agriculture

MDPI

Article

DDF-DETR: A Multi-Scale Spatial Context Method for Field Cotton Seedling Detection

Feng Xu [1,2,3], Huade Zhou [1,2,3], Yinyi Pan [1,2,3], Yi Lu [1,2,3] and Luan Dong [1,2,3,*]

[1] College of Computer and Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China
[2] Engineering Research Center of Intelligent Agriculture Ministry of Education, Urumqi 830052, China
[3] Xinjiang Agricultural Informatization Engineering Technology Research Center, Urumqi 830052, China
* Correspondence: dl@xjau.edu.cn

Abstract

Accurate assessment of cotton emergence rates is essential for precision agriculture management, and unmanned aerial vehicle (UAV) imagery provides a scalable means for field-level monitoring. However, cotton seedling detection from UAV images faces persistent challenges: individual seedlings appear as small targets with diverse morphologies across varying flight altitudes; strong plastic film reflections, weeds, and soil cracks introduce substantial background interference; and "missing seedling" targets, which manifest as negative space features, exhibit high similarity to background noise. Existing CNN–Transformer hybrid detection architectures are limited by fixed convolutional receptive fields that cannot adapt to multi-scale target variations, attention mechanisms that lack explicit directional geometric modeling, and interpolation-based upsampling that attenuates high-frequency edge details of small targets. To address these issues, this paper proposes DDF-DETR (Dynamic-Direction-Frequency Detection Transformer), a multi-scale spatial context detection method based on RT-DETR. The method incorporates three components: a Dynamic Gated Mixer Block (DGMB) for adaptive multi-scale feature extraction with background noise suppression, a Direction-Aware Adaptive Transformer Encoder (DAATE) for directional geometric feature modeling at linear computational complexity, and a Frequency-Aware Sub-pixel Upsampling Network (FASN) for high-frequency detail recovery in the feature pyramid. On the self-constructed Xinjiang cotton field dataset, DDF-DETR achieves 83.72% mAP@0.5 and 63.46% mAP@0.5:0.95, representing improvements of 2.38% and 5.28% over the baseline RT-DETR-R18, while reducing the parameter count by 30.6% and computational cost to 42.8 GFLOPs. Generalization experiments on the VisDrone2019 and TinyPerson datasets further validate the robustness of the proposed method for small target detection across different scenarios.

Keywords: cotton seedling detection; object detection; UAV remote sensing; Transformer; multi-scale feature fusion; small object detection

Check for updates

Received: 31 January 2026
Revised: 28 February 2026
Accepted: 4 March 2026
Published: 7 March 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 Attribution (CC BY) license.

1. Introduction

Cotton is one of the major economic crops in China. Xinjiang, as the largest cotton-producing region in China, accounts for over 90% of the national production [1]. The seedling stage is a critical period in cotton growth and development, with emergence quality directly affecting subsequent yield formation and fiber quality [2]. Traditional cotton seedling monitoring relies on manual field surveys, which suffer from low efficiency

and subjective bias, failing to meet the demands of large-scale precision agriculture management [3]. Therefore, developing rapid and accurate automated cotton seedling detection technology is valuable for precision agriculture applications.

UAV remote sensing has gained increasing attention in agricultural monitoring due to its low cost, operational flexibility, and high spatiotemporal resolution [4,5]. Nevertheless, aerial image acquisition inherently suffers from quality degradation challenges such as complex background interference and low signal-to-noise ratio [6], and deep learning-based methods integrating multi-source remote sensing data have demonstrated effectiveness in handling small-scale crop targets under such complex conditions [7]. UAV-based crop monitoring technology has been applied to maize seedling detection [8], vegetation species identification [9], and sorghum seedling counting [10]. However, cotton seedling detection faces the following challenges:

1. **Weak target features**: Individual cotton seedlings are small with diverse morphologies, appearing as small objects in UAV images. The contrast between cotton seedlings and weeds or soil background is low, and field environments contain interference factors such as illumination variations, vegetation occlusion, and plastic film reflections [11,12].
2. **Strong negative sample interference**: "Missing seedling" targets lack salient object-level visual features, primarily manifesting as local texture absence or negative space features. These targets are prone to confusion with background features such as shadows from plastic film wrinkles, natural film damage, and soil cracks, leading to false positive detections.

Object detection is one of the core tasks in computer vision. Deep learning-based object detection methods can be categorized into two-stage detectors and single-stage detectors [13]. Two-stage detectors, represented by the R-CNN series including R-CNN [14], Fast R-CNN [15], and Faster R-CNN [16], first generate region proposals and then perform classification and regression on these proposals. These methods have been applied to agricultural pest detection [17] and crop seedling localization [18]. Single-stage detectors, represented by the YOLO series [19] and SSD [20], adopt an end-to-end detection framework that directly predicts target categories and locations. The YOLO series has undergone multiple iterations, evolving from YOLOv1 to YOLOv8 [21], YOLOv10 [22], and YOLOv26 [23], with applications in weed detection [24], pest identification [25], and crop disease diagnosis [26].

Following its success in natural language processing, the Transformer architecture was introduced to computer vision tasks. DETR (Detection Transformer), the first model to apply the Transformer to object detection, adopts end-to-end learning and eliminates hand-designed components such as Non-Maximum Suppression (NMS) in traditional detectors [27]. Subsequently, researchers have proposed a series of improvements: Deformable DETR [28] accelerates convergence through deformable attention mechanisms; DAB-DETR [29] introduces dynamic anchor box queries; DINO [30] improves detection performance through contrastive denoising training. RT-DETR (Real-Time Detection Transformer) achieves a balance between speed and accuracy through an efficient hybrid encoder design [31]. RT-DETR-based improvements have been applied to agricultural scenarios including poultry detection [32], fruit growth monitoring [33], and crop disease identification [34].

Although the Transformer architecture can model global dependencies, its computational complexity scales quadratically with input sequence length, and it lacks explicit modeling of local spatial structures [35,36]. Consequently, researchers have explored hybrid CNN-Transformer architectures, aiming to combine the local feature extraction capability of CNNs with the global context modeling capability of Transformers. Such hybrid archi-

tectures typically employ CNN backbone networks to extract multi-scale feature maps, followed by Transformer encoders for global feature interaction, thereby enhancing model representation capability while maintaining computational efficiency. RT-DETR, as a representative hybrid architecture, extracts hierarchical features through ResNet or HGNet, which are fused via the hybrid encoder before being sent to the Transformer decoder. In agricultural remote sensing, hybrid architectures have been applied to crop hyperspectral classification [37], plant disease detection [38], and fruit growth monitoring, demonstrating superior performance compared to single CNN or Transformer architectures. However, existing hybrid architectures still exhibit notable limitations when processing tiny dense targets in UAV imagery. The fixed receptive field of standard CNN branches constrains multi-scale adaptation, as small convolutional kernels cover only a fraction of the theoretical receptive field and fail to accommodate the scale variations caused by different flight altitudes [39]. Standard self-attention mechanisms rely on isotropic dot-product similarity, which lacks explicit directional geometric modeling and leads to degraded performance on elongated or oriented objects [40]. Interpolation-based upsampling in feature pyramids, particularly nearest-neighbor interpolation, reconstructs spatial resolution through simple pixel replication without recovering high-frequency edge and texture details [41]. Content-aware upsampling methods such as CARAFE [42] and DySample [43] have partially alleviated this issue, yet their isotropic sampling patterns remain limited in recovering directional high-frequency information for anisotropic small targets.

To address the aforementioned problems, this paper proposes DDF-DETR (Dynamic-Direction-Frequency Detection Transformer), a multi-scale spatial context detection network for cotton seedling detection. The main contributions of this paper include:

- **A Dynamic Gated Mixer Block (DGMB) is designed**: Through the fusion of multi-branch dynamic convolution and gating mechanisms with local convolutional features, adaptive feature extraction for the anisotropic geometric morphology of cotton seedlings is achieved. The gating mechanism suppresses interference from background noise such as plastic film reflections and shadows on missing seedling detection.
- **A Direction-Aware Adaptive Transformer Encoder (DAATE) is proposed**: Through polarized linear attention mechanisms and spatial-aware feed-forward networks, the perception capability of the encoder for cotton seedling targets in complex backgrounds is enhanced, and contextual modeling assists in distinguishing missing seedling holes from background cracks.
- **A Frequency-Aware Sub-pixel Upsampling Network (FASN) is designed**: Through learnable sub-pixel convolution and direction-aware spatial shift mechanisms, high-frequency edge and texture information of small targets is recovered in the feature pyramid upsampling path, facilitating distinction between sharp-edged missing seedling holes and blurred plastic film shadows.

The remainder of this paper is organized as follows: Section 2 introduces the construction and annotation process of the cotton seedling detection dataset; Section 3 describes the network architecture and key module designs of DDF-DETR; Section 4 presents the experimental setup and result analysis; Section 5 concludes the paper.

2. Dataset Collection and Construction

2.1. Study Area and Collection Conditions

Cotton seedling images for this study were collected at Huaxing Farm, Daxiqu Township, Changji City, Xinjiang Uygur Autonomous Region, China (87°29′ E, 44°22′ N), a region with a temperate continental climate. The cotton variety planted was Zhongmian 113, which is cultivated using a "one film, three belts, and six rows" planting pattern with 2.05 m wide plastic mulch. Under each film, three paired planting belts were arranged. The

narrow row spacing within each pair was 10 cm, and the center-to-center distance between adjacent belts was 66 cm. The plant spacing within the row was 10 cm. The planting pattern schematic is illustrated in Figure 1. Data collection was conducted from May to June 2025, during the critical cotton seedling stage.

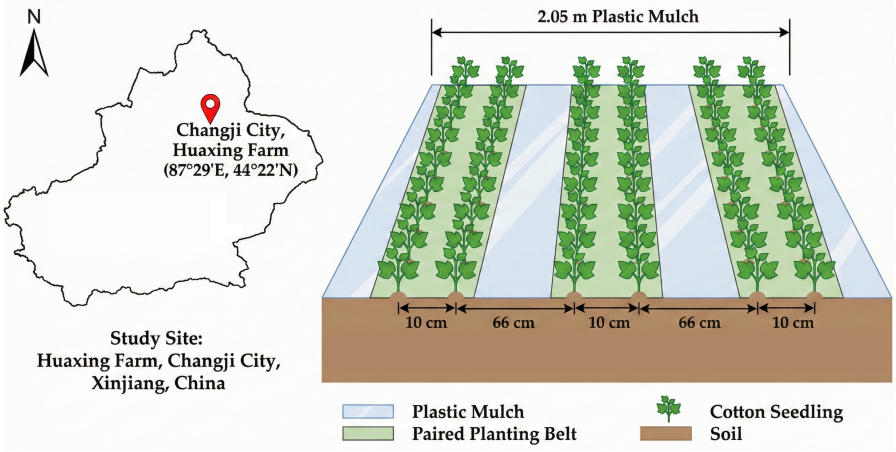


Figure 1. Schematic diagram of the "one film, three belts, and six rows" planting pattern for cotton cultivation.

2.2. Image Acquisition Equipment and Parameters

Image acquisition employed a DJI Mavic 3 Pro quadrotor UAV platform (DJI, Shenzhen, China) equipped with a Hasselblad L2D-20C aerial camera system (Hasselblad, Gothenburg, Sweden). The main camera features a 4/3-inch CMOS image sensor with 20 million effective pixels, 24 mm equivalent focal length, 84° field of view, f/2.8–f/11 aperture range, 12-bit RAW format support, and native dynamic range of 12.8 stops. The camera shutter speed ranges from 8 to 1/8000 s, with ISO sensitivity from 100 to 6400. The UAV positioning system employs GPS+GLONASS+Galileo multi-constellation combined navigation, coupled with a downward-facing binocular vision system, achieving hovering accuracy of $\pm$0.1 m vertical and $\pm$0.3 m horizontal. The original image resolution was 5280 $\times$ 3956 pixels.

To capture multi-scale cotton seedling features, the UAV acquired images at four altitudes: 15 m (high-altitude overview), 5 m (medium-altitude view), 3 m (low-altitude detail), and 1.5 m (close-range texture). The flight speed was set to 1 m/s with a capture interval of 0.5 s, resulting in a flight distance of 0.5 m between adjacent images. Gimbal attitude parameters were configured as follows: in vertical shooting mode, pitch angle was $-90°$ (vertical downward), yaw angle averaged 5.03°, and roll angle averaged 0.10°; in oblique shooting mode, pitch angle was $-45°$, yaw angle averaged 4.87°, and roll angle averaged 0.15°.

2.3. Image Annotation and Dataset Composition

Annotation was performed using the LabelImg tool in PASCAL VOC format (https://github.com/HumanSignal/labelImg (accessed on 15 January 2026)). The dataset comprises two categories: "seedling" (cotton) and "missing seedling" (no seeding). Based on the above planting pattern, annotators identified "missing seedling" targets by reconstructing the expected grid positions from visible seedlings and labeling a dark hole as "missing seedling" only when it appeared within $\pm$5 pixels of an expected grid node with approximately circular morphology; film tears and soil cracks at irregular positions were excluded as background.

To validate model robustness in real-world complex agricultural environments, the dataset specifically incorporated multiple typical field interference scenarios, including:

- **Biological interference**: Field weeds with colors similar to cotton seedlings, whose green features can cause false positive detections;
- **Non-biological interference**: Damaged areas on plastic film, reflections, and irregularly shaped soil cracks and clods. These background noises exhibit similarity to the "missing seedling" category in morphology and grayscale characteristics, representing the primary causes of misclassifying background as missing seedlings.

By incorporating these challenging negative samples, the dataset enables comprehensive evaluation of model feature extraction and discrimination capabilities in complex backgrounds. After screening and annotation processing, the dataset comprises 3372 images and 416,869 annotated instances. The dataset was randomly split at the image level in a 7:1.5:1.5 ratio. To prevent images from the same field plot appearing in both training and test sets, stratified sampling was performed at the plot level. Dataset split details are shown in Table 1.

Table 1. Statistics of the Xinjiang cotton field seedling dataset.

Dataset Split	Number of Images	Annotated Instances	Percentage (%)
Training Set	2360	291,808	70
Validation Set	506	62,530	15
Test Set	506	62,531	15
Total	3372	416,869	100

The average number of targets per image is 123.6, reflecting the dense distribution of targets in cotton field environments. To visualize dataset characteristics, Figure 2 presents typical samples and anchor box dimension distribution analysis. Figure 2a displays cotton field images at different capture altitudes and angles, with backgrounds containing complex factors such as plastic film reflections, exposed soil, and weed interference. The box plot in Figure 2b reveals the width-height distribution of annotated targets, with a mean target width of 13.1 pixels (standard deviation 9.5 pixels) and a mean height of 19.3 pixels (standard deviation 13.1 pixels). This indicates that cotton seedling targets generally exhibit tiny and elongated morphological characteristics, imposing high requirements on detector's small object perception capabilities.

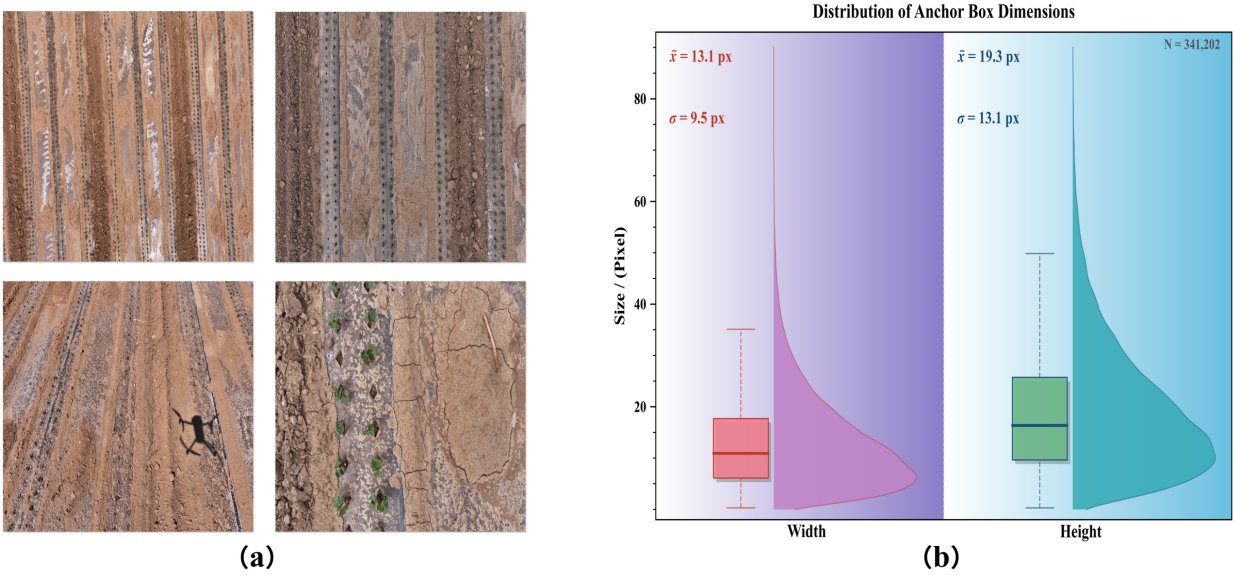


Figure 2. Typical samples and anchor box dimension distribution of the self-constructed Xinjiang cotton field seedling dataset. (**a**) Sample images; (**b**) Anchor box distribution.

To comprehensively evaluate detection model adaptability in real field environments, the dataset encompasses multiple challenging scenarios during collection (see Figure 3). Regarding illumination conditions, the dataset includes strong plastic film reflection scenarios where the film surface produces high-brightness regions under direct sunlight with grayscale values similar to cotton seedling leaves, which can cause false detections. It also includes partial shadow scenarios caused by cloud occlusion, creating contrast differences between light and dark regions within the same image. Regarding biological interference, the dataset incorporates samples of cotton seedlings intermixed with field weeds that exhibit similar color and texture features, challenging model category discrimination capability. Regarding non-biological interference, the dataset includes plastic film defects such as damage and wrinkles, as well as irregularly shaped soil cracks and clods, whose edge features can be confused with cotton seedling contours. In images captured at high altitude (15 m), some cotton seedlings have pixel areas smaller than 32×32, falling within the small object detection domain. The incorporation of these challenging scenarios enables the dataset to support robustness validation of models in complex agricultural environments.

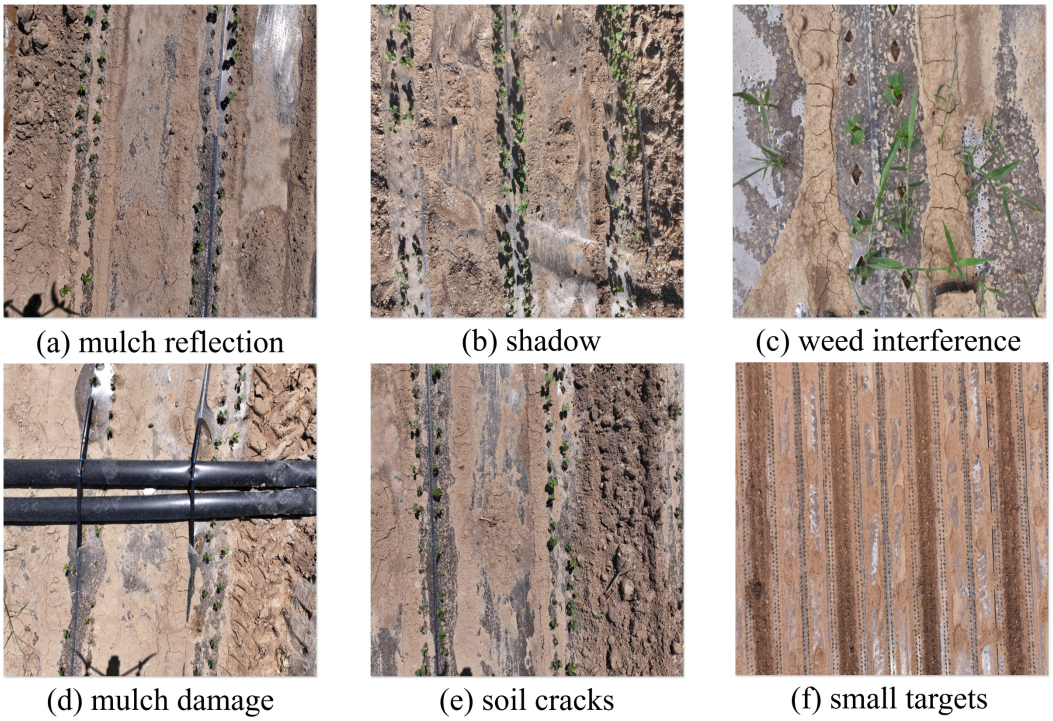


Figure 3. Dataset challenging scenario examples: (**a**) strong plastic film reflection; (**b**) illumination shadow; (**c**) weed interference; (**d**) plastic film damage; (**e**) soil crack; (**f**) small object.

3. Overall Architecture Design

3.1. RT-DETR Baseline Network Architecture

RT-DETR (Real-Time Detection Transformer) is a real-time end-to-end object detector proposed by the Baidu PaddlePaddle team in 2023. The RT-DETR network comprises three components: backbone network, hybrid encoder, and Transformer decoder. The backbone network typically employs ResNet or HGNet series to extract image feature representations. The hybrid encoder contains an Attention-based Intra-scale Feature Interaction (AIFI) module and a CNN-based Cross-scale Feature Fusion Module (CCFM), which decouple multi-scale feature interaction into two steps—intra-scale interaction and cross-scale fusion—progressively improving model accuracy while reducing computational cost. The Transformer decoder incorporates IoU-aware query selection mechanisms and detection heads, eliminating the Non-Maximum Suppression (NMS) post-processing pro-

cedure to improve inference speed. By integrating the local perception capability of CNNs with the global modeling capability of Transformers, RT-DETR achieves a balance between efficiency and accuracy, making it the preferred baseline network for tiny target detection in cotton fields in this study.

3.2. Improved DDF-DETR Model

To address feature degradation and spatial detail loss issues encountered by RT-DETR when detecting tiny cotton seedling targets in complex agricultural scenarios, this study proposes DDF-DETR (Dynamic-Direction-Frequency Detection Transformer). The overall architecture of this model is illustrated in Figure 4, incorporating three main improvements: Dynamic Gated Mixer Block (DGMB), Direction-Aware Adaptive Transformer Encoder (DAATE), and Frequency-Aware Sub-pixel Upsampling Network (FASN).

In the backbone network, DGMB modules replace the original BasicBlock structures of ResNet. In the neck network, DGMB similarly replaces the original RepC3 modules, achieving full-chain dynamic feature extraction. This module enables adaptive feature extraction for multi-scale cotton seedling morphologies through multi-branch dynamic convolution and gating mechanisms. In the hybrid encoder, DAATE modules replace the original AIFI modules, enhancing directional geometric feature representation through polarized linear attention mechanisms. In the upsampling path of the feature pyramid, FASN modules replace nearest neighbor interpolation, recovering high-frequency detail information of small targets through sub-pixel convolution.

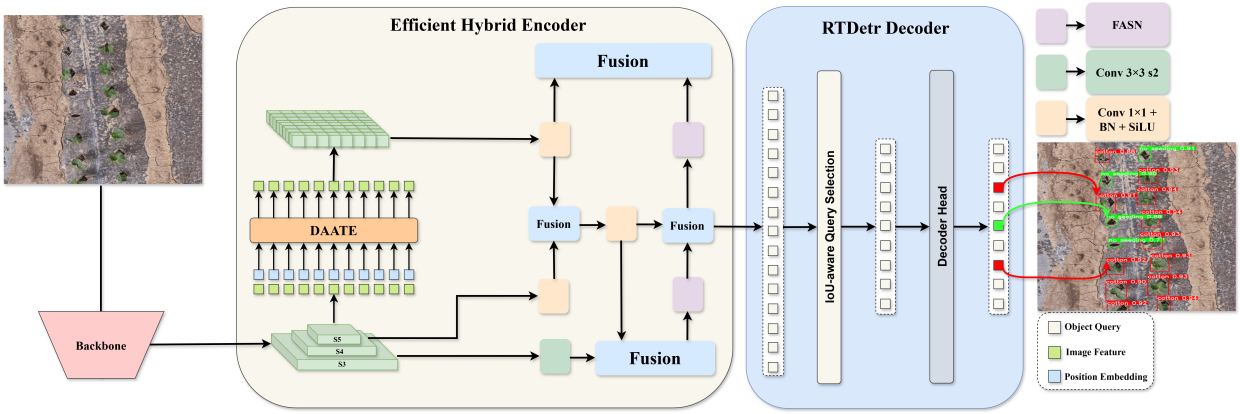


Figure 4. DDF-DETR model overall architecture diagram.

3.3. DGMB Module

The backbone network of RT-DETR-R18 employs a ResNet-style BasicBlock for feature extraction, with the detection head based on RepC3 modules. In field cotton seedling detection, the fixed 3×3 convolutional kernel of BasicBlock results in limited receptive fields that struggle to adapt to seedling scale variations caused by UAV multi-altitude acquisition. Its static computation pattern also causes resource redundancy in simple background regions. Additionally, RepC3 lacks adaptive fusion capability for cross-scale features. Existing alternatives each address only part of this problem. Large-kernel re-parameterization methods such as UniRepLKNet [44] expand the receptive field with isotropic square kernels that do not model the directional extension of elongated leaves. DCNv2 [45] learns data-driven sampling offsets but lacks structured geometric priors. Lightweight context-guided approaches [46,47] reduce model size at the cost of small-target accuracy. To jointly address receptive field adaptation, geometric modeling, and background suppression, this study proposes DGMB (Dynamic Gated Mixer Block), whose structure is illustrated in Figure 5. It introduces dynamic multi-scale convolution and adaptive channel modulation to reduce

computational complexity while enhancing the capability of the model to capture multi-scale seedling features. The Gated Linear Unit (GLU) mechanism in DGMB adaptively regulates information flow based on input feature content, suppressing response channels to background textures such as plastic film reflections and irregular cracks, thereby reducing background noise interference on missing seedling detection.

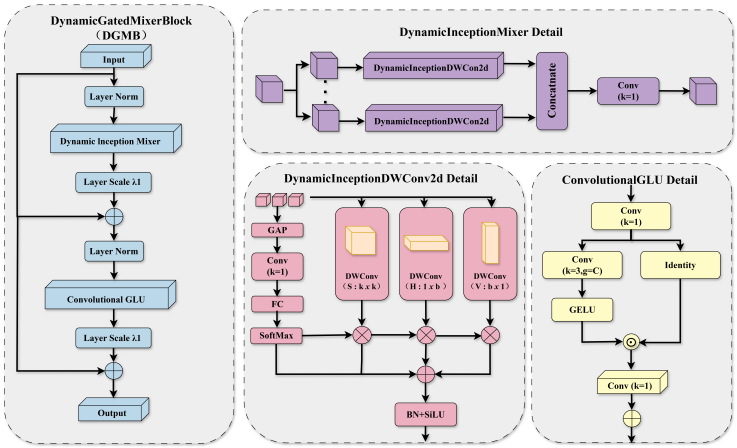


Figure 5. DGMB module architecture diagram.

DGMB follows the advanced MetaFormer architecture paradigm and is designed to enhance feature representation capability by decoupling spatial mixing and channel mixing processes. This module consists of two cascaded residual sub-modules: the front-end DynamicInceptionMixer captures long-range spatial dependencies, while the back-end ConvolutionalGLU focuses on refined channel feature selection. To accelerate deep network convergence and maintain the stability of training dynamics, learnable layer scaling factors are introduced before the residual connection of each substructure. Assuming the input feature tensor is X, the computational flow of the entire block can be formally described as:

$$Z = X + \lambda_1 \odot \mathcal{F}_{\mathrm{DIM}}(\mathrm{LN}(X)) \tag{1}$$

$$Y = Z + \lambda_2 \odot \mathcal{F}_{\mathrm{GLU}}(\mathrm{LN}(Z)) \tag{2}$$

where X and Y represent the input and output feature tensors, respectively; Z denotes the intermediate feature state; $\mathrm{LN}(\cdot)$ represents Layer Normalization for stabilizing feature distribution; $\mathcal{F}_{\mathrm{DIM}}$ and $\mathcal{F}_{\mathrm{GLU}}$ denote the DynamicInceptionMixer and ConvolutionalGLU transformation functions, respectively; $\lambda_1, \lambda_2 \in \mathbb{R}^C$ are learnable layer scaling parameter vectors, typically initialized to small values (e.g., 10^{-2}); $\odot$ represents element-wise Hadamard product.

3.3.1. DynamicInceptionMixer Workflow and Principles

DynamicInceptionMixer ($\mathcal{F}_{\mathrm{DIM}}$) adopts a multi-branch parallel processing strategy to adapt to multi-scale features. Input features are split along the channel dimension into G subgroups, each independently processed through dynamic depthwise convolutions with different kernel scales, and finally fused through linear projection. The mathematical principles are as follows:

$$\mathcal{F}_{\mathrm{DIM}}(X) = W_{\mathrm{proj}} \circledast \left(\|_{g=1}^{G} \mathrm{dw}_{(g)}^{\Phi}(X_g) \right) \tag{3}$$

where $\|_{g=1}^{G}$ denotes concatenation across the G groups along the channel dimension before projection; G is the number of groups (corresponding to the length of the kernels list in

implementation); X_g represents the feature subset of the g-th group; $\mathrm{dw}_{(g)}^{\Phi}$ represents the dynamic depthwise convolution operation with specific kernel scale on the g-th branch; W_{proj} is the weight matrix of 1×1 convolution for inter-channel information interaction; $\circledast$ represents standard convolution operation.

ConvolutionalGLU ($\mathcal{F}_{\mathrm{GLU}}$) introduces gating mechanisms to enhance nonlinear expressiveness. The input is first projected and split into "content stream" and "gate stream", where the content stream encodes contextual information through depthwise convolution and is then modulated by the gate stream:

$$\mathcal{F}_{\mathrm{GLU}}(X) = W_2 \circledast \left((\mathrm{DWConv}(X_{\mathrm{content}})) \odot X_{\mathrm{gate}} \right), \quad \text{where } [X_{\mathrm{content}}, X_{\mathrm{gate}}] = \mathrm{Split}(W_1 \circledast X) \tag{4}$$

where W_1 and W_2 are convolution weights for input projection and output projection, respectively; $\mathrm{Split}(\cdot)$ operation halves and splits the feature channels; X_{content} is the main stream carrying feature information; X_{gate} is the gating stream controlling information flow; DWConv represents 3×3 depthwise convolution; $\odot$ implements content-based attention weighting.

3.3.2. DynamicInceptionDWConv2d Workflow and Principles

DynamicInceptionDWConv2d is the core component for handling the anisotropic morphology of cotton seedlings in this algorithm. It constructs an orthogonal convolution group containing square kernels ($S : k \times k$), horizontal strip kernels ($H : 1 \times b$), and vertical strip kernels ($V : b \times 1$). To achieve adaptive receptive fields, this module first generates dynamic weights through an attention sub-network. Input feature X, after global average pooling compression, is mapped through two fully connected layers to generate a normalized weight vector α for three branches:

$$\alpha = \mathrm{Softmax}(\mathrm{Reshape}(W_{\mathrm{fc2}} \cdot \delta(W_{\mathrm{fc1}} \cdot \mathrm{GAP}(X)))) \tag{5}$$

where $\alpha = \{\alpha_S, \alpha_H, \alpha_V\}$ are channel-wise weight vectors corresponding to three branches, satisfying $\sum_j \alpha_j^{(c)} = 1$ for each channel c; $\mathrm{GAP}(\cdot)$ represents Global Average Pooling, which compresses the input feature $X \in \mathbb{R}^{C \times H \times W}$ into a channel descriptor $\mathbf{z} \in \mathbb{R}^C$; $W_{\mathrm{fc1}} \in \mathbb{R}^{3C \times C}$ is the weight matrix of the first fully connected layer, mapping the C-dimensional descriptor to a $3C$-dimensional hidden representation with an expansion ratio of $3\times$; δ denotes the ReLU activation function; $W_{\mathrm{fc2}} \in \mathbb{R}^{3C \times 3C}$ and the subsequent Reshape operation reorganize the $3C$-dimensional output into three groups of C-dimensional vectors, each assigned to one convolution branch; Softmax normalization is applied along the branch dimension to produce channel-wise weight vectors $\alpha_j \in \mathbb{R}^C$ ($j \in \{S, H, V\}$), enabling per-channel directional selectivity.

After obtaining weights, the module performs weighted aggregation of the three orthogonal branch outputs. Since α is computed independently for each input sample, the kernel emphasis shifts adaptively toward vertical, horizontal, or isotropic patterns in response to the actual morphology present in each input, rather than applying a fixed weighting after training. The final output Y_{dyn} is computed as:

$$Y_{\mathrm{dyn}} = \sigma\left(\mathrm{BN}\left(\sum_{j \in \{S,H,V\}} \alpha_j \cdot \mathrm{dw}_{(j)}^{W} \circledast X \right) \right) \tag{6}$$

where j iterates over three convolution kernel types $\{S, H, V\}$; $\mathrm{dw}_{(j)}^{W}$ represents depthwise convolution kernel weights of corresponding geometric shapes; $\alpha_j \in \mathbb{R}^C$ is the generated channel-wise dynamic weight vector, applied to each channel of the corresponding branch output via element-wise multiplication; $\sum$ represents element-wise addition fusion of

multi-branch features; BN denotes Batch Normalization; σ is SiLU activation function for introducing nonlinearity.

By deploying DGMB in the backbone and neck networks, the improved algorithm constructs a feature extraction framework with geometric adaptability. This framework leverages the synergy between dynamic kernel selection mechanisms and multi-scale mixing strategies to enhance model robustness to cotton seedling morphological diversity in complex cotton field environments. This design decouples horizontal and vertical texture features of leaves in the spatial dimension and suppresses interference from background noise such as plastic film reflections through gating mechanisms. By ensuring feature signal-to-noise ratio and discriminability at different resolutions, it provides more discriminative feature representations for subsequent detection heads.

3.4. Direction-Aware Adaptive Transformer Encoder (DAATE)

The AIFI module in the RT-DETR hybrid encoder employs standard multi-head self-attention, whose computational complexity scales quadratically with the number of spatial tokens. When processing cotton seedling features, this design has three limitations: scalar dot-product similarity lacks explicit two-dimensional geometric modeling for anisotropic leaf structures; softmax normalization constrains attention weights to non-negative values, preventing the encoding of suppressive relationships between spatial regions; and position-independent feed-forward networks discard local neighborhood structure. Existing linear attention methods reduce complexity to $O(N)$ through kernel decomposition, yet simple kernel functions such as ELU-based mappings lack the capacity to represent complex spatial relationships in visual features. To address these issues, this study designs the Direction-Aware Adaptive Transformer Encoder (DAATE), whose structure is illustrated in Figure 6. Through the synergy of polarized linear attention and spatial-aware feed-forward networks, DAATE models directional geometric features and complex backgrounds while maintaining linear computational complexity. DAATE leverages the long-range dependency modeling capability of the Transformer architecture to capture global contextual information such as overall row-column planting patterns in cotton fields, assisting the model in determining whether local black spot features conform to planting patterns, thereby distinguishing genuine missing seedling holes from randomly occurring background shadows.

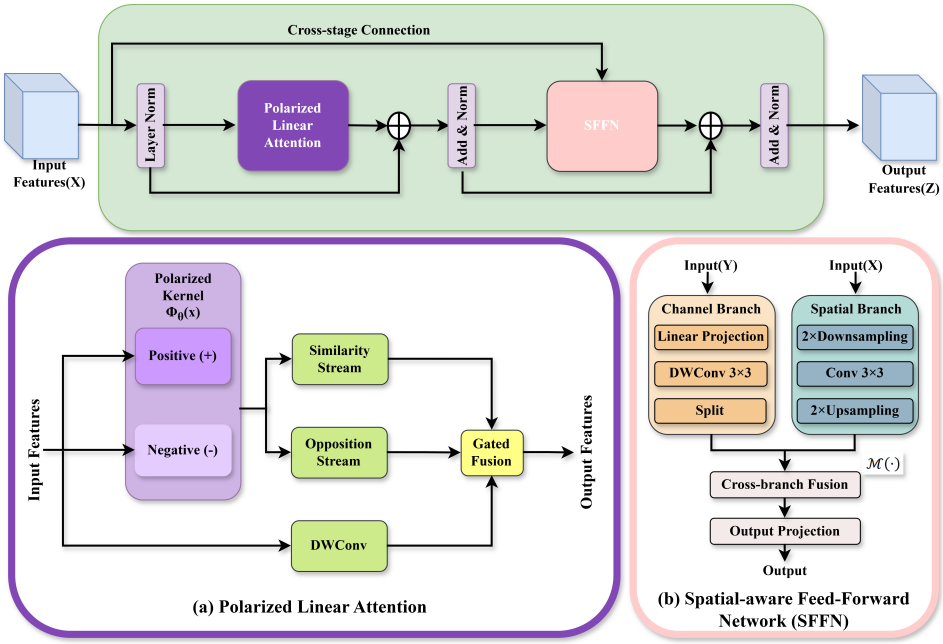


Figure 6. DAATE module architecture diagram.

3.4.1. Polarized Linear Attention Mechanism

The computational bottleneck of standard self-attention stems from the explicit construction of the attention matrix. Linear attention methods approximate softmax similarity functions through kernel decomposition, reducing complexity from $O(N^2)$ to $O(N)$. However, simple kernel functions lack sufficient expressive power to capture complex spatial relationships of visual features. This study introduces polarized kernel functions that decompose feature representations into positive and negative components, enabling the model to simultaneously capture excitatory and inhibitory relationships between spatial positions. Directional sensitivity is further achieved through an asymmetric positional encoding scheme: a learnable absolute positional encoding $\mathbf{PE} \in \mathbb{R}^{1\times N\times d}$ is applied exclusively to the key vectors K prior to the kernel transformation, while the query vectors Q remain unmodified, avoiding explicit polar-coordinate mapping. The polarized kernel transformation is defined as:

$$\phi(x) = [\mathrm{ReLU}(x)]^{\beta} \cdot \mathrm{softplus}(\gamma) \oplus [\mathrm{ReLU}(-x)]^{\beta} \cdot \mathrm{softplus}(\gamma) \tag{7}$$

where $\gamma \in \mathbb{R}^C$ is a learnable channel-level scaling parameter initialized to 1; $\omega \in \mathbb{R}^H$ is a learnable head-level modulation parameter, with H being the number of attention heads; $\beta = 1 + \bar{\gamma} \cdot \mathrm{sigmoid}(\omega)$ is the head-level exponent coefficient, where $\bar{\gamma}$ denotes the mean of γ across channels; $\oplus$ denotes feature dimension concatenation. Polarized decomposition explicitly separates features into positive and negative responses, enabling the model to represent signed similarities between spatial positions and allowing the encoding of inhibitory relationships. Based on polarized kernels, a similarity stream and an opposition stream—two complementary attention flows—are constructed: the similarity stream enhances intra-target feature aggregation through matching polarized patterns, while the opposition stream promotes target boundary discrimination through complementary polarity pairing. The linear attention operator reduces computational complexity from $O(N^2D)$ to $O(ND^2)$, while introducing a parallel local enhancement branch through depthwise convolution to compensate for local spatial structures potentially lost in kernel approximation. Global and local features are finally integrated through gating mechanisms. Under this asymmetric formulation, the attention score between positions i and j is expressed as:

$$A(i,j) = \phi(\mathbf{q}_i)^{\top} \cdot \phi(\mathbf{k}_j + \mathbf{pe}_j) \tag{8}$$

where $\mathbf{pe}_j$ denotes the positional embedding of key position j, and its interaction with query content $\mathbf{q}_i$ produces position-content cross-terms that guide the encoder to preferentially aggregate features along the dominant row-column spatial directions of cotton seedling planting patterns.

3.4.2. Spatial-Aware Feed-Forward Network

Standard Transformer feed-forward networks perform identical transformations independently at each spatial position, discarding spatial neighborhood relationships required for visual understanding. Although features processed by attention mechanisms are enriched in semantic content, fine-grained spatial structures are diluted through weighted averaging. This study designs a Spatial-aware Feed-Forward Network (SFFN) that employs a dual-branch architecture to process channel expansion and spatial context extraction in parallel. The channel branch projects features to an expanded hidden space and injects local spatial dependencies through depthwise convolution; the spatial branch extracts

multi-scale contextual information from cross-stage feature references (feature states before attention processing):

$$\mathrm{SFFN}(Y, X) = W_{\mathrm{out}} \cdot \mathcal{F}_{\mathrm{fuse}}(\mathrm{DWConv}(W_{\mathrm{in}} \cdot Y), \mathcal{M}(X)) \quad (9)$$

where Y represents post-attention features, X represents pre-attention feature references; $\mathcal{M}(\cdot)$ is a multi-scale spatial processing operator that expands the effective receptive field through downsampling-convolution-upsampling paths; $\mathcal{F}_{\mathrm{fuse}}$ is the cross-branch fusion operation. The theoretical basis of the cross-stage design lies in the fact that pre-attention features, without global mixing, retain spatial locality and geometric structure clarity inherited from convolutional backbone networks, providing spatially coherent reference supplements for semantically rich but spatially mixed attention outputs.

Through the synergy of these mechanisms, DAATE constructs a Transformer encoder optimized for cotton seedling morphological features: polarized linear attention reduces computational complexity to a linear level, with the dual-stream design achieving synergy between intra-target aggregation and boundary discrimination; spatial-aware feed-forward networks preserve spatial structural information attenuated during attention mixing through cross-stage feature references. These mechanisms collectively mitigate feature degradation issues faced by standard attention mechanisms when processing unstructured agricultural scenarios, enhancing detection capability for densely distributed tiny targets while maintaining computational efficiency.

3.5. Frequency-Aware Sub-Pixel Upsampling Network (FASN)

In the top-down path of feature pyramid networks, upsampling operations recover spatial resolution for multi-scale feature alignment. The RT-DETR baseline employs nearest-neighbor interpolation, which acts as a zero-order hold filter and attenuates high-frequency components corresponding to seedling edge contours and textures. In the spatial domain, it replicates a single source pixel to an $s \times s$ block, producing blocky artifacts that blur boundaries between adjacent targets. Content-aware reassembly methods such as CARAFE rely on large receptive field neighborhoods that can confuse features of adjacent individuals in dense scenarios, while dynamic point sampling methods such as DySample adopt isotropic sampling patterns that do not account for directional leaf textures. To overcome these limitations, this study proposes the Frequency-Aware Sub-pixel Upsampling Network (FASN), as illustrated in Figure 7. This network achieves frequency-aware feature reconstruction through learnable sub-pixel convolution and direction-aware spatial mechanisms, recovering high-frequency edge and texture information of small targets and improving model discrimination of detail features.

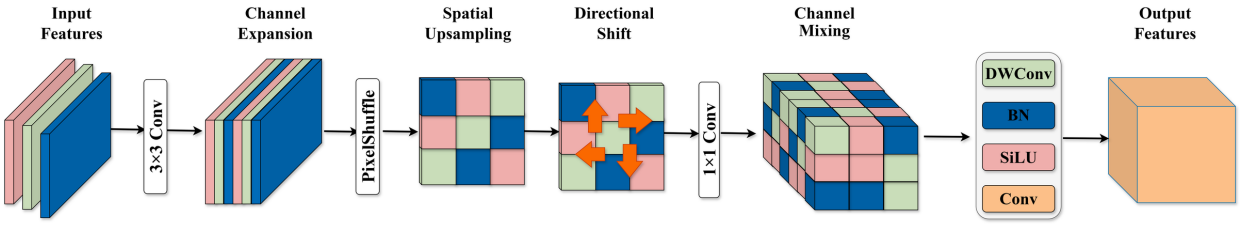


Figure 7. FASN module architecture diagram.

FASN is constructed based on sub-pixel convolution theory, the core idea of which is to learn predictive representations of high-resolution pixels in low-resolution feature

space. For upsampling factor s ($s = 2$ in this study), the network first expands input feature channel count to $C \cdot s^2$ through 3×3 convolution:

$$F_{\text{expand}} = \text{Conv}_{3\times3}(X) \tag{10}$$

where $X \in \mathbb{R}^{C\times H\times W}$ and $F_{\text{expand}} \in \mathbb{R}^{C\cdot s^2\times H\times W}$. The weights of this convolutional layer automatically learn how to reconstruct high-resolution pixel values from low-resolution features during training. Its frequency response function $\mathcal{H}_\omega$ can establish sufficient response intensity in high-frequency regions through backpropagation, thereby effectively compensating for high-frequency attenuation caused by traditional interpolation methods. Subsequently, through a periodic shuffle operation (PixelShuffle), channel dimension information is redistributed to spatial dimensions:

$$F_{\text{up}} = \text{PixelShuffle}(F_{\text{expand}}) \tag{11}$$

This operation transforms F_{expand} into $F_{\text{up}} \in \mathbb{R}^{C\times 2H\times 2W}$. In the frequency domain, this process is equivalent to mapping the convolution kernel frequency response to a wider frequency range, achieving information reconstruction beyond the original Nyquist frequency. To further enhance network modeling capability for spatial directions, FASN introduces direction-aware spatial shift mechanisms. The upsampled feature map F_{up} is uniformly divided into four groups along the channel dimension, with each group subjected to circular shift operations in four cardinal directions: up, down, left, and right. This operation enables convolution kernels to perceive neighborhood information in different directions across different channel groups—for example, for seedling leaves extending horizontally, left-right shift channel groups can capture their longitudinal extension features; for seedlings with vertical posture, up-down shift channel groups provide corresponding directional responses. The spatial shift mechanism breaks the translation equivariance constraint of standard convolution, enabling the network to distinguish semantic differences in absolute spatial positions, thereby effectively utilizing directional information from different hierarchical levels in feature pyramid fusion. Shifted features are mixed through 1×1 convolution:

$$F_{\text{mix}} = \text{Conv}_{1\times1}(\text{Concat}(F_{\text{shift1}}, F_{\text{shift2}}, F_{\text{shift3}}, F_{\text{shift4}})) \tag{12}$$

This operation performs a linear transformation in channel space, removing information redundancy between different directional groups and learning feature combination weights across directions. The 1×1 convolutional layer linearly combines the four directional groups with trained weights, so directional sensitivity emerges from learning rather than from a fixed structural bias. Finally, FASN employs a cascaded feature refinement module to further optimize upsampling quality. This module consists of depthwise separable convolution, first extracting spatial context through 3×3 depthwise convolution, then performing nonlinear transformation through batch normalization and SiLU activation, and finally achieving channel dimensionality reduction through 1×1 pointwise convolution:

$$Y = \text{Conv}_{1\times1}(\text{SiLU}(\text{BN}(\text{DWConv}_{3\times3}(F_{\text{mix}})))) \tag{13}$$

Depthwise separable convolution decomposes standard convolution into two steps—depthwise convolution and pointwise convolution—enabling feature extraction while maintaining receptive field coverage, suitable for deployment at multiple scale levels of the feature pyramid. FASN integrates frequency reconstruction of sub-pixel convolution, spatial enhancement of direction-aware shift, information aggregation of channel mixing, and cascaded refinement mechanisms, constructing an upsampling operator ori-

ented toward tiny target detection. Compared to nearest neighbor interpolation, this design recovers high-frequency details of seedling edges through frequency domain information reconstruction, maintains clarity of adjacent target boundaries, and suppresses upsampling artifacts, thereby providing multi-scale feature representations for subsequent detection heads.

4. Experimental Results and Analysis

4.1. Experimental Environment and Configuration

To validate the effectiveness of the proposed DDF-DETR model, all experiments were conducted on a high-performance computing platform. The experimental hardware comprised an NVIDIA RTX 6000 Ada Generation GPU (NVIDIA Corporation, Santa Clara, CA, USA) (96 GB GDDR6) and an Intel Xeon Platinum 8470Q processor (Intel Corporation, Santa Clara, CA, USA) (22 vCPU). The operating system was Ubuntu 22.04 LTS, with PyTorch 2.8.0 as the deep learning framework and CUDA version 12.8.

Model training employed the AdamW optimizer with an initial learning rate of 1×10^{-4}, incorporating a cosine annealing strategy for dynamic learning rate adjustment. To adapt to the tiny-scale features of cotton field seedlings, input images were resized to 800×800 pixels using a letterboxing strategy, with the shorter edge padded with gray pixels (value 114) to preserve the original aspect ratio. The training process was iterated for 150 epochs with a batch size of 4, except for TinyPerson, where a maximum of 500 epochs was applied due to its extremely sparse target distribution, with performance evaluated at the best checkpoint of each model.

Statistical analysis. Results in Tables 2–9 are reported as mean ± standard deviation (SD) over $n = 3$ independent training runs for each model and dataset, covering Precision, Recall, F1-score, mAP@0.5, and mAP@0.5:0.95; Parameters and GFLOPs are deterministic and therefore reported as single values. To enhance the objectivity and credibility of the study, significance analysis is used to provide a quantitative basis for model comparisons. In the tables, different lowercase superscript letters indicate significant differences ($p < 0.05$), whereas the same letter indicates no significant difference.

4.2. Evaluation Metrics

To objectively evaluate the performance of the DDF-DETR model in cotton field tiny target detection tasks, this paper selected Precision (P), Recall (R), mean Average Precision (mAP), and Giga Floating-Point Operations (GFLOPs) as primary evaluation metrics.

Precision and Recall: Precision reflects the accuracy of model predictions, while Recall reflects the ability of the model to identify positive samples. The computation formulas are as follows:

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

where TP denotes true positives; FP denotes false positives; FN denotes false negatives.

Mean Average Precision (mAP): mAP is the mean of Average Precision (AP) across all categories, measuring comprehensive detection performance of the model. It is mathematically defined as the average area under the P-R curve across all categories, computed as follows:

$$AP = \int_0^1 P(R)dR \tag{16}$$

$$mAP = \frac{1}{C}\sum_{i=1}^{C} AP_i \tag{17}$$

where C represents the total number of categories, and AP_i represents the Average Precision of the i-th category. This paper employs two mAP metrics: mAP@0.5 (IoU threshold 0.5) and mAP@0.5:0.95 (average IoU thresholds from 0.5 to 0.95, step 0.05).

Computational Complexity and Speed: GFLOPs is used to measure model computational complexity. Frames per second (FPS) is additionally reported to characterize inference speed during deployment.

4.3. Experimental Results Comparison

To comprehensively evaluate the overall performance of DDF-DETR, this study designed comparative experiments along three dimensions: first, validating model effectiveness in specific agricultural scenarios on the self-constructed Xinjiang cotton field dataset; second, verifying model generalization capability and detection robustness for extremely small targets on two public datasets, VisDrone2019-DET and TinyPerson; third, evaluating practical inference performance on an embedded edge platform to assess real-time deployability for field applications.

4.3.1. Module Ablation Study

To quantitatively evaluate the contribution of each core component in the DDF-DETR architecture to overall detection performance, this study designed progressive ablation experiments. Starting with RT-DETR-R18 as the baseline model, the Dynamic Gated Mixer Block (DGMB), Direction-Aware Adaptive Transformer Encoder (DAATE), and Frequency-Aware Sub-pixel Upsampling Network (FASN) were sequentially introduced, recording model performance in accuracy (mAP), parameter count (Params), and computational complexity (GFLOPs) at each stage. Experimental results are shown in Table 2.

Table 2. Ablation study results comparison.

Baseline	DGMB	DAATE	FASN	mAP@0.5 (%)	mAP@0.5:0.95 (%)	GFLOPs	Params (M)
✓				81.34 ± 0.12^{g}	58.18 ± 0.13^{g}	56.9	19.87
✓	✓			82.18 ± 0.17^{d}	60.20 ± 0.22^{e}	39.9	12.01
✓		✓		81.90 ± 0.04^{f}	59.04 ± 0.15^{f}	57.7	21.50
✓			✓	81.83 ± 0.09^{f}	59.48 ± 0.18^{f}	58.1	20.02
✓		✓	✓	82.39 ± 0.11^{c}	60.63 ± 0.21^{d}	58.8	21.64
✓	✓		✓	83.00 ± 0.21^{b}	61.91 ± 0.18^{b}	41.7	13.64
✓	✓	✓		82.85 ± 0.06^{b}	61.70 ± 0.12^{c}	41.0	12.16
✓	✓	✓	✓	83.72 ± 0.10^{a}	63.46 ± 0.06^{a}	42.8	13.78

Note: GFLOPs and Params are deterministic values reported as single values. Different superscript letters indicate significant differences ($p < 0.05$). The checkmarks indicate which modules are included in each experimental configuration.

Experimental results show that after integrating the DGMB module, the baseline model RT-DETR-R18 achieved a parameter reduction from 19.87 M to 12.01 M and a computational cost reduction from 56.9 GFLOPs to 39.9 GFLOPs, with mAP@0.5 improving by 0.84%. This indicates that the dynamic gating mechanism maintains feature extraction capability while reducing model complexity. Introducing FASN (Baseline + DGMB + FASN) yields 83.00% mAP@0.5, while adding DAATE (Baseline + DGMB + DAATE) yields 82.85%. With integration of all three modules (DGMB + DAATE + FASN), mAP@0.5 and mAP@0.5:0.95 reached 83.72% and 63.46%, respectively, with overall parameter count maintained at 13.78 M. The data indicate that the synergistic operation of each module improves detection accuracy in complex cotton field environments. The improvement in mAP metrics reflects that the model not only enhances the capture of seedling features but also reduces false positives for the missing seedling category through background noise suppression.

To quantitatively validate the kernel shape design in `DynamicInceptionDWConv2d`, we conducted ablation experiments on seven kernel shape configurations under the full model setup (Baseline+DGMB+DAATE+FASN). The results are summarized in Table 3.

Table 3. Ablation study on kernel shape configurations in DynamicInceptionDWConv2d.

Kernel Config	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)	GFLOPs
S only	81.93 ± 0.11^{e}	61.07 ± 0.14^{d}	13.51	42.3
H only	80.87 ± 0.24^{f}	59.84 ± 0.29^{f}	13.47	42.1
V only	81.08 ± 0.21^{ef}	60.12 ± 0.26^{e}	13.47	42.1
S+H	82.68 ± 0.12^{c}	62.14 ± 0.15^{c}	13.63	42.5
S+V	82.91 ± 0.11^{b}	62.43 ± 0.13^{b}	13.63	42.5
H+V	82.05 ± 0.15^{d}	61.58 ± 0.17^{d}	13.59	42.3
S+H+V (Ours)	83.72 ± 0.10^{a}	63.46 ± 0.06^{a}	13.78	42.8

Note: S: square kernel (3×3); H: horizontal strip (1×7); V: vertical strip (7×1). Results are reported as mean $\pm$ SD over $n = 3$ independent runs. GFLOPs and Params are deterministic single values. Different superscript letters indicate significant differences ($p < 0.05$).

The results exhibit a clear performance gradient across configurations. Among single-branch variants, H only achieves the lowest accuracy with the highest training variance, suggesting that the strictly anisotropic receptive field of horizontal strip kernels is inadequate for capturing isotropic target features; S only performs more stably owing to its inherent isotropic coverage. Among two-branch variants, S+V outperforms S+H, corroborating the dominant α_V weights at low altitude reported in Table 9 and confirming the greater contribution of vertical kernels to leaf morphology representation in near-ground cotton seedlings. H+V underperforms despite combining two directional kernels, which is attributable to the absence of isotropic coverage. The full S+H+V combination achieves the best detection accuracy with the lowest training variance, confirming the complementary roles of the three kernel shapes in multi-scale geometric modeling; removal of any single branch leads to statistically significant performance degradation.

4.3.2. Multi-Scale Method Comparison Experiments

To validate the effectiveness of the proposed DGMB and FASN in multi-scale feature fusion, DDF-DETR was compared with current mainstream advanced feature extraction modules. Comparison methods included DRB (UniRepLKNet) based on large kernel convolution, lightweight context-guided network ContextGuided, space-frequency selective convolution SFSCconv, and deformable convolution DCNv2. All experiments were conducted on the self-constructed Xinjiang cotton field dataset with identical training strategies and hyperparameter configurations. Experimental results are shown in Table 4.

Table 4. Performance comparison of different multi-scale methods.

Method	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Params (M)	GFLOPs
DRB (UniRepLKNet)	81.46 ± 0.10^{c}	61.13 ± 0.14^{c}	15.53	50.3
ContextGuided	79.72 ± 0.19^{e}	58.51 ± 0.26^{e}	12.23	38.2
SFSCconv	80.82 ± 0.18^{d}	60.08 ± 0.15^{d}	15.01	47.0
DCNv2	82.34 ± 0.18^{b}	62.15 ± 0.19^{b}	19.20	51.7
DGMB (Ours)	83.72 ± 0.10^{a}	63.46 ± 0.06^{a}	13.78	42.8

Note: GFLOPs and Params are deterministic values reported as single values. Different superscript letters indicate significant differences ($p < 0.05$).

The experimental results show that different feature extraction methods exhibit varying trade-offs between detection performance and computational cost. The DRB module, while enhancing receptive field through large kernel convolution, incurs high parameter count (15.53 M) and computational cost (50.3 GFLOPs), achieving only 81.46% mAP@0.5.

Although ContextGuided and SFSCconv reduce model complexity, they demonstrate suboptimal small target detection accuracy, with mAP@0.5 of 79.72% and 80.82%, respectively. In comparison, DCNv2 improves adaptability to irregular targets through deformable convolution, achieving 82.34% mAP@0.5, but incurs the highest computational cost (51.7 GFLOPs). The proposed DDF-DETR achieved optimal performance across all evaluation metrics, with mAP@0.5 and mAP@0.5:0.95 reaching 83.72% and 63.46%, respectively, representing improvements of 1.38% and 1.31% over DCNv2. DDF-DETR demonstrates superior feature preservation capability when processing visually weak missing seedling targets compared to methods such as large kernel convolution, benefiting from the efficiency of the dynamic gating mechanism with model parameters of only 13.78 M and computational cost controlled at 42.8 GFLOPs. These results validate that DDF-DETR can capture multi-scale features and high-frequency details of cotton seedlings while ensuring a lightweight design, achieving a balance between detection accuracy and efficiency.

4.3.3. Direction-Aware Linear Attention Encoder Comparison

To validate the directional modeling superiority of DAATE over existing linear attention encoders, this section conducts controlled comparisons by replacing only the attention mechanism while keeping DGMB, FASN, and SFFN fixed. Three representative methods are evaluated: Linear Attention achieves linear complexity through kernel decomposition; cosFormer incorporates relative positional re-weighting to enhance spatial awareness; and FLatten Transformer further improves visual feature modeling through locality-enhanced kernels. Results are shown in Table 5.

Table 5. Performance comparison of different linear attention encoders under the DGMB + FASN configuration.

Method	mAP@0.5 (%)	mAP@0.5:0.95 (%)	GFLOPs	Params (M)
Standard AIFI	83.00 ± 0.21^{c}	61.91 ± 0.18^{d}	41.7	13.64
Linear Attention [48]	82.31 ± 0.22^{e}	62.08 ± 0.25^{c}	41.5	13.71
cosFormer [49]	82.89 ± 0.18^{d}	62.71 ± 0.19^{b}	41.6	13.76
FLatten Transformer [50]	83.18 ± 0.15^{b}	63.02 ± 0.17^{b}	41.8	13.77
DAATE (Ours)	83.72 ± 0.10^{a}	63.46 ± 0.06^{a}	42.8	13.78

Note: All methods use the DGMB + FASN configuration with only the attention mechanism replaced. Results are mean ± SD over $n = 3$ independent runs. GFLOPs and Params are deterministic single values. Different superscript letters indicate significant differences ($p < 0.05$).

Linear Attention falls below the AIFI baseline, as the non-negativity constraint of ELU kernels prevents the encoder from forming effective inhibitory responses to background film textures, reducing the ability of the model to discriminate missing seedling holes from background interference. The relative positional re-weighting of cosFormer improves local structure awareness, but its symmetric formulation limits modeling of the directional asymmetry between vertical leaf structures and horizontal film edges. The locality-enhanced kernels of FLatten Transformer yield the closest performance to DAATE through finer neighborhood modeling, yet its isotropic attention patterns remain insensitive to the row-column planting directions, leaving feature confusion between cotton seedlings and cross-row backgrounds unresolved. The polarized kernel decomposition of DAATE simultaneously encodes excitatory and inhibitory spatial relationships, effectively separating foreground seedling responses from background texture activations; the asymmetric key-only positional encoding further prioritizes attention aggregation along the row-column directions, embedding the spatial regularity of the crop layout as a geometric prior into the feature aggregation process.

4.3.4. Upsampling Method Comparison Experiments

To verify the advantages of the FASN module in recovering high-frequency details of small targets, this section compares the impact of different upsampling operators on detection performance. Experiments were conducted on the foundation configured with DGMB and DAATE, with FASN replaced by nearest neighbor interpolation and mainstream learnable upsampling methods CARAFE and DySample for comparison. Experimental results are shown in Table 6.

Table 6. Performance comparison of different upsampling methods.

Method	P (%)	R (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
Nearest	94.54 ± 0.15^{b}	78.82 ± 0.11^{c}	83.00 ± 0.21^{b}	61.91 ± 0.18^{c}
CARAFE	93.95 ± 0.05^{c}	78.48 ± 0.13^{c}	81.84 ± 0.30^{d}	60.36 ± 0.19^{d}
DySample	94.60 ± 0.17^{ab}	79.69 ± 0.09^{b}	82.29 ± 0.08^{c}	62.52 ± 0.20^{b}
FASN (Ours)	94.85 ± 0.11^{a}	80.11 ± 0.08^{a}	83.72 ± 0.10^{a}	63.46 ± 0.06^{a}

Note: All experiments use the configuration with DGMB and DAATE, with only the upsampling module varied. Different superscript letters indicate significant differences ($p < 0.05$).

As shown in Table 6, nearest neighbor interpolation (baseline) achieved 83.00% mAP@0.5; however, its zero-order hold characteristic causes high-frequency attenuation, limiting edge detail recovery for small targets. CARAFE constructs upsampling weights via content-aware reassembly kernels, yet its mAP@0.5 dropped to 81.84% in cotton field scenarios, attributed to large receptive field neighborhoods that introduce feature confusion between adjacent targets in dense configurations. DySample achieved 62.52% mAP@0.5:0.95 through dynamic point sampling, which improves localization precision, but its isotropic sampling pattern does not capture the directional texture of cotton seedling leaves. The proposed FASN achieved the highest performance across all metrics, with Recall of 80.11%, mAP@0.5 of 83.72%, and mAP@0.5:0.95 of 63.46%. Feature map visualizations in Figure 8 show that nearest neighbor interpolation produces blocky artifacts and blurred edge responses at seedling boundaries; CARAFE reduces interpolation artifacts but exhibits response fusion at adjacent target boundaries; DySample yields concentrated feature responses with insufficient directional texture representation. FASN, combining sub-pixel convolution for frequency reconstruction with direction-aware spatial shift, produces distinct edge responses and preserves individual target boundaries in densely arranged regions. These results confirm that the direction-aware design of FASN recovers high-frequency information suited to the anisotropic morphology of cotton seedling leaves, providing spatial detail representations for detection heads in feature pyramid fusion.

4.3.5. Cotton Field Scenario Performance Comparison

To comprehensively evaluate the detection performance of the proposed DDF-DETR model in complex cotton field environments, this study conducted comparative experiments on the self-constructed cotton field dataset. The experiments selected multiple mainstream object detection algorithms as benchmarks, covering two-stage detectors (Faster R-CNN), single-stage detectors (YOLOv8, YOLOv11, YOLOv12, YOLOv26), and Transformer-based detectors (DETR, Deformable DETR, RT-DETR series, DEIM). Experimental results are shown in Table 7.

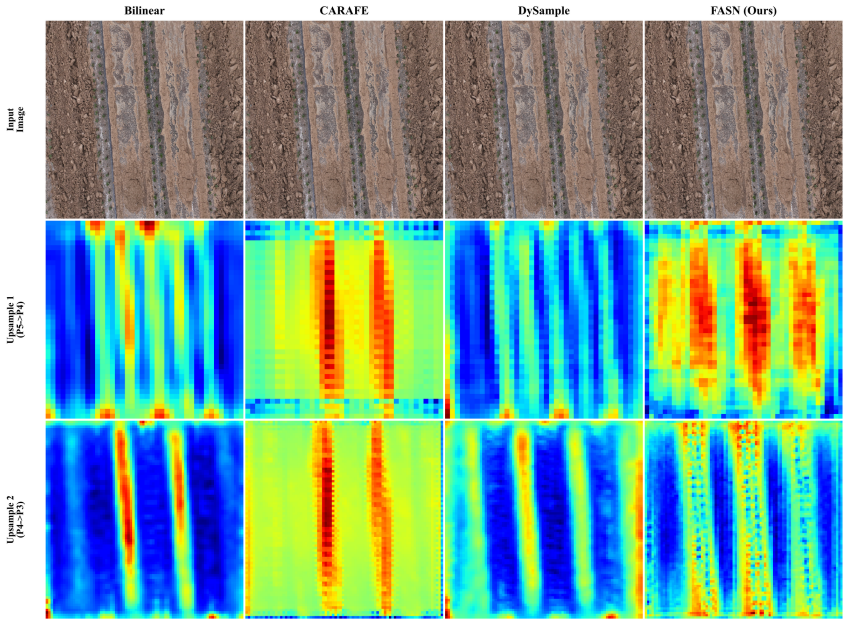


Figure 8. Feature map visualization comparison of different upsampling methods.

Table 7. Performance comparison of different detection models on the self-constructed cotton field dataset.

Method	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	GFLOPs	Params (M)
Faster R-CNN	53.50 ± 0.26^{h}	39.10 ± 0.40^{h}	45.27 ± 0.21^{h}	54.00 ± 0.20^{g}	39.23 ± 0.29^{g}	–	–
YOLO8n	92.28 ± 0.15^{f}	76.55 ± 0.09^{f}	83.64 ± 0.16^{e}	80.41 ± 0.27^{e}	58.33 ± 0.17^{e}	8.1	3.00
YOLO8m	94.43 ± 0.20^{bc}	78.82 ± 0.13^{de}	85.78 ± 0.12^{cd}	81.96 ± 0.21^{d}	60.42 ± 0.15^{cd}	78.7	25.85
YOLO11n	92.92 ± 0.08^{ef}	77.58 ± 0.17^{ef}	84.43 ± 0.13^{de}	80.99 ± 0.26^{e}	59.22 ± 0.14^{de}	6.3	2.58
YOLO11m	94.95 ± 0.11^{ab}	79.44 ± 0.12^{cd}	86.58 ± 0.06^{bc}	83.05 ± 0.11^{bc}	62.02 ± 0.13^{b}	67.7	20.03
YOLO12m	94.40 ± 0.11^{c}	78.17 ± 0.11^{e}	85.33 ± 0.13^{d}	81.31 ± 0.16^{de}	60.25 ± 0.05^{cd}	67.2	20.11
YOLO26m	95.00 ± 0.13^{ab}	80.88 ± 0.13^{ab}	87.35 ± 0.11^{ab}	83.78 ± 0.16^{ab}	63.84 ± 0.20^{a}	67.9	20.36
DETR	50.71 ± 0.36^{i}	34.92 ± 0.93^{i}	41.48 ± 0.06^{i}	44.70 ± 0.50^{h}	29.83 ± 0.77^{h}	187.2	41.56
RT-DETR-R18	95.05 ± 0.15^{ab}	77.86 ± 0.10^{ef}	85.52 ± 0.16^{cd}	81.34 ± 0.12^{de}	58.18 ± 0.13^{e}	56.9	19.87
RT-DETR-R34	94.10 ± 0.13^{cd}	79.64 ± 0.17^{bc}	86.43 ± 0.09^{bc}	83.01 ± 0.16^{bc}	61.56 ± 0.26^{bc}	88.8	31.11
RT-DETR-R50	94.69 ± 0.14^{bc}	80.49 ± 0.15^{ab}	87.08 ± 0.06^{ab}	83.75 ± 0.30^{ab}	64.05 ± 0.10^{a}	129.5	41.96
Deformable DETR	70.51 ± 0.76^{g}	55.18 ± 1.15^{g}	61.53 ± 0.26^{g}	68.54 ± 0.60^{f}	55.48 ± 1.42^{f}	201.2	40.60
DEIM [51]	94.09 ± 0.55^{cd}	81.15 ± 1.17^{a}	87.23 ± 0.26^{a}	84.12 ± 0.60^{a}	60.90 ± 0.68^{cd}	178.5	31.50
DDF-DETR (Ours)	94.85 ± 0.11^{ab}	80.11 ± 0.08^{bc}	86.87 ± 0.14^{b}	83.72 ± 0.10^{ab}	63.46 ± 0.06^{a}	42.8	13.78

Note: GFLOPs and Params are deterministic values reported as single values. "–" indicates that GFLOPs and Params for Faster R-CNN were not measured under the unified framework. Different superscript letters indicate significant differences ($p < 0.05$).

Experimental results show that different detection architectures exhibit distinct performance characteristics in cotton field tiny target detection tasks. Faster R-CNN and the original DETR, limited by their feature extraction capability for tiny targets, achieved only 39.23% and 29.83%, respectively, on the mAP@0.5:0.95 metric. Although the YOLO series detectors have advantages in inference speed, their feature discrimination capability in complex backgrounds is limited; for example, YOLO11m achieved 83.05% mAP@0.5. In comparison, the RT-DETR series, leveraging hybrid encoders to enhance multi-scale feature interaction, achieved 81.34% mAP@0.5 with RT-DETR-R18. The proposed DDF-DETR achieves a favorable trade-off between detection accuracy and computational efficiency, attaining competitive performance with mAP@0.5 and mAP@0.5:0.95 reaching 83.72% and 63.46%, respectively, and an F1 score of 86.87%. Compared with the baseline model RT-DETR-R18, DDF-DETR achieved a 2.38% mAP@0.5 improvement while reducing the parameter count by 30.6% (from 19.87 M to 13.78 M) and the computational cost (from 56.9 GFLOPs to 42.8 GFLOPs). For the challenging "missing seedling" category, the YOLO

series tends to produce false detections due to plastic film shadows and cracks, while DDF-DETR, benefiting from the capability of the dynamic gating mechanism to suppress background noise, distinguishes genuine missing seedling holes from similar background textures, thereby reducing the false alarm rate in missing seedling detection while ensuring the seedling detection rate and improving overall mAP. Figure 9 shows representative detection visualizations on the cotton field dataset. These results validate the effectiveness of dynamic gated mixer blocks and direction-aware mechanisms in processing dense tiny targets and suppressing background interference.

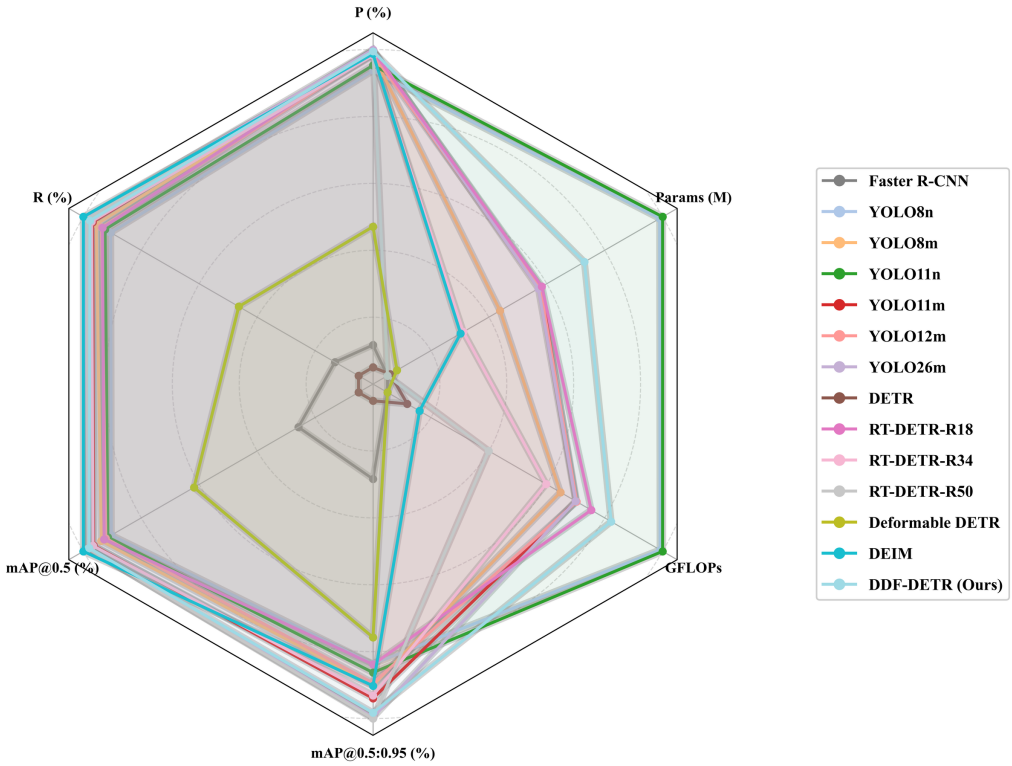


Figure 9. Visualization on the cotton field dataset.

To further investigate the category-level detection capability of DDF-DETR and quantify its background suppression performance, confusion matrices of RT-DETR-R18 and DDF-DETR on the cotton field test set are presented in Figure 10.

As shown in Figure 10, DDF-DETR reduces background misclassifications across both categories compared to RT-DETR-R18. For the "seedling" category, instances misclassified as background decrease from 2251 to 1891, a reduction of 16.0%. For the "missing seedling" category, background false negatives decrease from 6638 to 5282, a reduction of 20.4%. This reduction is of particular note given that "missing seedling" targets lack salient visual features and are prone to confusion with background textures such as plastic film shadows and soil cracks. The decrease in background misclassifications, combined with an increase in true positives from 10,680 to 11,410, indicates that the dynamic gating mechanism enables the model to discriminate genuine missing seedling holes from visually similar background noise, improving detection reliability in complex cotton field environments.

4.3.6. Generalization Capability Validation

To further evaluate the generalization performance of DDF-DETR across different scenarios, experiments were conducted on two public benchmarks: VisDrone2019-DET and TinyPerson.

VisDrone2019-DET [52] is a large-scale UAV object detection benchmark collected by the AISKYEYE team at Tianjin University, containing 10,209 images with over 2.6 million annotations across 10 object categories. The dataset is characterized by significant scale variations, high object density, and complex backgrounds from 14 cities under diverse conditions.

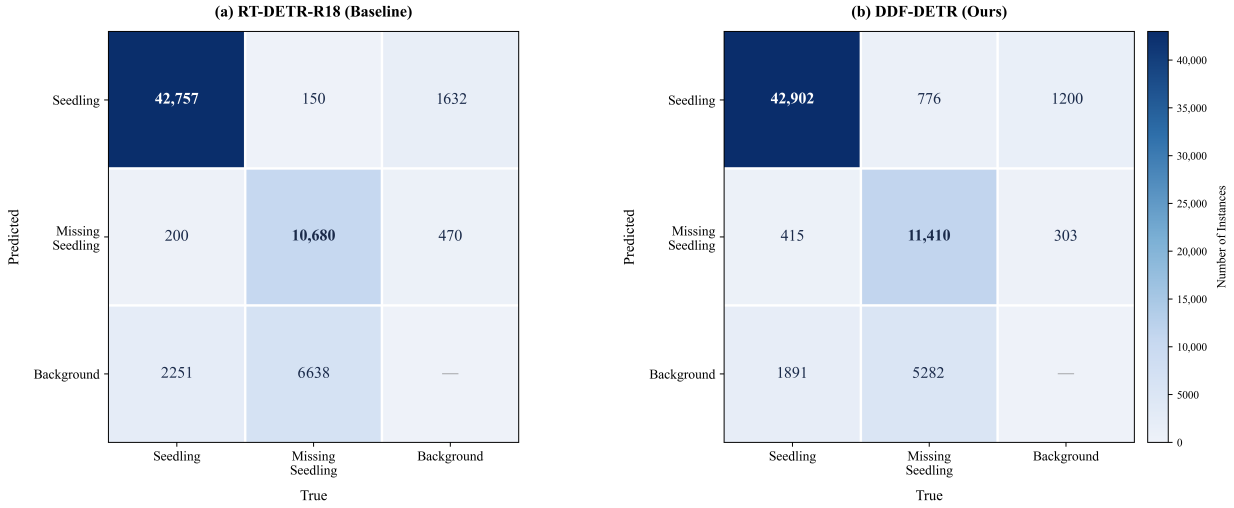


Figure 10. Confusion matrix comparison between RT-DETR-R18 and DDF-DETR on the cotton field test set. Results are from the representative run whose mAP@0.5 is closest to the mean across $n = 3$ independent training runs.

TinyPerson [53] is a benchmark for extremely small object detection introduced at WACV 2020, comprising 1610 images with 72,651 annotations. The target sizes range from 2 to 32 pixels, with most objects below 20 pixels in height, presenting challenges of limited visual information and massive background interference.

These two datasets validate model capability from complementary perspectives: multi-scale adaptation in general UAV scenarios and extremely small target detection. Results are shown in Table 8.

Table 8. Performance comparison of different models on generalization datasets.

Dataset	Model	P (%)	R (%)	F1 (%)	mAP@0.5 (%)	mAP@0.5:0.95 (%)
VisDrone2019	YOLO8m	52.15 ± 0.21^{e}	41.58 ± 0.23^{d}	45.98 ± 0.16^{d}	43.14 ± 0.20^{d}	26.41 ± 0.26^{e}
	YOLO11m	53.62 ± 0.43^{d}	42.20 ± 0.23^{d}	46.66 ± 0.21^{d}	44.30 ± 0.20^{c}	26.95 ± 0.14^{de}
	YOLO12m	53.29 ± 0.17^{d}	40.57 ± 0.34^{d}	45.68 ± 0.19^{d}	41.93 ± 0.24^{e}	25.22 ± 0.30^{f}
	YOLO26m	65.97 ± 0.18^{a}	45.06 ± 0.08^{c}	53.48 ± 0.30^{c}	50.41 ± 0.12^{b}	29.43 ± 0.20^{c}
	RT-DETR-R18	62.50 ± 0.14^{c}	48.97 ± 0.20^{b}	54.50 ± 0.22^{b}	50.94 ± 0.18^{b}	31.53 ± 0.08^{b}
	DDF-DETR	65.61 ± 0.24^{ab}	51.41 ± 0.21^{a}	57.60 ± 0.12^{a}	53.61 ± 0.12^{a}	33.07 ± 0.09^{a}
TinyPerson	YOLO8m	40.10 ± 0.25^{f}	19.77 ± 0.10^{f}	26.59 ± 0.14^{f}	19.94 ± 0.19^{e}	6.83 ± 0.16^{f}
	YOLO11m	45.09 ± 0.12^{d}	26.06 ± 0.18^{d}	32.80 ± 0.17^{d}	25.77 ± 0.17^{c}	9.17 ± 0.20^{d}
	YOLO12m	43.40 ± 0.06^{e}	25.49 ± 0.28^{d}	32.05 ± 0.09^{d}	24.15 ± 0.14^{d}	8.34 ± 0.15^{e}
	YOLO26m	53.40 ± 0.26^{b}	35.96 ± 0.25^{b}	43.00 ± 0.19^{b}	34.63 ± 0.20^{b}	11.19 ± 0.18^{b}
	RT-DETR-R18	50.66 ± 0.24^{c}	33.56 ± 0.38^{c}	40.53 ± 0.08^{c}	31.37 ± 0.28^{b}	10.69 ± 0.31^{bc}
	DDF-DETR	54.46 ± 0.18^{a}	38.94 ± 0.32^{a}	45.36 ± 0.05^{a}	35.75 ± 0.15^{a}	12.37 ± 0.19^{a}

Note: Significance letters are assigned independently within each dataset partition. Different superscript letters indicate significant differences ($p < 0.05$).

Table 8 presents a performance comparison of different models on the VisDrone2019 and TinyPerson datasets. The VisDrone2019 dataset features complex backgrounds and large-scale variations. DDF-DETR outperformed other models on mAP@0.5 and mAP@0.5:0.95 metrics, demonstrating adaptability when handling multi-category background interference and scale variations. Through dynamic convolution and adaptive feature modulation, the model reduces background interference and enhances target recog-

nition. The primary challenge of the TinyPerson dataset is extremely small targets. DDF-DETR also performed well on this dataset, particularly on mAP@0.5 and mAP@0.5:0.95, demonstrating detection and localization capability for extremely small targets. These results indicate that DDF-DETR maintains stable detection accuracy when facing complex backgrounds and extremely small targets, demonstrating its generalization capability across different scenarios and tasks. Figure 11 presents visualization examples on these datasets.



Figure 11. Visualization on generalization datasets.

4.3.7. Visualization Analysis

To qualitatively analyze the attention distributions of different detection models, Grad-CAM [54] was employed to generate heatmap visualizations of high-level feature responses. The comparison involved six models: YOLO8m, YOLO11m, YOLO12m, YOLO26m, RT-DETR-R18, and the proposed DDF-DETR. Test samples were selected from the self-constructed cotton field dataset, covering scenarios with plastic film reflections, dense seedling distributions, and small-scale targets. Results are presented in Figure 12.

As shown in Figure 12, the YOLO series models produce activation regions that extend along row textures, with responses distributed across film-covered areas rather than concentrated on individual targets. RT-DETR-R18 generates activation patterns with reduced spatial diffusion compared to the YOLO series, though non-target responses remain observable in regions with reflections or complex textures. The proposed DDF-DETR produces point-like activation peaks at cotton seedling locations, with boundaries between adjacent targets preserved in dense distribution scenarios. When target sizes are small or edges are blurred, DDF-DETR maintains localized response centers without redundant activation along row directions. These observations are consistent with the quantitative results, suggesting that the dynamic gating mechanism and direction-aware modeling contribute to discriminative feature learning, while frequency-aware upsampling preserves spatial details for small target discrimination.

To further quantify the input-adaptive behavior of DynamicInceptionDWConv2d in DGMB, forward hooks were registered on the Softmax output of the attention sub-network to extract the branch weight vector $\alpha = \{\alpha_S, \alpha_H, \alpha_V\}$ for each sample without modifying the inference pipeline. The mean and standard deviation of α were then computed on the validation set, stratified by UAV acquisition altitude. Results are presented in Table 9 and Figure 13.

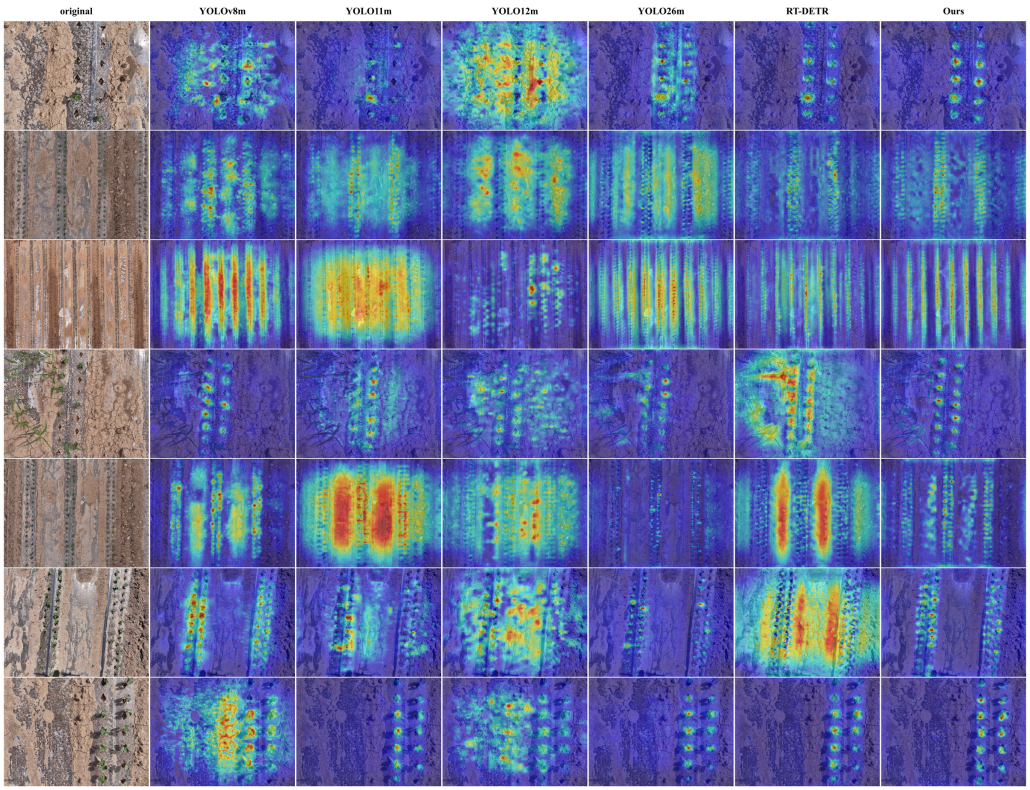


Figure 12. Grad-CAM heatmap visualization comparison of different detection models in cotton field scenarios.

Table 9. Mean branch weights (α_S, α_H, α_V) of DynamicInceptionDWConv2d on the validation set, stratified by UAV acquisition altitude.

Altitude	α_S	α_H	α_V
1.5 m	0.15 ± 0.06^{d}	0.30 ± 0.07^{a}	0.55 ± 0.08^{a}
3 m	0.29 ± 0.12^{c}	0.35 ± 0.11^{a}	0.36 ± 0.10^{b}
5 m	0.38 ± 0.11^{b}	0.30 ± 0.10^{a}	0.22 ± 0.10^{c}
15 m	0.81 ± 0.03^{a}	0.11 ± 0.02^{b}	0.08 ± 0.02^{d}

Note: Values are reported as mean ± SD over the validation set. Different superscript letters indicate significant differences ($p < 0.05$).

As shown in Table 9 and Figure 13, the dominant branch weight shifts systematically from α_V to α_S with increasing UAV acquisition altitude. At 1.5 m, α_V dominates, consistent with the vertically elongated leaf morphology resolved at close range. As altitude increases to 15 m, α_S progressively dominates, corresponding to the near-isotropic target appearance at long range where directional strip kernels become less discriminative. The non-trivial within-group standard deviations across all altitude strata further confirm that α varies continuously with each input sample rather than converging to fixed values, demonstrating genuine input-adaptive dynamic behavior of the module.

4.3.8. Edge Device Deployment Performance

To validate the practical deployability of DDF-DETR in field cotton seedling monitoring, inference performance was evaluated on an NVIDIA Jetson Orin Nano 8G embedded platform (NVIDIA Corporation, Santa Clara, CA, USA). All models were exported to TensorRT `.engine` format with FP16 precision. Measurements were conducted at 800 × 800 input resolution with batch size 1, using 50 warmup iterations followed by 200 timed iterations. Results are shown in Table 10.

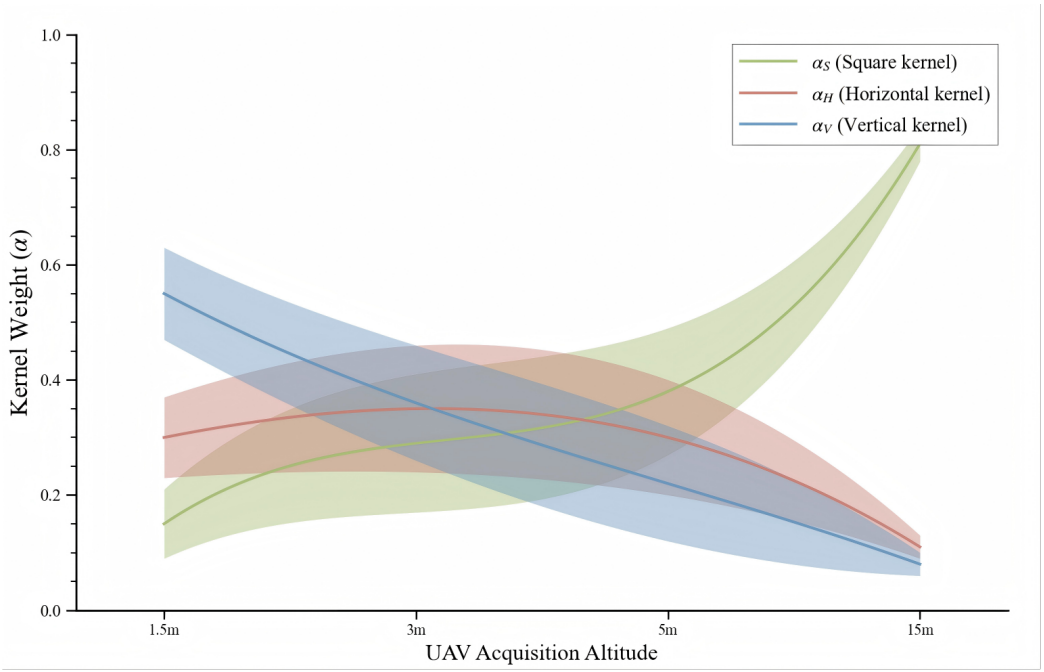


Figure 13. Mean branch weights (α_S, α_H, α_V) of DynamicInceptionDWConv2d on the validation set, stratified by UAV acquisition altitude. Solid lines denote mean values; shaded regions indicate $\pm$standard deviation.

Table 10. Inference performance comparison on NVIDIA Jetson Orin Nano 8G.

Model	Precision	Latency (ms)	FPS
RT-DETR-R18	FP16	41.22 ± 0.27	24.3
DDF-DETR	FP16	29.30 ± 0.24	34.1

DDF-DETR achieves 34.1 FPS on the Jetson Orin Nano 8G, improving throughput by 40.3% over the RT-DETR-R18 baseline at 24.3 FPS and reducing latency from 41.22 ms to 29.30 ms, corresponding to a 28.9% reduction. A prior study deploying an agricultural object detection model on the same platform achieved 26 FPS and explicitly confirmed this speed meets the requirements for real-time field detection [55]. The 34.1 FPS achieved by DDF-DETR exceeds this threshold, confirming its practical deployability on resource-constrained edge devices.

5. Conclusions

This paper proposes the DDF-DETR model for field cotton seedling detection, addressing feature degradation and background interference issues through three core components: Dynamic Gated Mixer Block (DGMB), Direction-Aware Adaptive Transformer Encoder (DAATE), and Frequency-Aware Sub-pixel Upsampling Network (FASN). On the self-constructed Xinjiang cotton seedling dataset, DDF-DETR achieved 83.72% mAP@0.5 and 63.46% mAP@0.5:0.95, representing improvements of 2.38% and 5.28%, respectively, over the baseline model RT-DETR-R18. The model parameter count is 13.78 M and computational cost is 42.8 GFLOPs, representing reductions of 30.6% and 24.8% compared to the baseline, respectively. The 5.28% improvement in mAP@0.5:0.95 indicates that the model has a more precise target localization capability, validating the effectiveness of FASN in recovering edge high-frequency information. Generalization experiments on VisDrone2019 and TinyPerson datasets further demonstrate model robustness. Deployment evaluation on an NVIDIA Jetson Orin Nano 8G embedded platform further demonstrated that DDF-DETR achieves 34.1 FPS in edge inference, confirming its practical suitability for real-time

field monitoring on resource-constrained devices. The current method relies solely on RGB imagery and does not exploit spectral signatures that could further separate seedlings from spectrally similar weeds. Future work will integrate multi-spectral channels into the detection pipeline and evaluate long-term deployment across different cotton growing seasons and regions.

Author Contributions: F.X.: Conceptualization, Methodology, Software, Investigation, Data curation, Writing—original draft, Writing—review and editing, Visualization. H.Z.: Data curation, Methodology, Validation. Y.P.: Data curation, Methodology. Y.L.: Data curation, Methodology. L.D.: Supervision, Methodology, Project administration, Funding acquisition, Writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Major Science and Technology Special Project of Xinjiang Uygur Autonomous Region (Grant No. 2023A02002) and the National College Students Innovation and Entrepreneurship Training Program of China (Grant No. 202510758024).

Data Availability Statement: The data presented in this study are available on request from the corresponding author on reasonable request.

Acknowledgments: We are appreciative of the reviewers' valuable suggestions on this manuscript and the editor's efforts in processing the manuscript.

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

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
DETR	Detection Transformer
RT-DETR	Real-Time Detection Transformer
DDF-DETR	Dynamic-Direction-Frequency Detection Transformer
DGMB	Dynamic Gated Mixer Block
DAATE	Direction-Aware Adaptive Transformer Encoder
FASN	Frequency-Aware Sub-pixel Upsampling Network
CNN	Convolutional Neural Network
mAP	mean Average Precision
IoU	Intersection over Union
GFLOPs	Giga Floating-point Operations
GLU	Gated Linear Unit
AIFI	Attention-based Intra-scale Feature Interaction
CCFM	CNN-based Cross-scale Feature Fusion Module
NMS	Non-Maximum Suppression

References

1. National Bureau of Statistics of China. *China Statistical Yearbook 2024*; China Statistics Press: Beijing, China, 2024.
2. Feng, A.; Zhou, J.; Vories, E.; Sudduth, K.A. Evaluation of cotton emergence using UAV-based imagery and deep learning. *Comput. Electron. Agric.* **2020**, *177*, 105711. [CrossRef]
3. Li, Z.; Guo, R.; Li, M.; Chen, Y.; Li, G. A review of computer vision technologies for plant phenotyping. *Comput. Electron. Agric.* **2020**, *176*, 105672. [CrossRef]
4. Tsouros, D.C.; Bibi, S.; Sarigiannidis, P.G. A review on UAV-based applications for precision agriculture. *Information* **2019**, *10*, 349. [CrossRef]
5. Sishodia, R.P.; Ray, R.L.; Singh, S.K. Applications of remote sensing in precision agriculture: A review. *Remote Sens.* **2020**, *12*, 3136. [CrossRef]
6. He, J.; Wang, Y.; Deng, L.; Li, Y.; Li, Z.L. Phase congruency image mosaicking approach for aerial mid-wave infrared low-overlap array scanning images. *ISPRS J. Photogramm. Remote Sens.* **2025**, *227*, 185–204. [CrossRef]

7. Wang, N.; Wu, Q.; Gui, Y.; Hu, Q.; Li, W. Cross-Modal Segmentation Network for Winter Wheat Mapping in Complex Terrain Using Remote-Sensing Multi-Temporal Images and DEM Data. *Remote Sens.* **2024**, *16*, 1775. [CrossRef]
8. Ren, L.; Li, C.; Yang, G.; Zhao, D.; Zhang, C.; Xu, B.; Li, H.; Yang, H. The detection of maize seedling quality from UAV images based on deep learning and Voronoi diagram algorithms. *Remote Sens.* **2024**, *16*, 3548. [CrossRef]
9. Ayhan, B.; Kwan, C. Application of deep learning with stratified k-fold for vegetation species discrimination in a protected mountainous region using Sentinel-2 image. *Remote Sens.* **2020**, *12*, 1808.
10. Chen, H.; Chen, H.; Huang, X.; Zhang, S.; Chen, S.; Cen, F.; He, T.; Zhao, Q.; Gao, Z. Estimation of sorghum seedling number from drone image based on support vector machine and YOLO algorithms. *Front. Plant Sci.* **2024**, *15*, 1399872. [CrossRef]
11. Feng, Y.; Chen, W.; Ma, Y.; Zhang, Z.; Gao, P.; Lv, X. Cotton seedling detection and counting based on UAV multispectral images and deep learning methods. *Remote Sens.* **2023**, *15*, 2680. [CrossRef]
12. Tian, F.; Ransom, C.J.; Zhou, J.; Wilson, B.; Sudduth, K.A. Assessing the impact of soil and field conditions on cotton crop emergence using UAV-based imagery. *Comput. Electron. Agric.* **2024**, *218*, 108738. [CrossRef]
13. Zou, Z.; Chen, K.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *Proc. IEEE* **2023**, *111*, 257–276. [CrossRef]
14. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
15. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 5–13 December 2015; pp. 1440–1448.
16. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 7–12 December 2015; p. 28.
17. Teng, Y.; Zhang, J.; Dong, S.; Zheng, S.; Liu, L. MSR-RCNN: A multi-class crop pest detection network based on a multi-scale super-resolution feature enhancement module. *Front. Plant Sci.* **2022**, *13*, 810546. [CrossRef]
18. Quan, L.; Feng, H.; Lv, Y.; Wang, Q.; Zhang, C.; Liu, J.; Yuan, Z. Maize seedling detection under different growth stages and complex field environments based on an improved Faster R-CNN. *Biosyst. Eng.* **2019**, *184*, 1–23. [CrossRef]
19. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
20. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
21. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLO. 2023. Available online: https://github.com/ultralytics/ultralytics (accessed on 15 January 2026).
22. 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.
23. Sapkota, R.; Cheppally, R.H.; Sharda, A.; Karkee, M. YOLO26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv* **2025**, arXiv:2509.25164. [CrossRef]
24. Gallo, I.; Rehman, A.U.; Dehkordi, R.H.; Landro, N.; La Grassa, R.; Boschetti, M. Deep object detection of crop weeds: Performance of YOLOv7 on a real case dataset from UAV images. *Remote Sens.* **2023**, *15*, 539. [CrossRef]
25. Dong, Q.; Sun, L.; Han, T.; Cai, M.; Gao, C. PestLite: A novel YOLO-based deep learning technique for crop pest detection. *Agriculture* **2024**, *14*, 228. [CrossRef]
26. Shahi, T.B.; Xu, C.Y.; Neupane, A.; Guo, W. Recent advances in crop disease detection using UAV and deep learning techniques. *Remote Sens.* **2023**, *15*, 2450. [CrossRef]
27. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020; pp. 213–229.
28. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable transformers for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Vienna, Austria, 3–7 May 2021.
29. Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; Zhang, L. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
30. Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023.
31. Lv, W.; Xu, S.; Zhao, Y.; Wang, G.; Wei, J.; Cui, C.; Du, Y.; Dang, Q.; Liu, Y. DETRs beat YOLOs on real-time object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 16965–16974.
32. Li, X.; Cai, M.; Tan, X.; Yin, C.; Chen, W.; Liu, Z.; Han, Y. An efficient transformer network for detecting multi-scale chicken in complex free-range farming environments via improved RT-DETR. *Comput. Electron. Agric.* **2024**, *224*, 109160. [CrossRef]

33. Wang, S.; Dong, S.; Chen, S.; Liu, M. HESP-DETR: An efficient scale-aware model based on the improved RT-DETR for pomegranate growth stage detection. *Neurocomputing* **2025**, *647*, 130462. [CrossRef]
34. Wang, H.; Nguyen, T.H.; Nguyen, T.N.; Dang, M. PD-TR: End-to-end plant diseases detection using a transformer. *Comput. Electron. Agric.* **2024**, *224*, 109123. [CrossRef]
35. Alsabhan, W.; Alotaiby, T. A survey of the vision transformers and their CNN-transformer based variants. *Artif. Intell. Rev.* **2024**, *57*, 281.
36. Miri Rekavandi, A.; Rashidi, S.; Boussaid, F.; Hoefs, S.; Akbas, E.; Bennamoun, M. Transformers in small object detection: A benchmark and survey of state-of-the-art. *ACM Comput. Surv.* **2025**, *58*, 1–33. [CrossRef]
37. Guo, X.; Feng, Q.; Guo, F. CMTNet: A hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture. *Sci. Rep.* **2025**, *15*, 12383. [CrossRef]
38. Xie, W.; Zhao, M.; Liu, Y.; Yang, D.; Huang, K.; Fan, C.; Wang, Z. Recent advances in Transformer technology for agriculture: A comprehensive survey. *Eng. Appl. Artif. Intell.* **2024**, *138*, 109412. [CrossRef]
39. Ding, X.; Zhang, X.; Han, J.; Ding, G. Scaling Up Your Kernels to 31 × 31: Revisiting Large Kernel Design in CNNs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 11963–11975.
40. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, M.M.; Yang, J.; Li, X. Large Selective Kernel Network for Remote Sensing Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 16794–16805.
41. Shi, W.; Caballero, J.; Huszár, F.; Totz, J.; Aitken, A.P.; Bishop, R.; Rueckert, D.; Wang, Z. Real-Time Single Image and Video Super-Resolution Using an Efficient Sub-Pixel Convolutional Neural Network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 1874–1883.
42. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. CARAFE: Content-Aware ReAssembly of FEatures. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 3007–3016.
43. Liu, W.; Lu, H.; Fu, H.; Cao, Z. DySample: Dynamic Upsampling for Semantic Segmentation. *arXiv* **2023**, arXiv:2308.14073.
44. Ding, X.; Zhang, Y.; Ge, Y.; Zhao, S.; Song, L.; Yue, X.; Shan, Y. UniRepLKNet: A Universal Perception Large-Kernel ConvNet for Audio, Video, Point Cloud, Time-Series and Image Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 28877–28888.
45. Zhu, X.; Hu, H.; Lin, S.; Dai, J. Deformable ConvNets v2: More Deformable, Better Results. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 3403–3412.
46. Wu, T.; Tang, S.; Zhang, R.; Cao, J.; Zhang, Y. CGNet: A Light-weight Context Guided Network for Semantic Segmentation. *arXiv* **2018**, arXiv:1811.08201. [CrossRef] [PubMed]
47. Li, K.; Wang, D.; Hu, Z.; Li, M.; Wang, Q.; Peng, S. Unleashing Channel Potential: Space-Frequency Selection Convolution for SAR Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–21 June 2024; pp. 17323–17332.
48. Katharopoulos, A.; Vyas, A.; Pappas, N.; Fleuret, F. Transformers are RNNs: Fast autoregressive transformers with linear attention. In Proceedings of the International Conference on Machine Learning (ICML), Virtual, 12–18 July 2020; pp. 5156–5165.
49. Qin, Z.; Zhang, Y.; Meng, S.; Cao, Z.; Peng, H. cosFormer: Rethinking softmax in attention. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual, 25–29 April 2022.
50. Han, X.; Wang, D.; Kong, L.; Shao, L. FLatten Transformer: Vision transformer using focused linear attention. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–6 October 2023; pp. 6011–6021.
51. Huang, S.; Lu, Z.; Cun, X.; Yu, Y.; Zhou, X.; Shen, X. DEIM: DETR with Improved Matching for Fast Convergence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 15162–15171.
52. Du, D.; Zhu, P.; Wen, L.; Bian, X.; Ling, H.; Hu, Q.; Peng, T.; Zheng, J.; Wang, X.; Zhang, Y.; et al. VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; pp. 213–226.
53. Yu, X.; Gong, Y.; Jiang, N.; Ye, Q.; Han, Z. Scale Match for Tiny Person Detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), Snowmass Village, CO, USA, 1–5 March 2020; pp. 1257–1265.

54. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 618–626.
55. Xie, Z.; Li, C.; Yang, Z.; Zhang, Z.; Jiang, J.; Guo, H. YOLOv5s-BiPCNeXt, a lightweight model for detecting disease in eggplant leaves. *Plants* **2024**, *13*, 2303. [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.