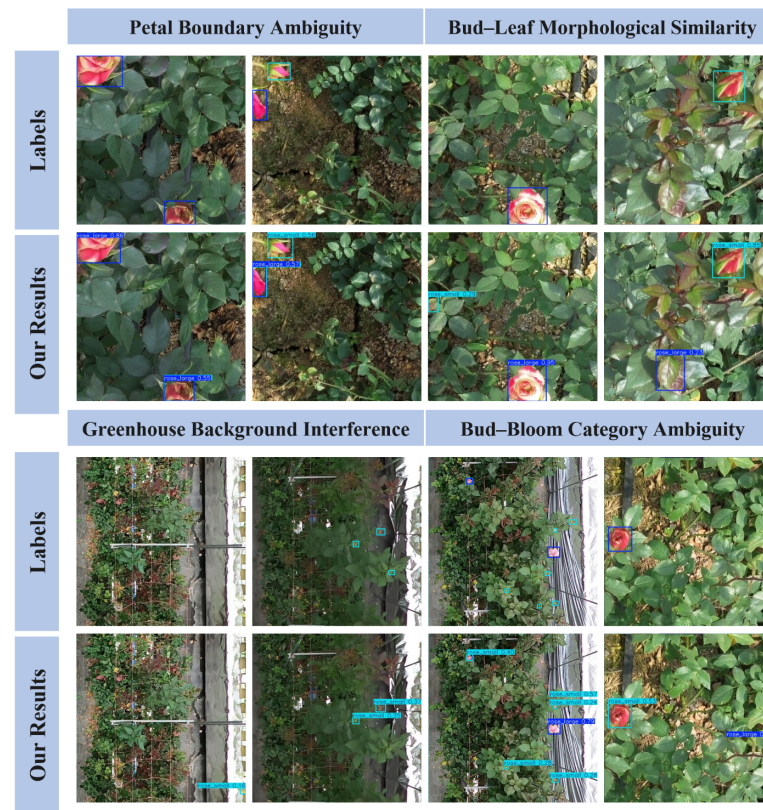


Figure 13. Failure case analysis of BloomRoseNet in challenging greenhouse rose scenes. The selected examples include petal boundary ambiguity, bud–leaf morphological similarity, greenhouse background interference, and bud–bloom category ambiguity. For each case, the ground-truth labels and the prediction results of BloomRoseNet are shown for comparison.

Petal boundary ambiguity mainly occurred when roses were heavily occluded or truncated near image boundaries. In such cases, incomplete contours caused inaccurate or low-confidence bounding boxes, revealing the limitation of rectangular detection for irregular and partially visible petals. Bud–leaf similarity arose because small buds, reddish young leaves, and folded leaf tips may share similar color, shape, and texture. Although S2C2f and AFFM improve fine-grained feature representation, these visually similar reproductive and vegetative structures remain difficult to distinguish under dense canopies.

Greenhouse background interference was associated with support frames, soil regions, shadows, reflective surfaces, dense foliage, and small reddish fragments that generated target-like responses. This problem was further intensified for distant roses because their weak texture and boundary cues were easily lost in wide-view UAV images. Bud–bloom category ambiguity was the most frequent error source because rose development is continuous, whereas the current annotation protocol uses only the two discrete classes *rose_small* and *rose_large*. Partially opened or side-view flowers may contain both bud-like and bloom-like characteristics, increasing confusion near the class boundary.

These errors also reveal limitations in the current data distribution and task definition. The primary UAV dataset mainly contains healthy red roses, while the supplementary multi-view images introduce additional colors and viewpoints but cannot fully represent cultivars with substantially different petal morphology, bud shape, or canopy structure.

Pest- and disease-affected roses were also outside the present scope. Consequently, the current results should not be interpreted as demonstrating uniform performance across all cultivars, plant-health conditions, or greenhouse environments.

Future work should prioritize failure-oriented data collection, including heavily occluded roses, reddish leaves, complex greenhouse structures, intermediate flowering stages, diverse cultivars, and diseased plants. Hard-negative mining may improve background suppression, while finer phenological labels or a hierarchical detection–classification scheme could reduce bud–bloom ambiguity. Class-balanced sampling, adaptive loss weighting, or focal loss may also be considered when more pronounced category imbalance is encountered.

At the model level, boundary-aware localization, instance segmentation, or mask-assisted detection could better represent incomplete petal contours, while high-resolution branches, tile-based inference, and multi-view or temporal information may improve the detection of distant or occluded targets. Uncertainty-aware classification and confidence calibration may further reduce overconfident predictions for transitional stages. For multi-greenhouse deployment, federated learning and image-security methods could support privacy-preserving model collaboration and data transmission [74–76], while environmental time-series modeling may complement visual observations in future flowering-dynamics analysis [77,78].

Overall, the quantitative failure analysis provides targeted guidance for improving boundary representation, hard-negative learning, small-object enhancement, and fine-grained rose developmental-stage recognition.

5. Conclusions

This study developed a lightweight and practical framework for rose detection in UAV-derived greenhouse imagery, with particular emphasis on distinguishing blooming roses from flower buds under complex cultivation conditions. To address the challenges of small targets, dense distribution, occlusion, and heterogeneous backgrounds, a multi-source dataset construction strategy and an improved YOLOv12n-based detector, BloomRoseNet, were jointly proposed. The *GreenHouse Rose dataset*, constructed by integrating self-collected UAV overhead images, screened RoseTracker images, and supplementary multi-view rose images, improved the diversity of rose appearance and provided a more representative benchmark for greenhouse rose detection. Experimental results further showed that the supplementary multi-view data improved precision, recall, and mAP@50, confirming the value of enhanced data diversity for detecting challenging rose targets, especially small buds in dense scenes.

Meanwhile, BloomRoseNet achieved the best overall performance among all compared methods. Compared with the baseline YOLOv12n, it improved precision, recall, mAP@50, and mAP@50:95 by 3.2, 3.3, 3.6, and 1.6 percentage points, respectively, while reducing the parameter count from 2.55 M to 2.08 M and the model size from 5.5 MB to 4.5 MB. In addition, the model maintained real-time inference capability and showed stronger robustness under Gaussian blur, motion blur, random occlusion, and brightness variation. Overall, the results demonstrate that combining multi-source data construction with targeted lightweight model improvement is an effective strategy for greenhouse rose detection in UAV imagery, and that the proposed framework provides useful support for large-area greenhouse phenotyping and intelligent cultivation management. The applicability of BloomRoseNet is currently bounded by greenhouse scenes with imaging conditions and rose appearances similar to those represented in the training data, while severe occlusion, complex background interference, extremely small distant targets, and ambiguous transitional flowering stages may still reduce its reliability. Future work will therefore

focus on improving hard-negative learning, small-object and boundary representation, and fine-grained recognition of continuous rose developmental stages.

Author Contributions: Conceptualization, Y.W. and F.Z.; methodology, Y.W. and F.Z.; software, Y.W. and J.H.; validation, P.Z., W.C. and S.S.; formal analysis, Y.W. and J.H.; investigation, Y.W., P.Z. and W.C.; resources, N.L. and F.Z.; data curation, Y.W. and S.S.; writing—original draft preparation, Y.W.; writing—review and editing, J.H., F.X., Z.Q., H.W., N.L. and F.Z.; visualization, Y.W. and P.Z.; supervision, N.L. and F.Z.; project administration, N.L. and F.Z.; funding acquisition, N.L. and F.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was partially supported by the Japan Science and Technology Agency’s SPRING Program (JST SPRING), Grant No. JPMJSP2108; the Industry-Education Integration and New Business Development Research Center Project, Grant No. 2025-KYJD07; the Southern Xinjiang Joint Youth Fund of Xinjiang Uygur Autonomous Region Natural Science Foundation, Grant No. 20261155210; and the 2026 University-level Scientific Research Fund of Xinjiang University of Science and Technology, Grant No. 2026-KYPT37. 2025 University-level Scientific Research Fund, Xinjiang University of Science and Technology, Grant No. 2025-KYPT33.

Data Availability Statement: The self-collected Greenhouse Rose image dataset can be accessed at the following link: https://github.com/ZhaoUTokyo/GreenHouse_Rose-dataset accessed on 21 July 2026. Other data will be made available on request.

Acknowledgments: The authors would like to thank all collaborators who supported data collection, annotation, and manuscript preparation.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Adebayo, I.A.; Pam, V.K.; Arsad, H.; Samian, M.R. The Global Floriculture Industry: Status and Future Prospects. In *The Global Floriculture Industry*; Apple Academic Press: Palm Bay, FL, USA, 2020; pp. 1–14. [\[CrossRef\]](#)
2. Shinoda, R.; Motoki, K.; Hara, K.; Kataoka, H.; Nakano, R.; Nakazaki, T.; Noguchi, R. RoseTracker: A system for automated rose growth monitoring. *Smart Agric. Technol.* **2023**, *5*, 100271. [\[CrossRef\]](#)
3. Costa, A.P.; Pôccas, I.; Cunha, M. Estimating the Leaf Area of Cut Roses in Different Growth Stages Using Image Processing and Allometrics. *Horticulturae* **2016**, *2*, 6. [\[CrossRef\]](#)
4. Fotouo Makouate, H.; Zude-Sasse, M. Advances in Growing Degree Days Models for Flowering to Harvest: Optimizing Crop Management with Methods of Precision Horticulture—A Review. *Horticulturae* **2025**, *11*, 1415. [\[CrossRef\]](#)
5. Proietti, S.; Scariot, V.; De Pascale, S.; Paradiso, R. Flowering Mechanisms and Environmental Stimuli for Flower Transition: Bases for Production Scheduling in Greenhouse Floriculture. *Plants* **2022**, *11*, 432. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Zheng, C.; Abd-Elrahman, A.; Whitaker, V. Remote Sensing and Machine Learning in Crop Phenotyping and Management, with an Emphasis on Applications in Strawberry Farming. *Remote Sens.* **2021**, *13*, 531. [\[CrossRef\]](#)
7. Li, J.; Li, M. Flowering Index Intelligent Detection of Spray Rose Cut Flowers Using an Improved YOLOv5s Model. *Appl. Sci.* **2024**, *14*, 9879. [\[CrossRef\]](#)
8. Körschens, M.; Bucher, S.F.; Römermann, C.; Denzler, J. Unified Automatic Plant Cover and Phenology Prediction. In *Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*; IEEE : New York, NY, USA, 2023; pp. 685–693. [\[CrossRef\]](#)
9. Tsouros, D.C.; Bibi, S.; Sarigiannidis, P.G. A Review on UAV-Based Applications for Precision Agriculture. *Information* **2019**, *10*, 349. [\[CrossRef\]](#)
10. You, S.; Li, B.; Chen, Y.; Ren, Z.; Liu, Y.; Wu, Q.; Tao, J.; Zhang, Z.; Zhang, C.; Xue, F.; et al. Rose-Mamba-YOLO: An enhanced framework for efficient and accurate greenhouse rose monitoring. *Front. Plant Sci.* **2025**, *16*, 1607582. [\[CrossRef\]](#) [\[PubMed\]](#)
11. Li, H.; Zhao, F.; Xue, F.; Wang, J.; Liu, Y.; Chen, Y.; Wu, Q.; Tao, J.; Zhang, G.; Kobayashi, H.H. Succulent-YOLO: Smart UAV-Assisted Succulent Farmland Monitoring with CLIP-Based YOLOv10 and Mamba Computer Vision. *Remote Sens.* **2025**, *17*, 2219. [\[CrossRef\]](#)
12. Gallmann, J.; Schüpbach, B.; Jacot, K.; Albrecht, M.; Winizki, J.; Kirchgessner, N.; Aasen, H. Flower Mapping in Grasslands with Drones and Deep Learning. *Front. Plant Sci.* **2022**, *12*, 774965. [\[CrossRef\]](#) [\[PubMed\]](#)
13. Schnalke, M.; Funk, J.; Wagner, A. Bridging Technology and Ecology: Enhancing Applicability of Deep Learning and UAV-Based Flower Recognition. *Front. Plant Sci.* **2025**, *16*, 1498913. [\[CrossRef\]](#) [\[PubMed\]](#)

14. Shajeena, J.; Govindasamy, B.; Gnanasundaram, M.; Robinson Joel, M. Mobile-Le Harmonic Fusion Network for Object Recognition and SiamMoT Based Multi-Object Tracking Using Video Surveillance. *Cybern. Syst.* **2026**, *57*, 866–896.
15. Zhao, F.; Liu, Y.; Wang, J.; Chen, Y.; Xi, D.; Shao, X.; Tabeta, S.; Mizuno, K. Riverbed Litter Monitoring Using Consumer-Grade Aerial-Aquatic Speedy Scanner (AASS) and Deep Learning Based Super-Resolution Reconstruction and Detection Network. *Mar. Pollut. Bull.* **2024**, *209*, 117030. [[CrossRef](#)] [[PubMed](#)]
16. Yang, X.; Yang, J.; Yan, J.; Zhang, Y.; Zhang, T.; Guo, Z.; Sun, X.; Fu, K. SCRDet: Towards More Robust Detection for Small, Cluttered and Rotated Objects. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 8231–8240. [[CrossRef](#)]
17. Vierbergen, W.; Willekens, A.; Dekeyser, D.; Cool, S.; Wyffels, F. Sim2real Flower Detection towards Automated Calendula Harvesting. *Biosyst. Eng.* **2023**, *234*, 125–139. [[CrossRef](#)]
18. Wang, J.; Gao, Z.; Zhang, Y.; Zhou, J.; Wu, J.; Li, P. Real-Time Detection and Location of Potted Flowers Based on a ZED Camera and a YOLO V4-Tiny Deep Learning Algorithm. *Horticulturae* **2022**, *8*, 21. [[CrossRef](#)]
19. Park, H.M.; Park, J.H. YOLO Network with a Circular Bounding Box to Classify the Flowering Degree of Chrysanthemum. *AgriEngineering* **2023**, *5*, 1530–1543. [[CrossRef](#)]
20. Zhao, W.; Wu, D.; Zheng, X. Detection of Chrysanthemums Inflorescence Based on Improved CR-YOLOv5s Algorithm. *Sensors* **2023**, *23*, 4234. [[CrossRef](#)] [[PubMed](#)]
21. Zeng, S.; Chang, X.; Xie, M.; Liu, X.; Bai, Y.; Pan, Z.; Xu, M.; Wei, X. FutureSightDrive: Thinking Visually with Spatio-Temporal CoT for Autonomous Driving. *arXiv* **2025**, arXiv:2505.17685.
22. Zhang, F.; Lv, Z.; Zhang, H.; Guo, J.; Wang, J.; Lu, T.; Zhangzhong, L. Verification of Improved YOLOX Model in Detection of Greenhouse Crop Organs: Considering Tomato as Example. *Comput. Electron. Agric.* **2023**, *205*, 107582. [[CrossRef](#)]
23. Li, X.; Du, J.; Chen, X.; Shi, F.; Li, S. Deep Learning-Based Kiwifruit Flower Recognition Method to Facilitate Automated Pollination. *Appl. Soft Comput.* **2025**, *184*, 113855. [[CrossRef](#)]
24. Uh, D.; Kumar, J.P. Enhanced Transfer Learning-Based CNN for Abnormal Human Activity Detection in Video Surveillance Using Spatial-Temporal Features. *Cybern. Syst.* **2025**, 1–30. [[CrossRef](#)]
25. Zhao, F.; Huang, B.; Wang, J.; Shao, X.; Wu, Q.; Xi, D.; Liu, Y.; Chen, Y.; Zhang, G.; Ren, Z.; et al. Seafloor Debris Detection Using Underwater Images and Deep Learning-Driven Image Restoration: A Case Study from Koh Tao, Thailand. *Mar. Pollut. Bull.* **2025**, *214*, 117710. [[CrossRef](#)] [[PubMed](#)]
26. Bhattarai, U.; Bhusal, S.; Zhang, Q.; Karkee, M. AgRegNet: A Deep Regression Network for Flower and Fruit Density Estimation, Localization, and Counting in Orchards. *Comput. Electron. Agric.* **2024**, *227*, 109534. [[CrossRef](#)]
27. Estrada, J.S.; Vasconez, J.P.; Fu, L.; Auat Cheein, F. Deep Learning Based Flower Detection and Counting in Highly Populated Images: A Peach Grove Case Study. *J. Agric. Food Res.* **2024**, *15*, 100930. [[CrossRef](#)]
28. Rahim, U.F.; Utsumi, T.; Mineno, H. Deep Learning-Based Accurate Grapevine Inflorescence and Flower Quantification in Unstructured Vineyard Images Acquired Using a Mobile Sensing Platform. *Comput. Electron. Agric.* **2022**, *198*, 107088. [[CrossRef](#)]
29. Xu, S.; Lu, J.; Wu, Y.; Liu, H.; Chen, F.; Zhang, F.; Chen, S.; Fang, W.; Guan, Z. Deep Learning-Based Approach for Extracting Inflorescence Morphology Features in Cut Chrysanthemum. *Smart Agric. Technol.* **2025**, *12*, 101265. [[CrossRef](#)]
30. Fatehi, F.; Bagherpour, H. Design, Construction, and Evaluation of an Autonomous Robot for Harvesting Damask Roses, Equipped with a Specialized End-Effector. *J. Agric. Food Res.* **2025**, *24*, 102432. [[CrossRef](#)]
31. Zhou, X.; Lee, W.S.; Ampatzidis, Y.; Chen, Y.; Peres, N.; Fraisse, C. Strawberry Maturity Classification from UAV and Near-Ground Imaging Using Deep Learning. *Smart Agric. Technol.* **2021**, *1*, 100001. [[CrossRef](#)]
32. Liu, Y.; Wang, G.; Sun, W.; Yang, S.; Feng, B.; Jia, S.; Wu, C. Honeysuckle Flower Stage Identification Based on Improved YOLOv5s. *Agron. J.* **2024**, *116*, 2511–2522. [[CrossRef](#)]
33. Xu, X.; Jin, J.; Xu, L.; Zheng, Z.; Zhang, Y.; Zhang, Z. A Semi-Supervised Method for Detecting Female Cucumber Flowers in Greenhouses Based on Unmanned Aerial Vehicle Images. *Comput. Electron. Agric.* **2025**, *239*, 111009. [[CrossRef](#)]
34. Zhao, F.; Ren, Z.; Wang, J.; Wu, Q.; Xi, D.; Shao, X.; Liu, Y.; Chen, Y.; Mizuno, K. Smart UAV-Assisted Rose Growth Monitoring with Improved YOLOv10 and Mamba Restoration Techniques. *Smart Agric. Technol.* **2025**, *10*, 100730. [[CrossRef](#)]
35. Zhang, P.; Li, X.; Yu, L.; Zhang, Z.; Dunkin, F.; Liu, H.; Li, Z. Boosting learning efficiency in few-shot tasks with layer-adaptive pid control. In *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence*; IEEE: New York, NY, USA, 2026.
36. Liu, H.; Li, H.; Wang, H.; Liu, C.; Qian, J.; Wang, Z.; Geng, C. Improved Detection and Location of Small Crop Organs by Fusing UAV Orthophoto Maps and Raw Images. *Remote Sens.* **2025**, *17*, 906. [[CrossRef](#)]
37. Yang, H.; Liu, Z.; Liu, W.; Wang, H.; Zhang, Y.; Wang, H. Graph-MDETR: A graph-guided Mamba-DETR network for UAV catenary support components detection in electrified railways. In *Proceedings of the IEEE Transactions on Intelligent Transportation Systems*; IEEE: New York, NY, USA, 2026.
38. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2016; pp. 779–788. [[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 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 16965–16974. [\[CrossRef\]](#)
40. Zhao, F.; Ma, B.; Xi, D.; Wang, J.; Chen, Y.; Liu, Y.; Shao, X.; Zhang, M.; Zhang, G.; Chen, J.; et al. Deep Learning-Based Super-Resolution Reconstruction and Improved YOLOv9 for Efficient Benthos Detection: A Case Study at Lake Hamana, Japan. *Remote Sens. Ecol. Conserv.* **2026**, *ahead of print*. [\[CrossRef\]](#)
41. 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. [\[CrossRef\]](#) [\[PubMed\]](#)
42. Hou, Y.; Zhang, L.; Wang, J.; Li, H.; Duan, X.; Huang, K.; Tian, Y. Efficientsarnet: A light-weight, high performance and fast sar image target recognition network. *IEEE Sens. J.* **2026**, *26*, 11413–11431. [\[CrossRef\]](#)
43. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242. [\[CrossRef\]](#)
44. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [\[CrossRef\]](#)
45. Duan, F.; Wang, H.; Yang, H.; Wei, C.; Zhang, C.; Song, Y.; Liu, Z. MRFM-IFCOS: An Anchor-Free Interactive Detector Based on Multireceptive Field Mamba for Detecting Catenary Support Components. *IEEE Trans. Instrum. Meas.* **2025**, *74*, 2553415. [\[CrossRef\]](#)
46. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [\[CrossRef\]](#)
47. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430. [\[CrossRef\]](#)
48. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976. [\[CrossRef\]](#)
49. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 7464–7475. [\[CrossRef\]](#)
50. Wang, C.Y.; Yeh, I.H.; Liao, H.Y.M. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Proceedings of the Computer Vision—ECCV 2024*; Springer: Cham, Switzerland, 2024; pp. 1–21.
51. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *arXiv* **2024**, arXiv:2405.14458. [\[CrossRef\]](#)
52. Zhao, F.; Xu, D.; Ren, Z.; Shao, X.; Wu, Q.; Liu, Y.; Wang, J.; Song, J.; Chen, Y.; Zhang, G.; et al. Mamba-Based Super-Resolution and Semi-Supervised YOLOv10 for Freshwater Mussel Detection Using Acoustic Video Camera: A Case Study at Lake Izunuma, Japan. *Ecol. Inform.* **2025**, *90*, 103324. [\[CrossRef\]](#)
53. 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. [\[CrossRef\]](#)
54. Liu, Z.; Wang, J.; Wu, H.; Xue, F.; Qin, Z.; Sun, S.; Guo, X.; Zhao, F. Water-Aware Real-Time Detection of Floating Plastic Debris via an Enhanced YOLOv13 Framework for Aquatic Pollution Monitoring. *Expert Syst. Appl.* **2026**, *313*, 131552. [\[CrossRef\]](#)
55. Vilcapoma, P.; Vásquez, I.N.; Prado, A.J.; Moya, V.; Vásquez, J.P. Drone-Based Marigold Flower Detection Using Convolutional Neural Networks. *Processes* **2025**, *13*, 3169. [\[CrossRef\]](#)
56. Tao, J.; Qiao, Q.; Song, J.; Sun, S.; Chen, Y.; Wu, Q.; Liu, Y.; Xue, F.; Wu, H.; Zhao, F. Deep Learning-Driven Automatic Segmentation of Weeds and Crops in UAV Imagery. *Sensors* **2025**, *25*, 6576. [\[CrossRef\]](#) [\[PubMed\]](#)
57. Zhao, F.; He, Y.; Song, J.; Wang, J.; Xi, D.; Shao, X.; Wu, Q.; Liu, Y.; Chen, Y.; Zhang, G.; et al. Smart UAV-Assisted Blueberry Maturity Monitoring with Mamba-Based Computer Vision. *Precis. Agric.* **2025**, *26*, 56. [\[CrossRef\]](#)
58. Bai, Y.; Yu, J.; Yang, S.; Ning, J. An Improved YOLO Algorithm for Detecting Flowers and Fruits on Strawberry Seedlings. *Biosyst. Eng.* **2024**, *237*, 1–12. [\[CrossRef\]](#)
59. Tao, J.; Zhao, F.; Chen, Y.; Liu, Y.; Xue, F.; Song, J.; Wu, H.; Chen, J.; Li, P.; Xu, N. Diffusion-Enhanced Underwater Debris Detection via Improved YOLOv12n Framework. *Remote Sens.* **2025**, *17*, 3910. [\[CrossRef\]](#)
60. Yuan, Z.; Gong, J.; Guo, B.; Wang, C.; Liao, N.; Song, J.; Wu, Q. Small Object Detection in UAV Remote Sensing Images Based on Intra-Group Multi-Scale Fusion Attention and Adaptive Weighted Feature Fusion Mechanism. *Remote Sens.* **2024**, *16*, 4265. [\[CrossRef\]](#)
61. Tian, Y.; Ye, Q.; Doermann, D. YOLOv12: Attention-Centric Real-Time Object Detectors. *arXiv* **2025**, arXiv:2502.12524. [\[CrossRef\]](#)
62. Nikouei, M.; Baroutian, B.; Nabavi, S.; Taraghi, F.; Aghaei, A.; Sajedi, A.; Moghaddam, M.E. Small Object Detection: A Comprehensive Survey on Challenges, Techniques and Real-World Applications. *Intell. Syst. With Appl.* **2025**, *27*, 200561. [\[CrossRef\]](#)
63. Wang, H.; Liu, J.; Zhao, J.; Zhang, J.; Zhao, D. Precision and Speed: LSOD-YOLO for Lightweight Small Object Detection. *Expert Syst. With Appl.* **2025**, *269*, 126440. [\[CrossRef\]](#)
64. Nghiem, V.Q.; Nguyen, H.H.; Hoang, M.S. LEAF-YOLO: Lightweight Edge-Real-Time Small Object Detection on Aerial Imagery. *Intell. Syst. With Appl.* **2025**, *25*, 200484. [\[CrossRef\]](#)

65. Dong, G.; Hong, Z.; Liu, S.; Wu, C.; Fang, Y.; Li, Z.; Zhang, X.; Liu, B.; Zhang, Y.; Ding, H.; et al. Dynamic Interaction-Aware and Causality-Disentangled Framework for Multimodal Sentiment Analysis. *arXiv* **2026**, arXiv:2605.30994.
66. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the Stars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2024; pp. 5694–5703.
67. Li, J.; Wu, H.; Qin, J. WeaveSeg: Iterative Contrast-weaving and Spectral Feature-refining for Nuclei Instance Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2025; pp. 21984–21993.
68. Jiang, Q.; Wang, Q.; Jin, X.; Wozniak, M.; Yao, S.; Zhou, W. MFmamba: A Multi-function Network for Panchromatic Image Resolution Restoration Based on State-Space Model. *Proc. AAAI Conf. Artif. Intell.* **2026**, *40*, 5406–5414. [\[CrossRef\]](#)
69. Wang, Y.; Lin, X.; Xiang, Z.; Su, W.H. VM-YOLO: YOLO with VMamba for Strawberry Flowers Detection. *Plants* **2025**, *14*, 468. [\[CrossRef\]](#) [\[PubMed\]](#)
70. Yang, G.; Jiang, Y.; Wang, S.; Chen, K. VinsFusion-Line: Binocular Vision-Inertial Navigation Real-Time SLAM System Based on Line Features. *Cybern. Syst.* **2026**, *57*, 350–375. [\[CrossRef\]](#)
71. Zhao, H.; Wang, B.; Fu, Y.; Zhou, M.; Li, N.; Gao, Z. Fixed-Time Adaptive Smooth Control with Relaxed Performance Constraints for a Canard Rotor/Wing UAV Subject to Aerodynamic Interferences and Actuator Faults. *Aerosp. Sci. Technol.* **2026**, *178*, 112959. [\[CrossRef\]](#)
72. Zhai, S.; Cheng, J.; Li, G.; Wang, Z.; Xing, Y.; Hou, M. Robust Auto-Landing Guidance and Control for Fixed-Wing Unmanned Aerial Vehicles with Strong Wind Disturbances. *Aerosp. Sci. Technol.* **2026**, *178*, 112413. [\[CrossRef\]](#)
73. Lv, H.; Li, C.; Dai, J.; Zhang, Y.; Fan, Z.; Tan, Y.; Wang, D.; Xie, B. Lightweight Framework for Underground Pipeline Recognition and Spatial Localization Based on Multiview 2-D GPR Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5110115. [\[CrossRef\]](#)
74. Gladson, V.B.; Deepa, G.; Jothi, S.F.E.; Mahalakshmi, L. Image Steganography with Security Using Massive Threefold Attentional Residual GAN Optimized by Chaotic PSO Algorithm. *Cybern. Syst.* **2025**, 1–32. [\[CrossRef\]](#)
75. Hou, Y.; Lv, D.; Liwang, M.; Wang, J.; Li, Z.; Xia, X.; Wu, D.; Tian, Y.; Quek, T.Q.S.; Wang, X. AsyncFed-SAR: Towards Communication-Efficient Asynchronous Federated Learning for SAR Image Target Recognition. In *Proceedings of the IEEE Transactions on Mobile Computing*; IEEE: New York, NY, USA, 2026.
76. Hou, Y.; Yu, B.; Yang, Z.; Wang, J.; Xiang, W.; Wu, D.; Liwang, M.; Xia, X.; Li, Z.; Tian, Y.; et al. Privacy-Preserving Federated SAR Image Target Recognition with Adaptive Resource Management in Space–Air–Ground Integrated Networks. *Pattern Recognit.* **2026**, *177*, 113253. [\[CrossRef\]](#)
77. Karaca, B.; Ünlü, K.D.; Türkan, S. Wavelet-Enhanced Sequence-to-Sequence Modeling with Attention Mechanism for Short-Term Wind Power Forecasting. *Cybern. Syst.* **2025**, *57*, 485–533. [\[CrossRef\]](#)
78. Yan, J.; Zhou, N.; Cheng, Y.; Zhang, F.; Wang, H.; Wang, M.; Jin, B.; Li, M.; Lu, Q.; Zhang, W. Application of Machine-Vision-Driven Physics-Informed Neural Networks in Pantograph–Catenary System State Detection. *Mech. Syst. Signal Process.* **2026**, *257*, 114577. [\[CrossRef\]](#)

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.