

Article

SFSMamba-DETR: Selective Feature Scanning with State Space Models and Dual-Scale Window Attention for Remote Sensing Object Detection

Yuanli Cai ^{1,*} , Junchao Zhao ¹ , Husheng Wu ²  and Rui Ma ¹

¹ School of Automation Science and Engineering, Xi'an Jiaotong University, Xi'an 710049, China; zjames@stu.xjtu.edu.cn (J.Z.); 3525236003@stu.xjtu.edu.cn (R.M.)

² College of Equipment Support and Management, Engineering University of PAP, Xi'an 710086, China; wuhusheng0421@163.com

* Correspondence: ylicai@mail.xjtu.edu.cn

Highlights

What are the main findings?

- SFSMamba-DETR integrates Selective Feature Scanning, Dual-Scale Window Attention, and hierarchical cross-scale feature aggregation to address scale variation, complex backgrounds, and densely distributed small objects in remote sensing imagery.
- On MAR20, UCAS-AOD, and the Jilin-1 Satellite Aircraft Detection Dataset, the proposed method achieved mAP@0.5 scores of 85.7%, 97.6%, and 82.3%, respectively, while maintaining an inference speed of 48.6 FPS on a single NVIDIA A100 GPU.

What is the implication of the main finding?

- The results indicate that selective state space scanning can efficiently model long-range, cross-scale dependencies without relying on additional dense global self-attention.
- The proposed framework provides an effective approach to horizontal-bounding-box detection of small, densely distributed objects in high-resolution remote sensing images with cluttered backgrounds.

Abstract

Object detection in remote sensing imagery remains challenging due to vast scale variations, complex backgrounds, and the prevalence of small, densely packed targets. Existing CNN-based detectors are limited by restricted receptive fields, while Transformer-based methods incur prohibitive computational overhead for high-resolution inputs. In this paper, we propose **SFSMamba-DETR**, a detection framework that integrates state space models with Dual-Scale Window Attention for efficient and accurate remote sensing object detection. Specifically, we design a *Selective Feature Scanning* (SFS) module that uses the Mamba-based 2D Selective Scan mechanism to model long-range spatial dependencies with linear computational complexity. To capture both fine-grained local patterns and broader contextual cues simultaneously, we introduce a *Dual-Scale Window Attention* (DSWA) mechanism that operates at two complementary window scales with multi-kernel convolution bridging. These modules are orchestrated within a *Cross-scale Feature Aggregation Module* (CFAM) that performs hierarchical multi-scale fusion in a hybrid encoder. Extensive experiments on three primary benchmarks (MAR20, UCAS-AOD, and the Jilin-1 Satellite Aircraft Detection Dataset), together with supplementary results on DOTA and DIOR, demonstrate that SFSMamba-DETR achieves strong detection accuracy while maintaining competitive inference speed.



Academic Editor: Peter Hofmann

Received: 18 May 2026

Revised: 16 July 2026

Accepted: 29 July 2026

Published: 21 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 Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: remote sensing; object detection; state space model; Mamba; dual-scale attention; feature fusion; aircraft detection

1. Introduction

Remote sensing object detection has attracted increasing research attention owing to its broad applications in military reconnaissance, urban planning, environmental monitoring, and disaster response [1,2]. At a general level, remote sensing detectors must cope with large image sizes, severe scale variation, small object footprints, and cluttered backgrounds [3,4]. These challenges require efficient long-range context modeling and robust multi-scale feature fusion. Aircraft detection further introduces more specialized difficulties: aircraft are often arranged according to airport layout priors such as runways, taxiways, and aprons; fine-grained aircraft categories may differ only in subtle wing or tail structures; and shadows, hangars, runway markings, and parked vehicles can easily produce confusing local patterns. These aircraft-specific issues require both structural scene reasoning and fine local discrimination, rather than only a larger receptive field.

Convolutional Neural Network (CNN)-based detectors have achieved remarkable progress in object detection. Two-stage methods such as Faster R-CNN [5] and Cascade R-CNN [6] deliver strong accuracy through region proposal refinement, while one-stage methods including SSD [7], RetinaNet [8], and the YOLO family [9–14] prioritize real-time inference speed. Despite their success, these methods inherently rely on local convolution operations with limited receptive fields, which restricts their ability to capture the global contextual information necessary for distinguishing small targets from complex backgrounds in remote sensing imagery [15,16].

The emergence of Vision Transformers (ViT) [17] and Detection Transformers (DETR) [18] has introduced powerful global modeling capabilities through self-attention mechanisms [19]. Subsequent improvements including Deformable DETR [20], DAB-DETR [21], DINO [22], and RT-DETR [23] have progressively addressed training convergence, query design, and real-time performance. However, the quadratic computational complexity of self-attention with respect to sequence length remains a fundamental bottleneck for processing the high-resolution, large-scale imagery typical in remote sensing applications.

Recently, state space models (SSMs), particularly the Mamba architecture [24], have emerged as a compelling alternative that achieves linear computational complexity while preserving the ability to model long-range dependencies. Building on the foundational Structured State Space (S4) model [25], Mamba introduces a selective scanning mechanism that dynamically adjusts its parameters based on input content, enabling efficient and context-aware sequence modeling. Several works have explored adapting Mamba for visual recognition, such as VMamba [26] and Vision Mamba [27]. However, directly applying standard Mamba scanning patterns to remote sensing detection is suboptimal, as the standard unidirectional or bidirectional scans fail to fully exploit the inherent multi-scale spatial structure of remote sensing targets.

A closer examination of these challenges reveals an underlying tension between three competing requirements that no single existing paradigm fully resolves. Effective aircraft detection in remote sensing imagery simultaneously demands (i) *global contextual reasoning* to associate aircraft with airfield layout cues such as runways, taxiways, and aprons, (ii) *fine-grained local sensitivity* to recover sub-meter targets and distinguish subtle aircraft structures, and (iii) *efficient inference* on imagery of substantially higher resolution than typical natural-scene benchmarks. The proposed SFS module is therefore designed for the first requirement,

injecting cross-scale context into the state space scanning process; DSWA is designed for the second requirement, preserving local shape details at complementary window scales; and CFAM is designed for the third requirement, organizing the two modules in an efficient hierarchical encoder. Pure CNN backbones satisfy (iii) but fall short on (i); Transformer-based detectors satisfy (i) and (ii) but quadratic attention compromises (iii); and recent vision Mamba variants restore (i) and (iii) but typically operate on a single feature stream and treat cross-scale fusion as an afterthought handled by an FPN-style decoder. As a result, multi-scale information is propagated only through summation or concatenation *after* the long-range modeling stage, rather than being injected *into* the scanning dynamics themselves. We argue that closing this gap requires rethinking how scale, context, and locality interact within a single backbone.

Motivated by these observations, we propose **SFSMamba-DETR**, a novel framework that bridges the efficiency of state space models with the discriminative power of multi-scale attention for remote sensing object detection. The core of our approach lies in three closely coupled innovations:

1. **Selective Feature Scanning (SFS) Module:** We design a tri-input Mamba-based module that takes main, guide, and auxiliary feature maps from different scales. Through a Vision State Space (VSS) module with 2D Selective Scan (SS2D), SFS enables adaptive long-range dependency modeling with linear complexity, while using cross-scale guidance to enhance feature discrimination.
2. **Dual-Scale Window Attention (DSWA) Mechanism:** We introduce a dual-branch attention module that partitions features into small and large windows simultaneously. A Dual Local–Global Perception Attention (DLGPA) mechanism with multi-kernel convolution bridges the two scales, capturing both fine-grained details critical for small aircraft detection and broader contextual patterns for suppressing background clutter.
3. **Cross-scale Feature Aggregation Module (CFAM):** We design a hierarchical feature fusion architecture within the hybrid encoder that orchestrates SFS modules and learnable fusion blocks across Feature Pyramid Network (FPN) levels, enabling systematic information exchange between scales.

We validate SFSMamba-DETR on three challenging primary benchmarks with complementary imaging conditions: MAR20 [28] (military aircraft with 20 fine-grained categories), UCAS-AOD [29] (aerial aircraft and vehicles), and the Jilin-1 Satellite Aircraft Detection Dataset [30] (real satellite imagery). Extensive experiments show that the proposed framework achieves strong performance across these benchmarks, outperforming 22 competitive baselines including recent methods such as YOLOv10 [13], YOLO11 [14], DINO [22], and RT-DETR [23]. To further examine category diversity, we also report supplementary HBB results on DOTA and DIOR, covering ships, bridges, storage tanks, buildings, and other remote sensing categories. The source code is publicly available at <https://github.com/Junchao-XJTU/SFSMamba-DETR> (accessed on 16 July 2026).

2. Related Work

2.1. CNN-Based Object Detection

CNN-based object detection has evolved along two main paradigms. Two-stage detectors, initiated by R-CNN and refined through Faster R-CNN [5] and Cascade R-CNN [6], generate region proposals followed by classification and regression. Feature Pyramid Networks (FPNs) [31] further enhanced multi-scale detection by constructing top-down feature hierarchies. One-stage detectors such as SSD [7], RetinaNet [8], FCOS [32], and CenterNet [33] eliminate the proposal generation stage for faster inference. Gold-YOLO [34]

proposed a gather-and-distribute mechanism for efficient multi-level feature fusion, and YOLO-World [35] extended the YOLO paradigm to open-vocabulary scenarios. The YOLO series has been particularly influential, with YOLOv7 [10] introducing efficient layer aggregation, YOLOv8 [11] adopting an anchor-free design, YOLOv9 [12] proposing programmable gradient information, YOLOv10 [13] achieving NMS-free end-to-end detection, and YOLO11 [14] further improving the accuracy-speed trade-off. Despite their remarkable performance on natural images, these methods are constrained by limited receptive fields when applied to large-scale remote sensing imagery with extreme scale variations.

2.2. Transformer-Based Detection

DETR [18] pioneered end-to-end object detection by formulating it as a set prediction problem using Transformer encoders and decoders with bipartite matching. Deformable DETR [20] addressed the slow convergence by introducing deformable attention that attends to a sparse set of sampling points. DAB-DETR [21] improved query formulation through dynamic anchor boxes, while DINO [22] combined denoising training with improved anchor initialization. Co-DETR [36] introduced collaborative hybrid assignments, Group DETR [37] accelerated training via group-wise one-to-many assignment, Diffusion-Det [38] formulated detection as a denoising diffusion process, and Lite DETR [39] proposed interleaved multi-scale encoding for efficiency. RT-DETR [23] achieved real-time performance through a hybrid encoder design with AIFI (Attention-based Intra-scale Feature Interaction) and CCFM (CNN-based Cross-scale Feature-fusion Module). Our work builds upon the RT-DETR framework but fundamentally redesigns the cross-scale fusion paradigm by introducing Mamba-based selective scanning and dual-scale window attention.

2.3. Multi-Scale Feature Fusion in Detection

Multi-scale feature fusion has been a long-standing focus in object detection, since aircraft and other remote sensing targets exhibit large variations in pixel size across imaging altitudes. The Feature Pyramid Network (FPN) [31] pioneered top-down lateral connections to combine high-resolution shallow features with semantically rich deep features. PANet [40] extended this design with an additional bottom-up path that strengthens the localization signal at deep levels, and BiFPN [41] introduced learnable per-edge weights together with repeated bidirectional fusion. More recent works further enrich the fusion topology: Gold-YOLO [34] replaces FPN/PAN with a gather-and-distribute mechanism that performs global feature aggregation followed by selective re-injection across scales, while RT-DETR [23] adopts a hybrid encoder in which intra-scale attention (AIFI) and a CNN-based cross-scale fusion module (CCFM) are decoupled to balance accuracy and latency. In remote sensing, multi-scale fusion has been revisited specifically to address dense small-target distributions and large background context [16,42]. Despite their effectiveness, the above schemes mainly perform fusion through linear combinations or local convolutions *after* feature extraction, leaving long-range cross-scale dependencies underexploited. Our CFAM differs in two key aspects: (i) it embeds cross-scale guidance *inside* the state space scanning loop via the tri-input SFS module, so multi-scale evidence is fused while sequential dependencies are being modeled; and (ii) it pairs the scanning-based fusion with a Dual-Scale Window Attention that recovers fine-grained spatial detail without resorting to dense global attention.

2.4. State Space Models for Vision

State space models (SSMs) originate from control theory and have been adapted for deep sequence modeling. The S4 model [25] demonstrated that Structured State Spaces with HiPPO initialization can efficiently model long-range dependencies. Mamba [24] further introduced input-dependent selective scanning that enables dynamic, content-aware

processing with linear complexity. In the vision domain, VMamba [26] adapted Mamba for image classification through Cross-Scan mechanisms, while Vision Mamba [27] proposed bidirectional state space modeling for visual representation learning. Recent efficient variants, including EfficientVMamba [43] and PlainMamba [44], further improve visual scanning continuity and computational efficiency. RSMamba [45] explored SSMs for remote sensing image classification, while Mamba-in-Mamba [46] proposed centralized cross-scan for hyperspectral image classification. CgMamba [47] demonstrated the effectiveness of state space models for fine-grained visual recognition. In the detection domain, MambaY-OLO [48] integrated SSM blocks into the YOLO framework. These studies motivate the use of SSMs in visual detection, but most of them either focus on image-level recognition or insert Mamba blocks into a single feature stream. They therefore do not directly address the multi-level feature interaction required by dense remote sensing detection. Our SFS module addresses this gap by designing a tri-input scanning mechanism that incorporates cross-scale feature guidance within the state space framework.

2.5. Object Detection in Remote Sensing

Remote sensing object detection has received growing attention with the development of dedicated benchmarks [1,3,4] and specialized methods. Oriented R-CNN [49] and RoI Transformer [50] handle oriented bounding boxes common in aerial imagery. R3Det [51] and ReDet [52] further address rotation-equivariant detection. LSKNet [53] introduced large selective kernels to dynamically adjust receptive fields. Recent works have also explored multi-modal fusion for remote sensing, including infrared-visible approaches [54,55] and multi-scale UAV detection [16]. Lightweight attention mechanisms have also been investigated for efficient remote sensing processing [56]. Multi-scale feature fusion has also been revisited for remote sensing detection [42]. DETR-based approaches have shown strong potential for remote sensing [57], motivating our design of SFSMamba-DETR that combines Mamba's efficiency with DETR's end-to-end formulation.

3. Proposed Method

3.1. Overall Architecture

The overall architecture of SFSMamba-DETR is illustrated in Figure 1. Given an input remote sensing image $I \in \mathbb{R}^{H \times W \times 3}$, our framework consists of four main components: (1) an EfficientNet [58] backbone for multi-scale feature extraction; (2) a hybrid encoder incorporating Attention-based Intra-scale Feature Interaction (AIFI) and our proposed Cross-scale Feature Aggregation Module (CFAM); (3) a Dual-Scale Window Attention (DSWA) module for multi-scale spatial reasoning; and (4) a decoder with IoU-aware query selection and separable dynamic decoder heads.

The EfficientNet backbone extracts a pyramid of feature maps $\{P_2, P_3, P_4, P_5\}$ at strides $\{4, 8, 16, 32\}$ respectively. We adopt EfficientNet for its optimal accuracy–efficiency trade-off achieved through compound scaling of depth, width, and resolution. The highest-level feature P_5 is first processed by the AIFI module, which applies multi-head self-attention to model intra-scale interactions within the most semantically rich feature level:

$$F_5 = \text{AIFI}(P_5) = \text{MHSA}(P_5 + \text{PE}) + P_5, \quad (1)$$

where $\text{MHSA}(\cdot)$ denotes multi-head self-attention and PE is the sinusoidal positional encoding.

The enriched feature F_5 is then fed into our CFAM along with $\{P_2, P_3, P_4\}$ for cross-scale aggregation. The aggregated multi-scale features are further refined by the DSWA module. Finally, the fused features are concatenated and passed to the decoder, which

employs IoU-aware query selection [23] to initialize object queries, followed by a separable dynamic decoder that produces classification scores, bounding box coordinates, and IoU predictions.

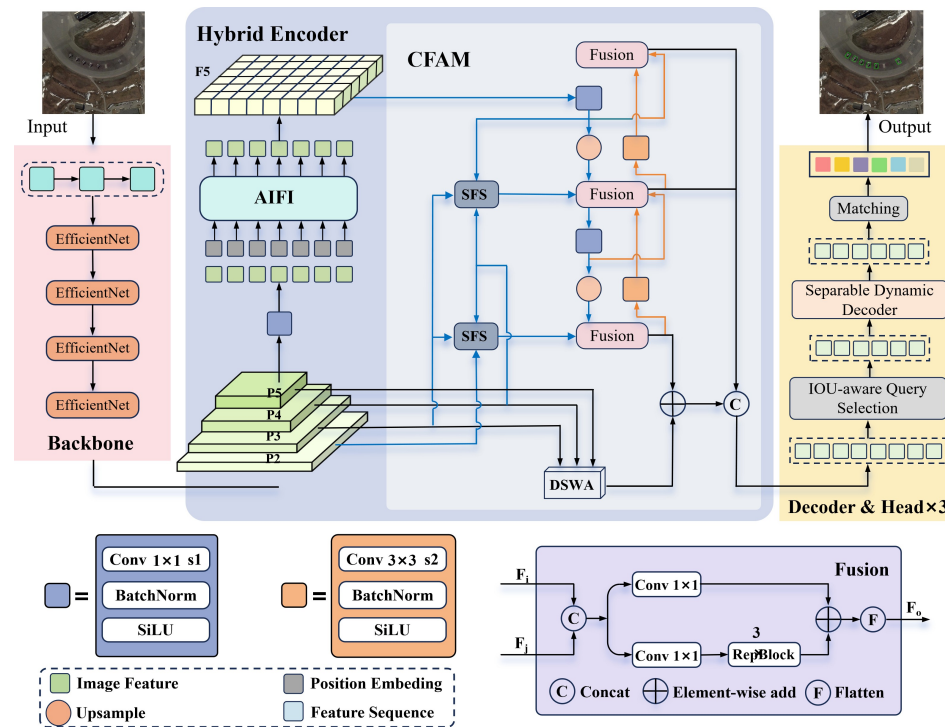


Figure 1. Overall architecture of the proposed SFSMamba-DETR. The framework comprises an EfficientNet backbone, a hybrid encoder with AIFI and CFAM modules, a DSWA module for dual-scale spatial reasoning, and a decoder with IoU-aware query selection. **SFS**: Selective Feature Scanning module. **DSWA**: Dual-Scale Window Attention.

3.2. Selective Feature Scanning Module

The Selective Feature Scanning (SFS) module is the core component enabling efficient long-range dependency modeling across multiple feature scales. Unlike standard attention mechanisms that incur quadratic complexity, or conventional Mamba blocks designed for single-input sequences, our SFS module accepts three distinct inputs (main, guide, and auxiliary features), enabling cross-scale information exchange within the state space framework. The detailed structure is shown in Figure 2.

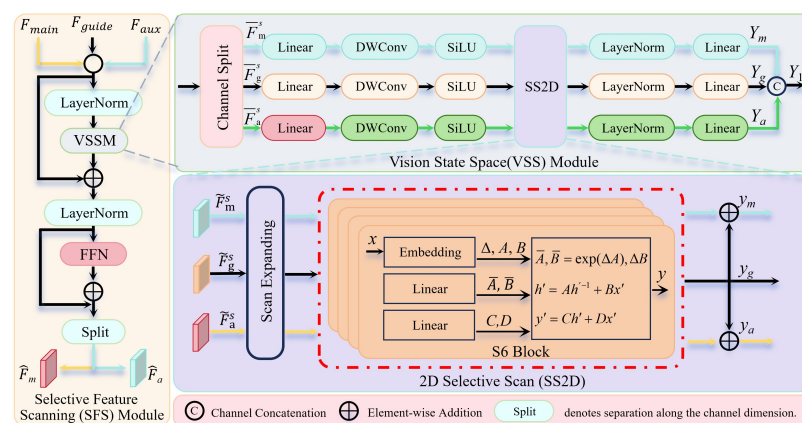


Figure 2. Detailed architecture of the Selective Feature Scanning (SFS) module. **Left**: The overall SFS pipeline with tri-input design (main, guide, auxiliary features). **Top-right**: The Vision State Space (VSS) module with Channel Split and parallel processing branches. **Bottom-right**: The 2D Selective Scan (SS2D) mechanism based on the S6 block.

3.2.1. Tri-Input Design Rationale

In the CFAM hierarchy, features at different FPN levels carry complementary information: lower-level features preserve fine spatial details while higher-level features encode rich semantic context. Our tri-input design explicitly models this complementarity. Given main feature $F_{\text{main}} \in \mathbb{R}^{H' \times W' \times C}$, guide feature $F_{\text{guide}} \in \mathbb{R}^{H' \times W' \times C}$, and auxiliary feature $F_{\text{aux}} \in \mathbb{R}^{H' \times W' \times C}$ (spatially aligned through interpolation), the SFS module first normalizes and concatenates them along the channel dimension:

$$F_{\text{cat}} = [F_{\text{main}}; F_{\text{guide}}; F_{\text{aux}}] \in \mathbb{R}^{H' \times W' \times 3C}. \quad (2)$$

To avoid passing a tripled channel dimension directly into the SSM, F_{cat} is projected by a shared linear bottleneck and then split into three same-width branches:

$$[\bar{F}_m^s, \bar{F}_g^s, \bar{F}_a^s] = \text{Split}(\text{Linear}_{3C \rightarrow 3C_s}(\text{LN}(F_{\text{cat}}))), \quad (3)$$

where $\bar{F}_m^s, \bar{F}_g^s, \bar{F}_a^s \in \mathbb{R}^{H' \times W' \times C_s}$. In our implementation, $C_s = C$ unless otherwise specified. Thus, concatenation is used to expose the S6 gates to cross-scale evidence, while the actual selective scan is performed on fixed-width branch features, preserving dimensional consistency with the state space parameters.

3.2.2. Vision State Space Module

Each sub-feature is independently processed through a sequence of Linear projection, Depthwise Convolution (DWConv), and SiLU activation:

$$\hat{F}_i^s = \text{SiLU}(\text{DWConv}(\text{Linear}(\bar{F}_i^s))), \quad i \in \{m, g, a\}. \quad (4)$$

The activated features are then fed into the 2D Selective Scan (SS2D) module. Based on the selective state space mechanism from Mamba [24], SS2D extends the one-dimensional selective scan to two-dimensional feature maps through a four-directional scanning strategy (left-to-right, right-to-left, top-to-bottom, bottom-to-top).

3.2.3. 2D Selective Scan Mechanism

The continuous-time state space model is defined by the following ordinary differential equation:

$$h'(t) = Ah(t) + Bx(t), \quad y(t) = Ch(t) + Dx(t), \quad (5)$$

where $A \in \mathbb{R}^{N \times N}$ is the state transition matrix for an N -dimensional hidden state, $B \in \mathbb{R}^{N \times 1}$ and $C \in \mathbb{R}^{1 \times N}$ are input and output projection matrices, $D \in \mathbb{R}$ is the feedthrough parameter, and $h(t) \in \mathbb{R}^N$ is the hidden state. For a visual branch with C_s channels, this formulation is applied channel-wise: each channel owns an N -dimensional state, so the implementation can be viewed as storing $\{A_c, B_{t,c}, C_{t,c}, \Delta_{t,c}\}_{c=1}^{C_s}$. Equivalently, practical Mamba implementations store a diagonal or log-parameterized version of A with shape $C_s \times N$, while B_t and C_t are generated with shape $C_s \times N$ for each spatial token. Through zero-order hold (ZOH) discretization with step size Δ , the discrete-time formulation becomes

$$\bar{A} = \exp(\Delta A), \quad \bar{B} = (\Delta A)^{-1}(\exp(\Delta A) - I) \cdot \Delta B, \quad (6)$$

$$h_t = \bar{A}h_{t-1} + \bar{B}x_t, \quad y_t = Ch_t + Dx_t. \quad (7)$$

The key innovation of Mamba's S6 (Selective S4) block is making B , C , and Δ input-dependent:

$$B_t = \text{Linear}_B(x_t), \quad C_t = \text{Linear}_C(x_t), \quad \Delta_t = \text{Softplus}(\text{Linear}_\Delta(x_t)). \quad (8)$$

In SFS, the token x_t is produced after the tri-scale projection in Equation (3). Consequently, the selective parameters are conditioned not only on the current-level feature but also on guide and auxiliary evidence. A larger $\Delta_{t,c}$ allows the corresponding channel to update its hidden state more aggressively when a spatial location is supported by cross-scale cues, whereas a smaller value retains longer contextual memory when the local evidence is ambiguous. The input-dependent $B_{t,c}$ controls how strongly the current multi-scale token enters the state, and $C_{t,c}$ controls how much of the hidden state is exposed to the output. This gives the S6 block a concrete mechanism for prioritizing information across the three heterogeneous scales instead of treating concatenation as a purely static fusion step.

For 2D feature maps, we flatten each spatial feature into four directional sequences and apply the S6 block independently, then merge the outputs:

$$Y_{\text{SS2D}} = \frac{1}{4} \sum_{d=1}^4 S6_d(\text{Scan}_d(\hat{F}^s)), \quad (9)$$

where $\text{Scan}_d(\cdot)$ denotes the d -th directional scanning operation.

After SS2D, each branch output passes through LayerNorm and Linear projection to produce Y_m , Y_g , and Y_a . The main and guide outputs are concatenated as $Y_1 = [Y_m; Y_g]$, followed by a Feed-Forward Network (FFN). The Selective Feature Scanning module then performs a channel-wise Split operation that separates the processed features into refined main feature $\tilde{F}_m$ and auxiliary feature $\tilde{F}_a$, which serve as inputs to subsequent fusion stages.

3.3. Dual-Scale Window Attention

While the SFS module excels at capturing long-range dependencies with linear complexity, local spatial patterns at multiple scales are equally critical for detecting objects with diverse sizes in remote sensing imagery. We propose the Dual-Scale Window Attention (DSWA) mechanism, illustrated in Figure 3, which operates at two complementary window scales to simultaneously capture fine-grained local details and broader contextual patterns.

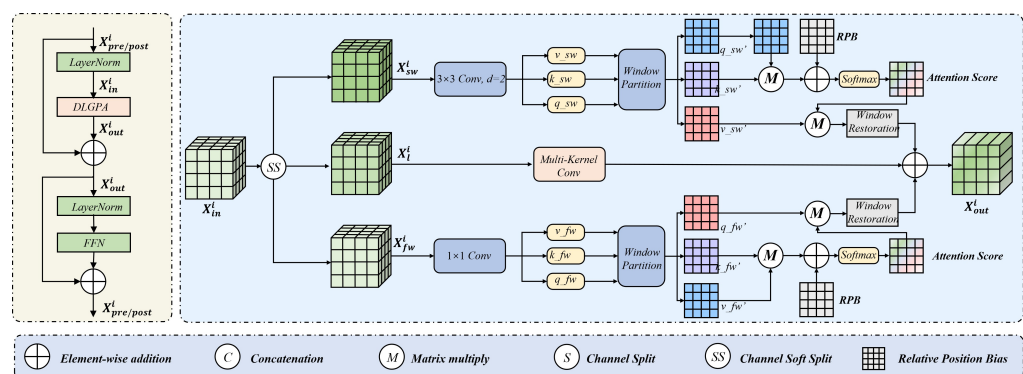


Figure 3. Architecture of the Dual-Scale Window Attention (DSWA) mechanism. **(Left):** Overall flow with LayerNorm, DLGPA, and FFN. **(Right):** Detailed Dual Local–Global Perception Attention (DLGPA) with small-window and large-window branches, multi-kernel convolution bridge, and relative position bias (RPB); SS denotes Channel Soft Split.

Dual Local–Global Perception Attention

Given input features $\mathbf{X}_{\text{in}}^l \in \mathbb{R}^{H' \times W' \times C}$, DSWA first applies LayerNorm and then performs a Channel Soft Split (SS) operation that softly partitions the features into a small-window branch $\mathbf{X}_{sw}^l$ and a large-window branch $\mathbf{X}_{fw}^l$, along with an intermediate representation $\mathbf{X}_t^l$:

$$\mathbf{X}_{sw}^l, \mathbf{X}_t^l, \mathbf{X}_{fw}^l = \text{ChannelSoftSplit}(\text{LN}(\mathbf{X}_{\text{in}}^l)). \quad (10)$$

Small-window branch. The small-window features are processed through a 3×3 dilated convolution (dilation rate $d = 2$) to generate queries, keys, and values. Window partitioning is then applied with a small window size s_w :

$$\mathbf{q}_{sw}, \mathbf{k}_{sw}, \mathbf{v}_{sw} = \text{Split}(\text{Conv}_{3 \times 3, d=2}(\mathbf{X}_{sw}^l)), \quad (11)$$

$$\text{Attn}_{sw} = \text{Softmax}\left(\frac{\mathbf{q}_{sw}\mathbf{k}_{sw}^\top}{\sqrt{d_k}} + \text{RPB}_{sw}\right)\mathbf{v}_{sw}, \quad (12)$$

where RPB_{sw} denotes the learnable relative position bias for the small-window scale.

Large-window branch. The large-window features use 1×1 convolution for projection and are partitioned with a larger window size s_{fw} :

$$\mathbf{q}_{fw}, \mathbf{k}_{fw}, \mathbf{v}_{fw} = \text{Split}(\text{Conv}_{1 \times 1}(\mathbf{X}_{fw}^l)), \quad (13)$$

$$\text{Attn}_{fw} = \text{Softmax}\left(\frac{\mathbf{q}_{fw}\mathbf{k}_{fw}^\top}{\sqrt{d_k}} + \text{RPB}_{fw}\right)\mathbf{v}_{fw}. \quad (14)$$

Multi-Kernel Convolution bridge. The intermediate representation $\mathbf{X}_t^l$ passes through a Multi-Kernel Conv module that applies parallel convolutions with different kernel sizes to capture multi-scale local patterns, bridging information between the two window scales.

The outputs from both branches are concatenated after Window Restoration and combined with the multi-kernel features:

$$\mathbf{X}_{\text{out}}^l = \text{WindowRestore}(\text{Attn}_{sw}) \oplus \text{WindowRestore}(\text{Attn}_{fw}) \oplus \text{MKConv}(\mathbf{X}_t^l), \quad (15)$$

where $\oplus$ denotes element-wise addition. The final output passes through a residual connection with LayerNorm and FFN:

$$\mathbf{X}_{\text{pre/post}}^l = \text{FFN}(\text{LN}(\mathbf{X}_{\text{out}}^l + \mathbf{X}_{\text{in}}^l)) + \mathbf{X}_{\text{out}}^l + \mathbf{X}_{\text{in}}^l. \quad (16)$$

3.4. Cross-Scale Feature Aggregation Module

The Cross-scale Feature Aggregation Module (CFAM) orchestrates the interaction between SFS modules and learnable fusion blocks across FPN levels within the hybrid encoder. As shown in Figure 1, CFAM operates in a top-down manner:

Stage 1. The AIFI-enriched feature $\mathbf{F}_5$ serves as the guide for processing $\mathbf{P}_4$:

$$\mathbf{F}_4^{\text{sfs}} = \text{SFS}(\mathbf{P}_4, \text{Up}(\mathbf{F}_5), \mathbf{P}_3), \quad \mathbf{F}_4 = \text{Fusion}(\mathbf{F}_4^{\text{sfs}}, \text{Down}(\mathbf{F}_5)), \quad (17)$$

where $\text{Up}(\cdot)$ and $\text{Down}(\cdot)$ denote bilinear upsampling and strided convolution downsampling respectively.

Stage 2. The fused feature F_4 guides the processing of P_3 :

$$F_3^{\text{sfs}} = \text{SFS}(P_3, \text{Up}(F_4), P_2), \quad F_3 = \text{Fusion}(F_3^{\text{sfs}}, \text{Down}(F_4)). \quad (18)$$

The **Fusion** block concatenates two input feature maps, applies 1×1 convolution for channel reduction, followed by three RepBlocks [59] for feature refinement, and adds a parallel 1×1 convolution skip connection:

$$\text{Fusion}(F_i, F_j) = \text{RepBlock}^3(\text{Conv}_{1 \times 1}([F_i; F_j])) \oplus \text{Conv}_{1 \times 1}([F_i; F_j]), \quad (19)$$

where $[\cdot; \cdot]$ denotes channel-wise concatenation.

The DSWA module is applied to the aligned aggregated features to enhance multi-scale spatial reasoning. Specifically, the multi-scale features are first resized to a common stride and combined by element-wise addition to form a context-enhanced feature. DSWA then refines this feature, and the refined DSWA output is concatenated with the aligned aggregation before being flattened into a feature sequence for the decoder.

3.5. Loss Function and Training

Following the DETR paradigm, we employ Hungarian matching [18] for bipartite assignment between predictions and ground truth. The total training loss consists of four components:

Classification loss. We adopt Focal Loss [8] to handle class imbalance:

$$\mathcal{L}_{\text{cls}} = -\alpha_t(1 - p_t)^\gamma \log(p_t), \quad (20)$$

where α_t and γ are the balancing factor and focusing parameter, set to 0.25 and 2.0 respectively.

Bounding box regression loss. We combine ℓ_1 loss and Generalized IoU (GIoU) loss [60]:

$$\mathcal{L}_{\text{box}} = \lambda_{L1} \|\hat{\mathbf{b}} - \mathbf{b}\|_1 + \lambda_{\text{giou}} \mathcal{L}_{\text{GIoU}}(\hat{\mathbf{b}}, \mathbf{b}). \quad (21)$$

IoU-aware loss. The IoU prediction branch is supervised by Binary Cross-Entropy:

$$\mathcal{L}_{\text{iou}} = -[\text{IoU}^* \log(\hat{s}) + (1 - \text{IoU}^*) \log(1 - \hat{s})], \quad (22)$$

where $\hat{s}$ is the predicted IoU score and IoU^* is the actual IoU between predicted and ground-truth boxes.

The total loss is

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{L1} + \lambda_3 \mathcal{L}_{\text{GIoU}} + \lambda_4 \mathcal{L}_{\text{iou}}, \quad (23)$$

where $\lambda_1 = 2.0$, $\lambda_2 = 5.0$, $\lambda_3 = 2.0$, and $\lambda_4 = 1.0$.

Algorithm 1 formalizes the end-to-end training pipeline. After backbone feature extraction and AIFI enhancement on P_5 (lines 4–5), CFAM operates in a top-down manner where each SFS call receives a tri-input tuple of current-level, deeper-level (upsampled), and lower-level features (lines 6 and 8), so that long-range dependency modeling and cross-scale information exchange take place *jointly* inside the state space scanning loop rather than as two separate stages. The aligned multi-scale features are then added, refined by DSWA, concatenated with the DSWA output, and passed to the decoder (lines 10–12). The predictions are supervised by the joint loss in Equation (23) and optimized end-to-end with cosine learning-rate annealing (lines 15–19).

Algorithm 1 Training procedure of SFSMamba-DETR.

Input: Training dataset $\mathcal{D} = \{(I_i, Y_i)\}_{i=1}^N$, pretrained EfficientNet weights θ_{bb}
Output: Trained model parameters θ^*

- 1 Initialize backbone with θ_{bb} ; randomly initialize encoder, CFAM, DSWA, decoder
- 2 **for** epoch = 1 to N_{epoch} **do**
- 3 **for each** mini-batch $\{(I, Y)\} \subset \mathcal{D}$ **do**
- 4 $\{P_2, P_3, P_4, P_5\} \leftarrow \text{EfficientNet}(I)$ (*Multi-scale feature extraction*)
- 5 $F_5 \leftarrow \text{AIFI}(P_5)$ (*Intra-scale feature interaction*)
- 6 $F_4^{fs} \leftarrow \text{SFS}(P_4, \text{Up}(F_5), P_3)$ (*Selective Feature Scanning*)
- 7 $F_4 \leftarrow \text{Fusion}(F_4^{fs}, \text{Down}(F_5))$
- 8 $F_3^{fs} \leftarrow \text{SFS}(P_3, \text{Up}(F_4), P_2)$
- 9 $F_3 \leftarrow \text{Fusion}(F_3^{fs}, \text{Down}(F_4))$
- 10 $F_{add} \leftarrow F_3 \oplus \text{Align}(F_4) \oplus \text{Align}(F_5)$ (*Element-wise aggregation*)
- 11 $F_{dswa} \leftarrow \text{DSWA}(F_{add})$ (*Dual-scale attention*)
- 12 $F_{enc} \leftarrow \text{Concat}(F_{add}, F_{dswa})$ (*Encoder output*)
- 13 $\mathcal{Q} \leftarrow \text{IoUAwareSelection}(F_{enc})$ (*Query initialization*)
- 14 $\{\hat{c}, \hat{b}, \hat{s}\} \leftarrow \text{Decoder}(\mathcal{Q}, F_{enc})$ (*Prediction*)
- 15 $\sigma^* \leftarrow \text{HungarianMatch}(\{\hat{c}, \hat{b}\}, Y)$ (*Bipartite matching*)
- 16 Compute $\mathcal{L}_{total}$ via Equation (23)
- 17 Update θ by $\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}_{total}$
- 18 **end for**
- 19 Apply learning rate scheduling (cosine annealing)
- 20 **end for**
- 21 **return** $\theta^* = \theta$

4. Experiments

4.1. Datasets

We evaluate SFSMamba-DETR on three remote sensing aircraft detection benchmarks that span complementary imaging conditions, from high-resolution Google Earth imagery to sub-meter satellite observations. Their key statistics are summarized in Table 1.

MAR20 [28] is the largest publicly available military aircraft recognition benchmark, released by Northwestern Polytechnical University. It contains 3842 high-resolution optical images collected from 60 military airports worldwide via Google Earth, covering 20 fine-grained aircraft models (e.g., A-10, B-2, F-15, F/A-18) with 22,341 instances annotated by both horizontal and oriented bounding boxes. The benchmark is challenging due to high inter-class similarity among aircraft types and large intra-class variability caused by climate, illumination, and viewing angle. We use the HBB annotations with the official 70%/30% training/testing split. The dataset is available at <https://gcheng-nwpu.github.io/> (accessed on 16 July 2026).

UCAS-AOD [29] is a widely used aerial object detection benchmark from the University of Chinese Academy of Sciences. It comprises 1510 high-resolution aerial images of airports, parking lots, and urban areas, with two categories: airplanes (1000 images, 7482 instances) and cars (510 images, 7114 instances). Targets are annotated with oriented bounding boxes and exhibit arbitrary orientations, dense arrangements, and substantial scale variation. We adopt the standard 5-fold cross-validation protocol. The dataset is available at <https://github.com/Lbx2020/UCAS-AOD-dataset> (accessed on 16 July 2026).

The Jilin-1 Satellite Aircraft Detection Dataset [30] consists of panchromatic satellite imagery acquired by the Jilin-1 commercial constellation operated by Chang Guang Satellite Technology Co., Ltd., Changchun, China, with a ground sampling distance of approximately 0.75 m. The dataset contains 1286 image patches covering civilian and military airports worldwide, with 8547 aircraft instances annotated by horizontal bounding boxes. Compared with the airborne MAR20 and UCAS-AOD benchmarks, the satellite imagery introduces substantially more challenging conditions, including extremely small target sizes (median aircraft area <600 pixels), low foreground–background contrast in single-channel panchromatic data, heterogeneous terrain (grassland, desert, snow, urban), and occasional cloud and shadow interference. We use a 70%/30% training/testing split.

The dataset is available from Chang Guang Satellite Technology Co., Ltd. upon reasonable request at <https://www.jl1mall.com/> (accessed on 16 July 2026).

Although MAR20 and UCAS-AOD also provide oriented bounding box annotations, this study evaluates the horizontal bounding box setting for three reasons. First, the base detector follows the RT-DETR horizontal-box formulation, which allows a controlled comparison against the strongest HBB baselines. Second, the Jilin-1 dataset is annotated with horizontal boxes only, and a unified HBB protocol avoids mixing detection heads and evaluation metrics across datasets. Third, the phrase “arbitrary orientations” in this paper refers to object appearance and spatial arrangement in overhead imagery rather than to oriented-box regression. We agree that OBB detection is important for elongated remote sensing targets, and we discuss this extension explicitly in Section 5.

Table 1. Summary of the three remote sensing aircraft detection benchmarks used in this study. HBB and OBB denote horizontal and oriented bounding box annotations, respectively; # denotes the number of images, categories, or instances.

Dataset	Source	#Images	#Categories	#Instances	Annotation
MAR20	Google Earth (NWPU)	3842	20	22,341	HBB + OBB
UCAS-AOD	Aerial (UCAS)	1510	2	14,596	OBB
Jilin-1	Satellite (~ 0.75 m GSD)	1286	1	8547	HBB

Representative detection examples from MAR20 and UCAS-AOD are shown in Figure 4, while examples from the Jilin-1 satellite dataset are illustrated in Figure 5.

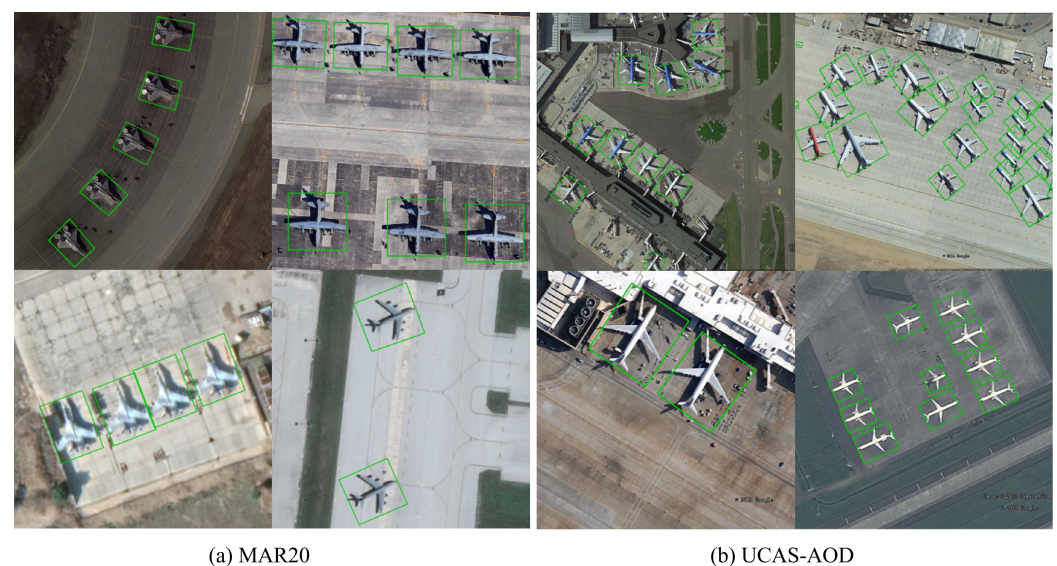
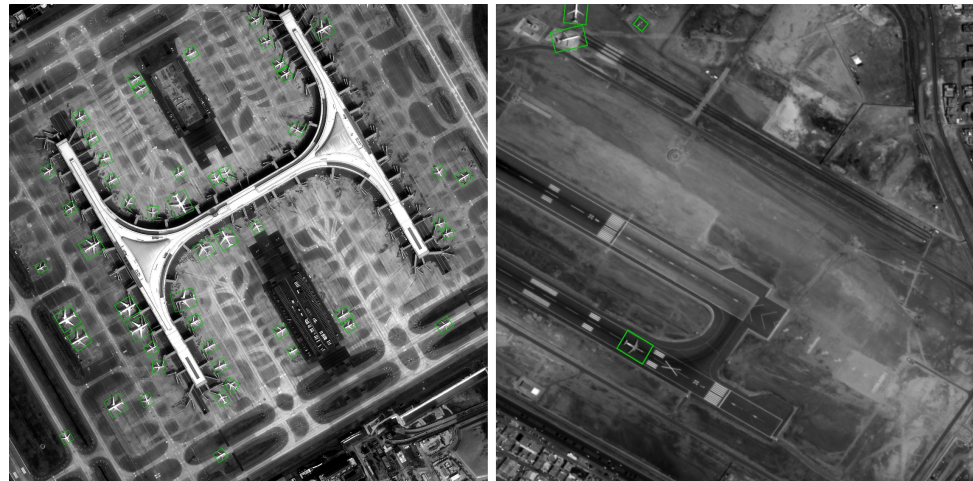


Figure 4. Detection examples from (a) MAR20 and (b) UCAS-AOD datasets. Green bounding boxes denote ground-truth annotations. Both datasets feature diverse aircraft types and complex airport backgrounds.

4.2. Implementation Details

All experiments are conducted on 4 NVIDIA A100 GPUs with 80 GB memory. We use EfficientNet-B4 pretrained on ImageNet as the backbone. Input images are resized to 640×640 with random horizontal/vertical flipping, random cropping, and mosaic augmentation during training. We train for 150 epochs using AdamW optimizer [61] with an initial learning rate of 1×10^{-4} , weight decay of 1×10^{-4} , and cosine annealing schedule. The batch size is set to 16. For the DSWA module, we set the small window size $s_w = 7$ and the large window size $s_{fw} = 14$. The number of decoder layers is 3, and the number of object queries is 300. For fair comparison, all reproduced baselines are trained under

the same data split, input resolution, number of epochs, augmentation policy, optimizer type, and hardware setting unless their official implementation requires an incompatible setting; in that case, we keep the official model structure but align the training schedule as closely as possible. We report mAP@0.5 and mAP@0.5:0.95 following the COCO evaluation protocol [62], along with model parameters (Params), floating-point operations (FLOPs), and frames per second (FPS) measured on a single A100 GPU.



Jilin-1 Satellite Aircraft Detection Dataset

Figure 5. Detection examples from the Jilin-1 Satellite Aircraft Detection Dataset. The panchromatic imagery presents unique challenges including low contrast, small target sizes, and complex terrain backgrounds.

4.3. Comparison with State-of-the-Art Methods

4.3.1. Results on MAR20 Dataset

Table 2 presents detailed comparison results on the MAR20 dataset. We compare with 22 representative methods spanning five categories: two-stage detectors, one-stage anchor-based detectors, anchor-free detectors, Transformer-based detectors, and remote-sensing-specific methods.

SFSMamba-DETR achieves the best mAP@0.5 of 85.7% and mAP@0.5:0.95 of 55.9%, outperforming the strongest baseline YOLO11-L by 2.3 and 2.8 absolute percentage points, respectively, and surpassing RT-DETR-L by 2.6 and 3.1 absolute percentage points on the two metrics. The proposed framework outperforms all YOLO-series variants while maintaining competitive inference speed (48.6 FPS). Compared with the heavyweight DINO (80.6%), SFSMamba-DETR yields a substantial +5.1 percentage-point improvement with significantly fewer parameters (34.8 M vs. 47.2 M). These gains suggest that Mamba-based scanning and dual-scale attention are particularly useful for fine-grained military aircraft recognition, where global airfield context and subtle local shape cues must be modeled together.

4.3.2. Results on UCAS-AOD Dataset

Table 3 reports comparison results on the UCAS-AOD dataset. SFSMamba-DETR achieves a mAP@0.5 of 97.6% and mAP@0.5:0.95 of 73.2%, with per-category performance of 98.5% for airplanes and 96.7% for cars. Compared to the second-best RT-DETR-L, the proposed detector achieves +1.4 and +2.0 absolute percentage-point gains on mAP@0.5 and mAP@0.5:0.95, respectively. The improvement is particularly notable for the car category (+1.4 percentage points over RT-DETR-L), where objects exhibit greater scale variation and denser arrangements. The corresponding visual comparison is discussed in Section 4.5.

Table 2. Comparison results on the MAR20 dataset. Best results are in **bold** and second-best are underlined. [†] denotes results reproduced with official code under the unified training setting described in Section 4.2.

Category	Method	mAP@0.5	mAP@0.5:0.95	Params (M)	FLOPs (G)	FPS
<i>Two-stage</i>	Faster R-CNN [5]	72.4	42.1	41.1	134.6	18.2
	Cascade R-CNN [6]	75.8	45.3	69.2	186.3	14.5
<i>One-stage</i>	RetinaNet [8]	70.1	40.5	36.3	119.8	21.3
	YOLOX-L [†] [63]	79.2	48.8	54.2	155.6	42.1
	YOLOv5-L [†] [9]	78.6	48.2	46.1	109.1	52.8
	YOLOv7 [†] [10]	80.3	49.8	36.5	103.2	48.3
	YOLOv8-L [†] [11]	81.5	51.2	43.6	165.2	45.6
	YOLOv9-C [†] [12]	82.1	52.0	25.3	102.1	50.1
	YOLOv10-L [†] [13]	82.8	52.5	24.4	120.3	46.8
	YOLO11-L [†] [14]	<u>83.4</u>	<u>53.1</u>	25.3	86.9	55.2
<i>Anchor-free</i>	RTMDet-L [†] [64]	81.8	51.5	52.3	160.4	40.2
	FCOS [32]	71.8	41.3	32.1	108.5	23.6
<i>Transformer</i>	CenterNet [33]	73.5	43.8	35.6	125.2	20.1
	DETR [18]	68.5	38.2	41.3	86.1	28.4
	Deformable DETR [20]	76.3	46.8	39.8	173.3	22.1
	DAB-DETR [21]	77.8	47.9	43.7	180.2	20.5
	DINO [22]	80.6	50.8	47.2	194.5	18.8
	Co-DETR [36]	81.2	51.3	48.1	198.2	17.5
	Lite DETR [39]	79.5	49.2	36.8	148.5	30.2
	RT-DETR-L [23]	83.1	52.8	32.0	110.0	54.8
<i>RS-specific</i>	Oriented R-CNN [49]	77.2	47.6	41.1	139.6	16.8
	LSKNet-S [53]	82.0	51.7	31.0	98.5	38.5
<i>Ours</i>	SFSMamba-DETR	85.7	55.9	34.8	118.5	48.6

Table 3. Comparison results on the UCAS-AOD dataset. Best results are in **bold** and second-best are underlined.

Category	Method	mAP@0.5	mAP@0.5:0.95	AP _{airplane}	AP _{car}	Params (M)	FLOPs (G)	FPS
<i>Two-stage</i>	Faster R-CNN [5]	89.3	62.5	91.2	87.4	41.1	134.6	18.2
	Cascade R-CNN [6]	91.5	65.0	93.1	89.9	69.2	186.3	14.5
<i>One-stage</i>	YOLOv5-L [9]	92.8	66.8	94.5	91.1	46.1	109.1	52.8
	YOLOv7 [10]	93.5	67.3	95.2	91.8	36.5	103.2	48.3
	YOLOv8-L [11]	94.2	68.5	95.8	92.6	43.6	165.2	45.6
	YOLOv9-C [12]	94.6	69.1	96.0	93.2	25.3	102.1	50.1
	YOLOv10-L [13]	94.9	69.3	96.2	93.6	24.4	120.3	46.8
	YOLO11-L [14]	95.3	69.6	96.5	94.1	25.3	86.9	55.2
<i>Transformer</i>	DETR [18]	85.4	58.1	87.6	83.2	41.3	86.1	28.4
	Deformable DETR [20]	91.8	65.2	93.5	90.1	39.8	173.3	22.1
	DINO [22]	93.7	67.5	95.4	92.0	47.2	194.5	18.8
	Co-DETR [36]	94.4	68.9	95.9	92.9	48.1	198.2	17.5
	RT-DETR-L [23]	<u>96.2</u>	<u>71.2</u>	<u>97.2</u>	<u>95.3</u>	32.0	110.0	54.8
<i>RS-specific</i>	LSKNet-S [53]	94.8	69.2	96.1	93.5	31.0	98.5	38.5
<i>Ours</i>	SFSMamba-DETR	97.6	73.2	98.5	96.7	34.8	118.5	48.6

4.3.3. Results on Jilin-1 Satellite Aircraft Detection Dataset

Table 4 presents results on the most challenging Jilin-1 dataset. The satellite imagery characteristics, including low contrast, small target sizes, and complex terrain, push all methods to lower performance levels, highlighting the difficulty of real-world satellite-based detection. SFSMamba-DETR achieves 82.3% mAP@0.5 and 50.7% mAP@0.5:0.95, outperforming RT-DETR-L by 3.8 and 3.9 absolute percentage points, respectively. The

improvements on AP_S (36.8% vs. 32.5%) indicate the effectiveness of SFS and DSWA for small object detection in satellite imagery. Visual comparisons are discussed in Section 4.5.

Table 4. Comparison results on the Jilin-1 Satellite Aircraft Detection Dataset. Best results are in **bold** and second-best are underlined. Params and FLOPs follow the same calculation protocol as Tables 2 and 3.

Category	Method	mAP@0.5	mAP@0.5:0.95	AP_S	Recall	Params (M)	FLOPs (G)	FPS
Two-stage	Faster R-CNN [5]	68.2	35.6	22.3	72.1	41.1	134.6	18.2
	Cascade R-CNN [6]	71.5	38.4	25.1	75.3	69.2	186.3	14.5
One-stage	YOLOv5-L [9]	73.8	41.6	27.5	77.6	46.1	109.1	52.8
	YOLOv8-L [11]	76.5	44.8	30.2	80.3	43.6	165.2	45.6
	YOLOv9-C [12]	77.3	45.6	31.1	81.2	25.3	102.1	50.1
	YOLOv10-L [13]	77.8	46.0	31.6	81.7	24.4	120.3	46.8
	YOLO11-L [14]	78.1	46.3	32.0	82.1	25.3	86.9	55.2
Transformer	DETR [18]	63.5	31.2	18.6	67.4	41.3	86.1	28.4
	Deformable DETR [20]	72.8	40.1	26.5	76.5	39.8	173.3	22.1
	DINO [22]	76.2	44.3	29.8	80.0	47.2	194.5	18.8
	Co-DETR [36]	77.0	45.2	30.8	80.8	48.1	198.2	17.5
	RT-DETR-L [23]	<u>78.5</u>	<u>46.8</u>	<u>32.5</u>	<u>82.6</u>	32.0	110.0	54.8
RS-specific	LSKNet-S [53]	77.5	45.8	31.3	81.4	31.0	98.5	38.5
Ours	SFSMamba-DETR	82.3	50.7	36.8	86.2	34.8	118.5	48.6

4.3.4. Supplementary Results on Multi-Category Benchmarks

To examine whether the proposed design is limited to aircraft- and vehicle-centered scenes, we further conduct supplementary HBB experiments on two multi-category remote sensing benchmarks, DOTA [3] and DIOR [1]. DOTA annotations are converted to horizontal bounding boxes for consistency with the detector head used in this paper. As shown in Table 5, SFSMamba-DETR maintains consistent gains over strong one-stage and DETR-based baselines on both datasets. The improvement is smaller than on the primary aircraft benchmarks, which is reasonable because DOTA and DIOR contain more diverse object shapes and scene contexts. Nevertheless, the results indicate that the SFS and DSWA modules also benefit broader remote sensing categories such as ships, bridges, storage tanks, expressway service areas, and buildings.

Table 5. Supplementary HBB results on multi-category remote sensing benchmarks. DOTA OBB annotations are converted to HBB annotations for consistency with the detector head. Bold values indicate the best performance within each dataset.

Dataset	Method	mAP@0.5	mAP@0.5:0.95	FPS
DOTA-v1.0 (HBB)	YOLO11-L [14]	73.1	45.6	49.7
	RT-DETR-L [23]	72.8	45.1	48.9
	SFSMamba-DETR	74.8	47.0	43.6
DIOR	YOLO11-L [14]	77.0	49.2	51.4
	RT-DETR-L [23]	77.2	49.6	50.1
	SFSMamba-DETR	78.9	51.1	44.8

4.4. Ablation Studies

4.4.1. Module Effectiveness

To verify the contribution of each proposed component, we conduct systematic ablation experiments on the Jilin-1 dataset with RT-DETR-L as the baseline. Results are reported in Table 6.

Table 6. Progressive ablation study on the effectiveness of the main modules on the Jilin-1 dataset. ✓ indicates that the module is enabled, and × indicates that the module is disabled.

SFS	DSWA	CFAM	EfficientNet	mAP@0.5	mAP@0.5:0.95
×	×	×	×	78.5	46.8
×	×	✓	×	79.0	47.4
×	×	×	✓	79.2	47.6
✓	×	×	×	80.1	48.5
×	✓	×	×	79.8	48.1
✓	✓	×	×	81.0	49.3
✓	✓	✓	×	81.5	49.8
✓	✓	✓	✓	82.3	50.7

Starting from the RT-DETR-L baseline (78.5% mAP@0.5), adding CFAM alone improves mAP@0.5 to 79.0%, indicating that explicit cross-scale aggregation is already beneficial even without the proposed scanning and attention modules. Replacing the original backbone with EfficientNet-B4 alone yields 79.2%, showing that the backbone contributes a moderate but not dominant gain. Adding the SFS module yields a larger +1.6 percentage-point improvement, demonstrating the benefit of Mamba-based long-range dependency modeling for satellite image detection. The DSWA module independently contributes +1.3 percentage points, confirming that Dual-Scale Window Attention effectively captures multi-scale local patterns. Combining SFS and DSWA achieves +2.5 percentage points, indicating their complementary nature: SFS provides efficient global scanning while DSWA enhances local multi-scale perception. CFAM further adds +0.5 percentage points by coordinating cross-scale feature aggregation, and replacing the HGNetv2 backbone with EfficientNet-B4 yields an additional +0.8 percentage points, reaching the final 82.3% mAP@0.5. The progressive improvements are visually validated through the heatmap visualization in Figure 6.

4.4.2. Effect of the Number of SFS Inputs

To further isolate the tri-input design, we compare SFS variants using one, two, and three input streams while keeping the remaining detector unchanged. As shown in Table 7, the single-input variant improves over the RT-DETR-L baseline but is weaker than the cross-scale variants, since it performs selective scanning only on the current FPN level. Adding the guide feature provides additional semantic context and improves mAP@0.5 from 79.3% to 79.7%. The full tri-input design achieves the best performance of 80.1% mAP@0.5 and 48.5% mAP@0.5:0.95, supporting the use of main, guide, and auxiliary features in SFS.

Table 7. Ablation study on the number of input streams in the SFS module on the Jilin-1 dataset. ✓ indicates that the input stream is used, and × indicates that it is not used.

Main	Guide	Auxiliary	mAP@0.5	mAP@0.5:0.95
✓	×	×	79.3	47.6
✓	✓	×	79.7	48.0
✓	✓	✓	80.1	48.5

4.4.3. Effect of Window Sizes in DSWA

We investigate the impact of window size configurations in the DSWA module. Table 8 shows that the combination of small window size $s_w = 7$ and large window size $s_{fw} = 14$ achieves the best performance. Smaller windows alone miss broader context, while larger windows alone lose fine-grained details, confirming the necessity of dual-scale design.

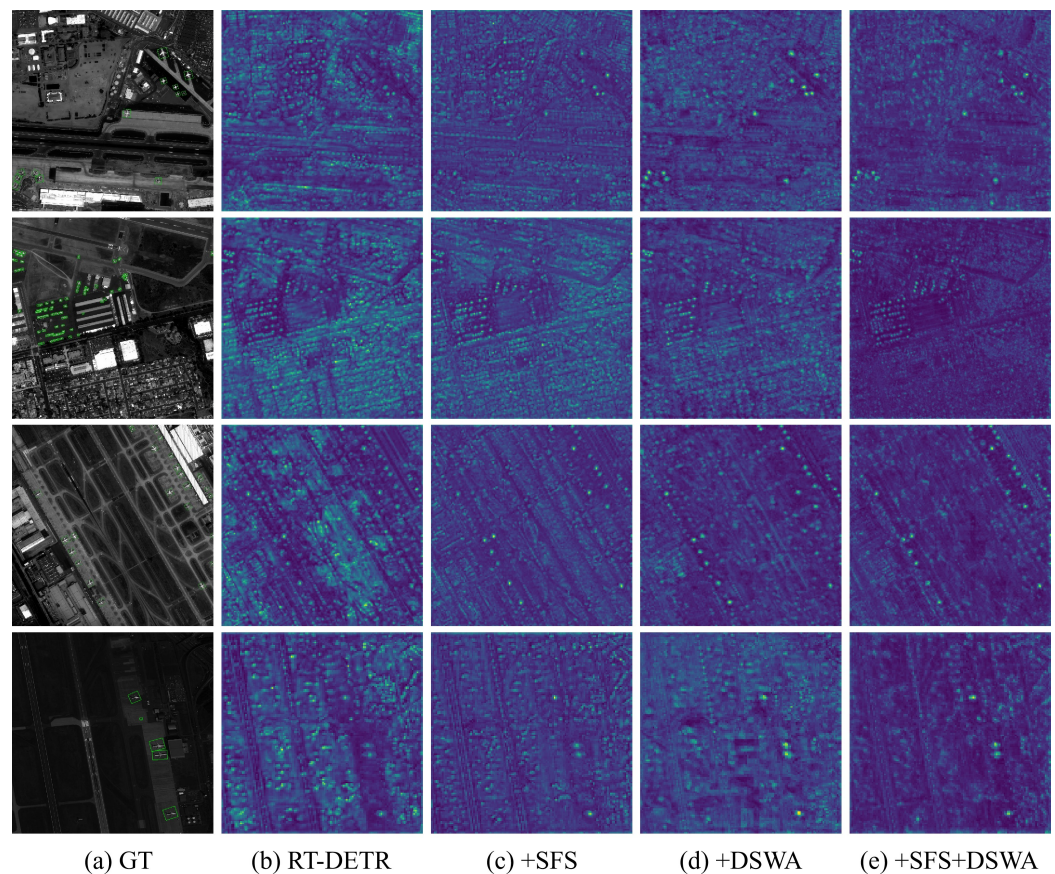


Figure 6. Heatmap visualization of ablation experiments on the Jilin-1 dataset. From left to right: (a) Ground truth, (b) RT-DETR baseline, (c) +SFS, (d) +DSWA, (e) +SFS+DSWA. The progressive enhancement in target localization and background suppression validates the contribution of the proposed modules.

Table 8. Ablation study on window size configurations in DSWA on the Jilin-1 dataset.

Small Window s_w	Large Window s_{fw}	mAP@0.5	mAP@0.5 :0.95	FPS
5	10	81.4	49.8	51.2
5	14	81.7	50.1	49.8
7	14	82.3	50.7	48.6
7	21	82.0	50.3	44.1
14	28	81.2	49.5	40.3

4.4.4. Effect of Scanning Directions in SS2D

Table 9 examines the number of scanning directions in the 2D Selective Scan module. The four-directional scanning (horizontal and vertical, both forward and backward) achieves the best mAP@0.5 of 82.3%, as it captures spatial dependencies from all orientations, which is critical for aircraft targets that appear at arbitrary orientations in remote sensing imagery.

Table 9. Effect of scanning directions in SS2D on the Jilin-1 dataset.

Directions	Description	mAP@0.5	mAP@0.5 :0.95
1	Left-to-right only	80.5	48.3
2	Horizontal bidirectional	81.2	49.2
2	Vertical bidirectional	81.0	49.0
4	All four directions	82.3	50.7

4.5. Qualitative Analysis

4.5.1. Detection Visualization

Figures 7–9 present visual comparisons of detection results across the three datasets. We compare SFSMamba-DETR against three representative baselines spanning two-stage CNN (Faster R-CNN), one-stage anchor-free (YOLO11), and Transformer-based (RT-DETR) paradigms.

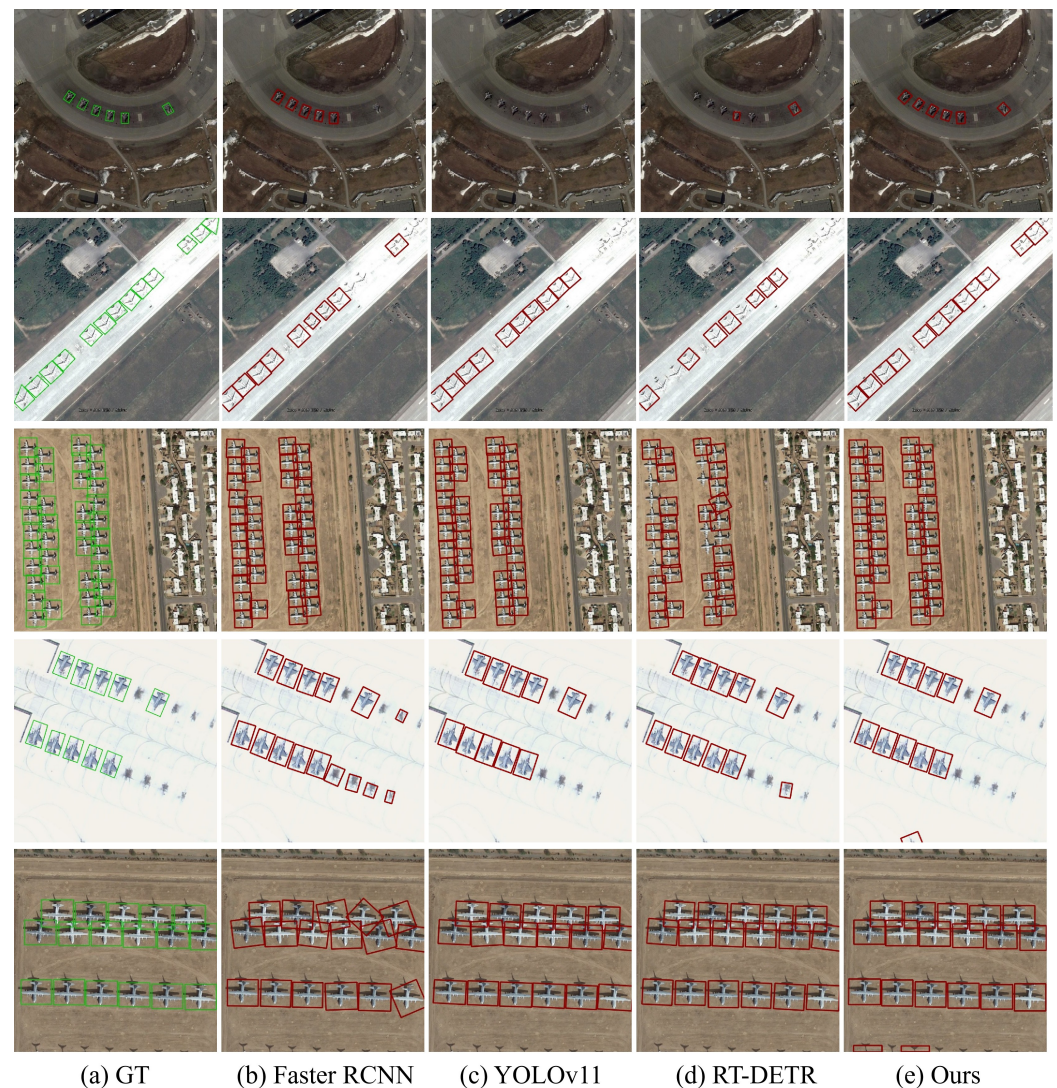


Figure 7. Visual comparison of detection results on the MAR20 dataset. (a) Ground truth, (b) Faster R-CNN, (c) YOLO11, (d) RT-DETR, (e) SFSMamba-DETR. Green boxes denote ground-truth annotations, and red boxes denote model predictions. The proposed detector produces more accurate bounding boxes with fewer missed detections in dense aircraft arrangements.

The qualitative results highlight three key advantages of the proposed framework: (1) *Dense target regions*: SFSMamba-DETR accurately detects closely packed aircraft with minimal false positives and missed detections; (2) *Small targets*: In the Jilin-1 satellite imagery, the detector successfully identifies small aircraft that other methods fail to detect; (3) *Complex backgrounds*: The framework effectively suppresses false alarms from background structures such as buildings and taxiways.

As shown in Figure 7, the MAR20 dataset contains diverse fine-grained military aircraft parked closely on airfield aprons. Faster R-CNN exhibits multiple missed detections in the upper rows where aircraft are partially occluded by adjacent shadows, and produces a

notable false positive over a hangar roof that visually resembles a small aircraft fuselage. YOLO11 achieves better recall but generates duplicated boxes around overlapping wingtips and occasionally misclassifies bombers as fighters due to insufficient long-range context. RT-DETR substantially reduces duplicates owing to its Hungarian matching, yet several elongated aircraft (e.g., B-52 in the top-right region) are localized with imprecise bounding boxes that exclude wing tips.



Figure 8. Visual comparison of detection results on the UCAS-AOD dataset. (a) Ground truth, (b) Faster R-CNN, (c) YOLO11, (d) RT-DETR, (e) SFSMamba-DETR. Green boxes denote ground-truth annotations, and red boxes denote model predictions. SFSMamba-DETR achieves more precise localization especially for small and overlapping targets.

In contrast, SFSMamba-DETR (column e) yields tight, well-aligned bounding boxes for all instances. The SFS module's tri-input state space scanning effectively propagates global airfield context, allowing the model to disambiguate visually similar but structurally distinct aircraft types. The DSWA module's small-window branch captures fine-grained shape cues such as wing sweep and tail configuration, while its large-window branch incorporates surrounding tarmac patterns to suppress confounding structures. The qualitative gains are consistent with the +2.6 percentage-point mAP@0.5 improvement over RT-DETR-L and +5.1 percentage-point improvement over DINO reported in Table 2.

The UCAS-AOD visualizations in Figure 8 highlight detection performance in two contrasting scenarios: aircraft on aprons (top rows) and vehicles in parking lots (bottom

rows). Vehicle detection is particularly demanding due to densely packed targets with arbitrary orientations and high inter-instance similarity. Faster R-CNN suffers from severe missed detections in tightly packed parking rows, while YOLO11 produces oversized boxes that merge adjacent vehicles. RT-DETR mitigates the merging issue but still misses several small vehicles near the image boundary where contextual cues are limited.

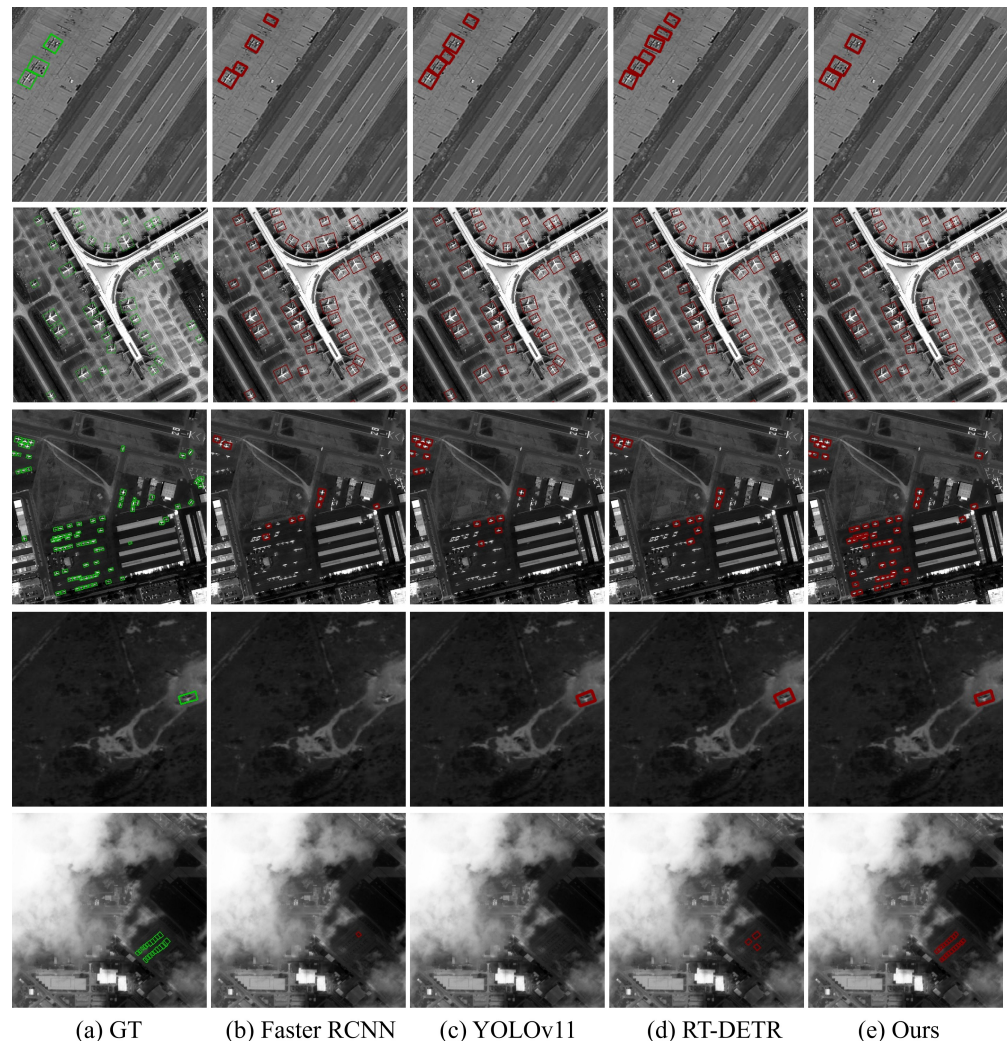


Figure 9. Visual comparison of detection results on the Jilin-1 Satellite Aircraft Detection Dataset. (a) Ground truth, (b) Faster R-CNN, (c) YOLO11, (d) RT-DETR, (e) SFSMamba-DETR. Green boxes denote ground-truth annotations, and red boxes denote model predictions. The proposed detector demonstrates strong capability in detecting small aircraft in low-contrast satellite imagery.

SFSMamba-DETR achieves near-perfect recall and produces compact, instance-level bounding boxes even in the most crowded regions. This is enabled by the complementary action of SFS and DSWA: SFS captures the periodic spatial pattern of parking rows, providing a strong prior for distinguishing individual vehicles, while DSWA's Dual-Scale Window Attention preserves the precise rectangular footprints essential for accurate car localization. The visual quality matches the quantitative improvements (97.6% mAP@0.5, +1.4 percentage points over RT-DETR-L) and supports our claim that the proposed framework generalizes well from large aircraft targets to small ground vehicles within the same model.

The Jilin-1 satellite imagery in Figure 9 presents the most challenging visualization scenario. The panchromatic single-channel images suffer from low contrast between aircraft and surrounding tarmac, and target instances often occupy fewer than 30×30 pixels. Faster R-CNN and YOLO11 both miss numerous small aircraft, particularly those positioned

against bright tarmac or near shadow boundaries. RT-DETR detects more targets but produces false positives over runway markings and taxiway joints that locally resemble aircraft silhouettes.

SFSMamba-DETR consistently outperforms all baselines on this challenging benchmark, recovering aircraft instances that are nearly invisible to other methods. We attribute this to two factors. First, the SFS module's four-directional 2D Selective Scan accumulates evidence along entire runways, allowing the network to exploit the strong contextual prior that aircraft typically appear in linear formations on aprons. Second, the EfficientNet-B4 backbone with compound scaling preserves higher spatial resolution at deeper layers, which is critical for retaining fine-grained signatures of sub-meter-scale targets. These observations align with the substantial +3.8 percentage-point $mAP@0.5$ and +4.3 percentage-point AP_5 gains over RT-DETR-L documented in Table 4.

4.5.2. Failure Mode Analysis

Although SFSMamba-DETR improves both quantitative accuracy and visual localization, several failure modes remain. First, aircraft located inside heavy cast shadows may still be missed when the target texture is almost indistinguishable from the surrounding tarmac. This is most visible in low-contrast Jilin-1 scenes, where the panchromatic imagery provides limited color cues. Second, elongated runway markings and taxiway joints occasionally trigger false positives because their local geometry resembles aircraft fuselage-wing patterns. DSWA reduces such cases by incorporating larger-window context, but it cannot fully remove them when the background structure is repeated over a long runway segment. Third, very dense apron scenes may still produce slightly merged boxes around overlapping wing tips, especially for small aircraft instances close to the image boundary. These observations suggest that future work should combine the current HBB detector with oriented detection heads and, where available, multi-modal or multi-temporal imagery to better separate true aircraft from shadows and airport infrastructure.

4.5.3. Performance–Efficiency Trade-Off

Figure 10 illustrates the $mAP@0.5$ versus FPS trade-off on all three datasets (MAR20, UCAS-AOD, and Jilin-1). SFSMamba-DETR achieves the best balance, consistently positioned in the upper-right region of all plots. While some YOLO variants achieve higher FPS, they sacrifice significant accuracy. Transformer-based methods like DINO and Co-DETR achieve reasonable accuracy but at much lower inference speeds. The proposed framework bridges this gap, demonstrating that the linear-complexity Mamba scanning effectively reduces computational overhead while the dual-scale attention maintains strong detection performance.

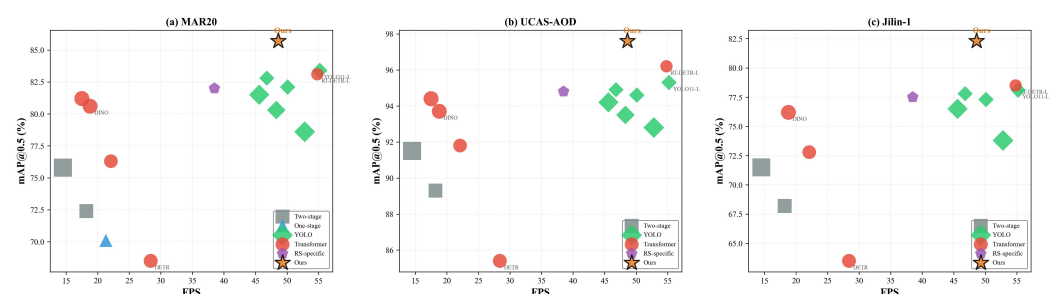


Figure 10. Performance versus efficiency comparison across three datasets. (a) $mAP@0.5$ vs. FPS on MAR20. (b) $mAP@0.5$ vs. FPS on UCAS-AOD. (c) $mAP@0.5$ vs. FPS on Jilin-1. Bubble area is linearly proportional to the number of model parameters, with a uniform visualization scale used across all three subplots. SFSMamba-DETR (★) is denoted by the yellow star and achieves the best accuracy–speed trade-off across all datasets.

4.5.4. Feature Activation Visualization

To intuitively demonstrate how each module contributes to target localization, we visualize the 3D feature activation maps at different stages of SFSMamba-DETR in Figure 11. The input features exhibit broad, noisy activations with weak target responses. After backbone extraction, target regions become more prominent but background noise persists. The SFS module sharpens target peaks through long-range selective scanning, while the DSWA module further refines localization through dual-scale local attention. After CFAM aggregation, the final output presents precise spike-like activations exclusively at target locations, with near-zero background response. This progressive feature refinement validates our design philosophy of combining global Mamba scanning with local multi-scale attention.

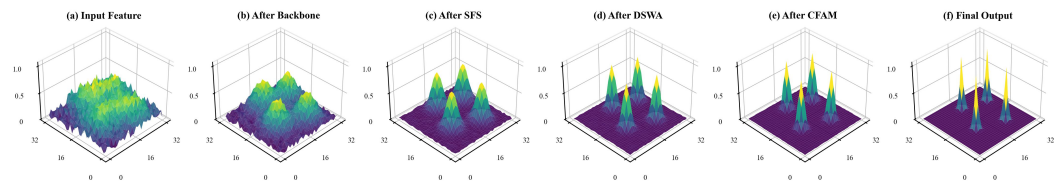


Figure 11. 3D feature activation visualization at different stages of SFSMamba-DETR. Colors encode normalized activation intensity, with purple/blue indicating lower responses and yellow indicating higher responses. (a) Input feature map with broad noisy activations. (b) After backbone extraction. (c) After SFS module, where long-range dependencies sharpen target responses. (d) After DSWA, where dual-scale attention further refines localization. (e) After CFAM aggregation. (f) Final output with precise spike-like activations at target locations. The progressive suppression of background noise and sharpening of target peaks confirm the contribution of each proposed module.

4.5.5. Feature Representation Analysis

To investigate the quality of learned feature representations, we employ t-SNE [65] to visualize the encoder output features on the Jilin-1 dataset. As shown in Figure 12, our SFSMamba-DETR produces more discriminative and well-clustered feature embeddings compared to RT-DETR and DINO. The foreground (aircraft) and background features are more clearly separated, indicating that our SFS and DSWA modules effectively enhance the discriminability of learned representations for remote sensing object detection.

4.5.6. Convergence Analysis

Figure 13 compares the training convergence curves of SFSMamba-DETR with RT-DETR and DINO on the Jilin-1 dataset. SFSMamba-DETR converges faster and achieves a higher final mAP, benefiting from the efficient information flow enabled by the SFS module's linear-complexity scanning and the DSWA module's structured local attention.

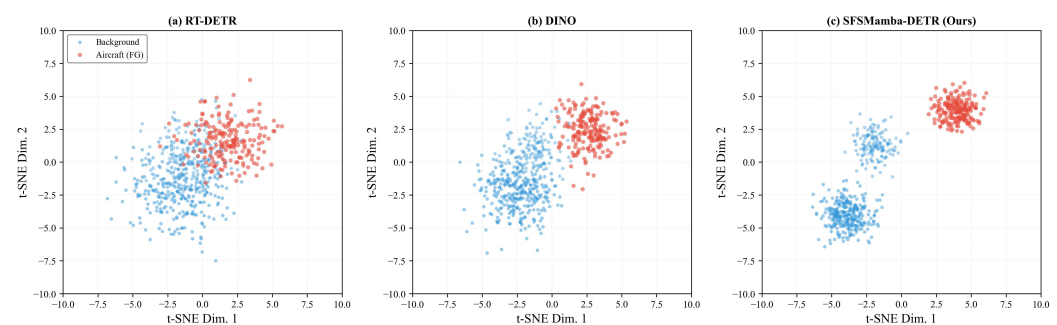


Figure 12. t-SNE visualization of encoder output features on the Jilin-1 dataset. (a) RT-DETR, (b) DINO, (c) SFSMamba-DETR. The proposed framework produces more discriminative feature embeddings with clearer foreground–background separation.

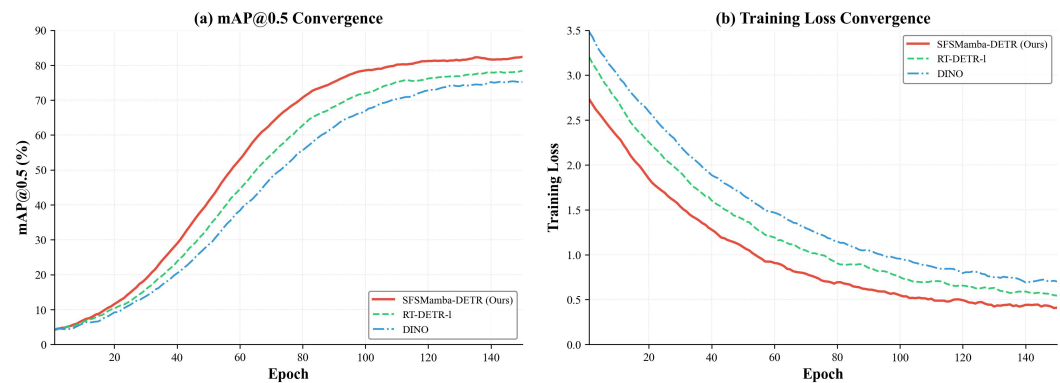


Figure 13. Training convergence comparison on the Jilin-1 dataset. (a) mAP@0.5 curves during training. (b) Training loss curves. SFSMamba-DETR converges faster and achieves higher final performance.

5. Discussion

The experimental results validate the effectiveness of our SFSMamba-DETR framework from multiple perspectives. Several key observations merit further discussion.

Complementarity of SFS and DSWA. The ablation results (Table 6) reveal that the SFS and DSWA modules provide complementary benefits: SFS excels at capturing long-range spatial dependencies through its linear-complexity selective scanning, which is particularly beneficial for understanding the global layout of airport scenes and distinguishing target objects from distant background elements. DSWA, on the other hand, focuses on local multi-scale patterns through its dual-window design, which is critical for resolving fine-grained details of small aircraft targets. The combined improvement (+2.5 percentage points) is close to the sum of individual gains (+1.6 + 1.3 = 2.9 percentage points), with the slight sub-additivity indicating that the two modules share partially overlapping representational capacity while still providing largely complementary benefits.

Tri-input SFS design. The tri-input design of SFS is a deliberate architectural choice that distinguishes our approach from standard Mamba blocks. By accepting main, guide, and auxiliary features from different FPN levels, SFS inherently performs cross-scale information exchange during the state space scanning process. This is fundamentally different from post hoc feature fusion, as the scanning dynamics are jointly influenced by multi-scale context, enabling more informed feature selection.

Scalability to challenging scenarios. The Jilin-1 results demonstrate that SFSMamba-DETR maintains substantial advantages even under extreme conditions with low-contrast satellite imagery and small targets. The consistent +3.8 percentage-point improvement on mAP@0.5 over RT-DETR-L suggests that the combination of Mamba’s global context modeling and DSWA’s multi-scale local attention is particularly well-suited for the unique challenges of satellite-based remote sensing detection.

Efficiency considerations. Despite introducing additional modules (SFS, DSWA, CFAM), SFSMamba-DETR maintains competitive inference speed (48.6 FPS) compared to RT-DETR-L (54.8 FPS), a modest trade-off for the significant accuracy gains. This is attributable to Mamba’s linear complexity, which avoids the quadratic overhead of additional self-attention layers. The model parameters (34.8 M) remain moderate, making the approach practical for deployment in real-world remote sensing applications.

Limitations and future work. While SFSMamba-DETR achieves strong results on horizontal bounding box detection, remote sensing often requires oriented bounding box predictions for elongated objects. Extending the framework with oriented detection heads [49] represents a promising future direction. In addition, although the supplementary DOTA and DIOR experiments broaden the category coverage beyond aircraft and vehicles,

they are still evaluated under the HBB protocol. Future work will extend the proposed modules to full OBB evaluation on large-scale multi-class datasets such as DOTA and FAIR1M, where ships, bridges, buildings, and densely arranged small objects appear with stronger rotation variation. Exploring multi-modal fusion [54,55] with Mamba-based scanning could also enhance detection robustness under varying imaging conditions.

6. Conclusions

In this paper, we have presented SFSMamba-DETR, a framework for remote sensing object detection that integrates state space models with Dual-Scale Window Attention. The key contributions include: (1) the Selective Feature Scanning (SFS) module that uses the Mamba-based 2D Selective Scan for efficient tri-input long-range dependency modeling with linear complexity; (2) the Dual-Scale Window Attention (DSWA) mechanism that captures both fine-grained local details and broader contextual patterns through complementary window scales; and (3) the Cross-scale Feature Aggregation Module (CFAM) that orchestrates hierarchical multi-scale feature fusion. Extensive experiments on MAR20, UCAS-AOD, and the Jilin-1 Satellite Aircraft Detection Dataset, together with supplementary results on DOTA and DIOR, demonstrate that SFSMamba-DETR achieves strong detection performance while maintaining competitive inference efficiency. The consistent improvements across different imaging conditions and target categories validate the effectiveness of the proposed design for efficient remote sensing detection.

Author Contributions: Conceptualization, Y.C. and J.Z.; methodology, Y.C. and J.Z.; software, J.Z.; validation, J.Z. and R.M.; formal analysis, J.Z.; investigation, J.Z. and H.W.; resources, Y.C. and H.W.; data curation, J.Z. and R.M.; writing, original draft preparation, Y.C. and J.Z.; writing, review and editing, Y.C., J.Z. and H.W.; visualization, J.Z.; supervision, Y.C.; project administration, Y.C. and H.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets used in this study are publicly accessible or available through the corresponding data provider. The **MAR20** dataset is available at <https://gcheng-nwpu.github.io/> (accessed on 16 July 2026). The **UCAS-AOD** dataset is available at <https://github.com/Lbx2020/UCAS-AOD-dataset> (accessed on 16 July 2026). The **Jilin-1 Satellite Aircraft Detection Dataset** is available from Chang Guang Satellite Technology Co., Ltd. upon reasonable request at <https://www.jl1mall.com/> (accessed on 16 July 2026). The source code of SFSMamba-DETR is publicly available at <https://github.com/Junchao-XJTU/SFSMamba-DETR> (accessed on 16 July 2026).

Acknowledgments: The authors would like to thank the anonymous reviewers for their constructive comments and suggestions.

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

Abbreviations

The following abbreviations are used in this manuscript:

SFS	Selective Feature Scanning
DSWA	Dual-Scale Window Attention
CFAM	Cross-scale Feature Aggregation Module
SSM	State Space Model
SS2D	2D Selective Scan
VSS	Vision State Space
AIFI	Attention-based Intra-scale Feature Interaction

DLGPA	Dual Local–Global Perception Attention
FPN	Feature Pyramid Network
mAP	mean Average Precision
FPS	Frames Per Second

References

- Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and A New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296–307. [CrossRef]
- Cheng, G.; Yuan, X.; Yao, X.; Yan, K.; Zeng, Q.; Xie, X.; Han, J. Towards Large-Scale Small Object Detection: Survey and Benchmarks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2023**, *45*, 13467–13488. [CrossRef] [PubMed]
- Xia, G.S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-Scale Dataset for Object Detection in Aerial Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 3974–3983.
- Sun, X.; Wang, P.; Yan, Z.; Xu, F.; Wang, R.; Diao, W.; Chen, J.; Li, J.; Feng, Y.; Xu, T.; et al. FAIR1M: A Benchmark Dataset for Fine-grained Object Recognition in High-Resolution Remote Sensing Imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *184*, 116–130. [CrossRef]
- 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*; Curran Associates, Inc.: Red Hook, NY, USA, 2015; Volume 28.
- Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 6154–6162.
- 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*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
- Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2017; pp. 2980–2988.
- Jocher, G. YOLOv5 by Ultralytics. 2021. Available online: <https://github.com/ultralytics/yolov5> (accessed on 16 July 2026).
- 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 7464–7475.
- Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 16 July 2026).
- 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 European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21.
- Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 107984–108011. [CrossRef]
- Jocher, G.; Qiu, J. YOLO11 by Ultralytics. 2024. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 16 July 2026).
- Liu, W.; Li, Y.; Zhang, S.; Mao, R. Towards Smart City Supervision: A Detection Pipeline for Illegal Buildings. *Eng. Appl. Artif. Intell.* **2026**, *163*, 113052. [CrossRef]
- Yan, Z.; Li, Y. AMSRDet: An Adaptive Multi-Scale UAV Infrared-Visible Remote Sensing Vehicle Detection Network. *Sensors* **2026**, *26*, 817. [CrossRef] [PubMed]
- Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations*, Virtual, 3–7 May 2021.
- 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*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229.
- Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is All You Need. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2017; Volume 30.
- 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*, Virtual, 3–7 May 2021.
- 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*, Virtual, 25–29 April 2022.
- 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*, Virtual, 1–5 May 2023.
- Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 16965–16974.

24. Gu, A.; Dao, T. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv* **2024**, arXiv:2312.00752.
25. Gu, A.; Goel, K.; Ré, C. Efficiently Modeling Long Sequences with Structured State Spaces. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022.
26. Liu, Y.; Tian, Y.; Zhao, Y.; Yu, H.; Xie, L.; Wang, Y.; Ye, Q.; Liu, Y. VMamba: Visual State Space Model. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103031–103063. [[CrossRef](#)]
27. Zhu, L.; Liao, B.; Zhang, Q.; Wang, X.; Liu, W.; Wang, X. Vision Mamba: Efficient Visual Representation Learning with Bidirectional State Space Model. In Proceedings of the International Conference on Machine Learning, Virtual, 21–27 July 2024.
28. Sun, P.; Zheng, Y.; Zhou, Z.; Xu, W.; Liu, Q. MAR20: A Benchmark for Military Aircraft Recognition from Remote Sensing Images. *Natl. Remote Sens. Bull.* **2024**, *28*, 996–1009.
29. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation Robust Object Detection in Aerial Images Using Deep Convolutional Neural Network. In *IEEE International Conference on Image Processing*; IEEE: Piscataway, NJ, USA, 2015; pp. 3735–3739.
30. Chang Guang Satellite Technology Co., Ltd. Jilin-1 Satellite Remote Sensing Image Dataset for Aircraft Detection. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5.
31. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 2117–2125.
32. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019; pp. 9627–9636.
33. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint Triplets for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019; pp. 6569–6578.
34. Wang, C.; He, W.; Nie, Y.; Guo, J.; Liu, C.; Wang, Y.; Han, K. Gold-YOLO: Efficient Object Detector via Gather-and-Distribute Mechanism. *Adv. Neural Inf. Process. Syst.* **2024**, *36*, 51094–51112. [[CrossRef](#)]
35. Cheng, T.; Song, L.; Ge, Y.; Liu, W.; Wang, X.; Shan, Y. YOLO-World: Real-Time Open-Vocabulary Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2024; pp. 16901–16911.
36. Zong, Z.; Song, G.; Liu, Y. DETRs with Collaborative Hybrid Assignments Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 6748–6758.
37. Chen, Q.; Chen, X.; Wang, J.; Zhang, S.; Yao, K.; Feng, H.; Han, J.; Ding, E.; Zeng, G.; Wang, J. Group DETR: Fast DETR Training with Group-Wise One-to-Many Assignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 6633–6642.
38. Chen, S.; Sun, P.; Song, Y.; Luo, P. DiffusionDet: Diffusion Model for Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 19830–19843.
39. Li, F.; Zeng, A.; Liu, S.; Zhang, H.; Li, H.; Zhang, L.; Ni, L.M. Lite DETR: An Interleaved Multi-Scale Encoder for Efficient DETR. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2023; pp. 18558–18567.
40. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 8759–8768.
41. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 10781–10790.
42. Li, H.; Deng, Z.; Li, Y.; Zhu, H. Rethinking Multi-scale Feature Fusion for Object Detection in Remote Sensing Imagery. In *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*; IEEE: Piscataway, NJ, USA, 2024; pp. 2783–2786.
43. Pei, X.; Huang, T.; Xu, C. EfficientVMamba: Atrous Selective Scan for Light Weight Visual Mamba. *arXiv* **2024**, arXiv:2403.09977.
44. Yang, C.; Chen, Z.; Espinosa, M.; Ericsson, L.; Wang, Z.; Liu, J.; Crowley, E.J. PlainMamba: Improving Non-Hierarchical Mamba in Visual Recognition. *arXiv* **2024**, arXiv:2403.17695.
45. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z. RSMamba: Remote Sensing Image Classification with State Space Model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 8002605. [[CrossRef](#)]
46. Li, W.; Zhu, J.; Wen, J.; Wang, D. Mamba-in-Mamba: Centralized Mamba-Cross-Scan in Tokenized Mamba Model for Hyperspectral Image Classification. *arXiv* **2024**, arXiv:2405.12003.
47. Li, Y.; Wang, T.; Luo, N.; Zhou, L.; Chen, Q. Cgmamba: Intelligent Identification of Counterfeit Goods Based on State Space Models. *Int. J. Intell. Syst.* **2025**, *2025*, 9939880. [[CrossRef](#)]
48. Huang, Z.; Chen, N.; Xie, M.; Wan, S. MambaYOLO: SSMS-Based YOLO For Object Detection. *arXiv* **2024**, arXiv:2406.05835.
49. Xie, X.; Cheng, G.; Wang, J.; Yao, X.; Han, J. Oriented R-CNN for Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 3520–3529.
50. Ding, J.; Xue, N.; Long, Y.; Xia, G.S.; Lu, Q. Learning RoI Transformer for Oriented Object Detection in Aerial Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 2849–2858.
51. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined Single-Stage Detector with Feature Refinement for Rotating Object. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual, 2–9 February 2021; Volume 35, pp. 3163–3171. [[CrossRef](#)]

52. Han, J.; Ding, J.; Xue, N.; Xia, G.S. ReDet: A Rotation-equivariant Detector for Aerial Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 2786–2795.
53. Li, Y.; Hou, Q.; Zheng, Z.; Cheng, 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*; IEEE: Piscataway, NJ, USA, 2023; pp. 16794–16805.
54. Qin, Z.; Li, Y. DCAM-DETR: Dual Cross-Attention Mamba Detection Transformer for RGB-Infrared Anti-UAV Detection. *Information* **2026**, *17*, 103. [\[CrossRef\]](#)
55. Yuan, H.; Li, Y. CSFADet: Dual-Modal Anti-UAV Detection via Cross-Spectral Feature Alignment and Adaptive Multi-Scale Refinement. *Algorithms* **2026**, *19*, 254. [\[CrossRef\]](#)
56. Yang, Z.; Li, Y.; Shao, J.; Li, W.; Luo, N. Lightweight Hybrid Attention for Channel, Spatial, and Token-level Enhancement. In *Proceedings of the 2025 8th International Conference on Computer Information Science and Artificial Intelligence*; IEEE: Piscataway, NJ, USA, 2025; pp. 212–217.
57. Zhou, D.; Li, Y. SoccerDETR: Real-Time Soccer Object Detection via Visual State Space Models with Semantic-Aware Feature Fusion. *Technologies* **2026**, *14*, 142. [\[CrossRef\]](#)
58. Tan, M.; Le, Q. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the International Conference on Machine Learning*; PMLR: Long Beach, CA, USA, 2019; pp. 6105–6114.
59. Ding, X.; Zhang, X.; Ma, N.; Han, J.; Ding, G.; Sun, J. RepVGG: Making VGG-style ConvNets Great Again. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2021; pp. 13733–13742.
60. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 658–666.
61. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In *Proceedings of the International Conference on Learning Representations*, New Orleans, LA, USA, 6–9 May 2019.
62. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft COCO: Common Objects in Context. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755.
63. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO Series in 2021. *arXiv* **2021**, arXiv:2107.08430.
64. Lyu, C.; Zhang, W.; Huang, H.; Zhou, Y.; Wang, Y.; Liu, Y.; Zhang, S.; Chen, K. RTMDet: An Empirical Study of Designing Real-Time Object Detectors. *arXiv* **2022**, arXiv:2212.07784.
65. Van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.

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.