

Article

SPA-DETR: An Enhanced RT-DETR with Spatial-Preserving Attention and Adaptive Loss for UAV Spectrogram Signal Detection

Conghao Fu, Lu Xu * and Yijia Zhang

School of Information Science and Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China; fu_conghao@163.com (C.F.); waiting@zstu.edu.cn (Y.Z.)

* Correspondence: xulu@zstu.edu.cn

Abstract

Rapid detection of unauthorized unmanned aerial vehicles (UAVs) via radio frequency (RF) spectrograms is critical for low-altitude security. However, standard object detectors struggle to locate transient, frequency-hopping UAV signals because their microscopic spatial footprints are easily discarded by conventional lossy downsampling and overwhelmed by complex background noise. To overcome this limitation, we propose SPA-DETR, a custom architecture based on the RT-DETR framework. The core of our design is the Spatial-Preserving Attention (SPA) block, which integrates Space-to-Depth Convolution (SPDConv) with a Parallel Patch-Aware Attention (PPA) module. By replacing traditional pooling mechanisms, the SPA block preserves the spatial details of weak signals without information loss, while the PPA module concurrently filters out ambient background interference. Furthermore, to address the severe foreground–background imbalance in RF spectrograms, we introduce an Adaptive Threshold Focal Loss (ATFL). Operating exclusively during training, ATFL prevents background noise gradients from dominating the learning process, forcing the network to focus on hard-to-detect signal patches without adding computational overhead during inference. Experiments on our public RFUAV dataset validate the approach. SPA-DETR achieves an $mAP_{50:95}$ of 86.8% and an AP_S of 85.6%, improving upon the baseline RT-DETR-R18 by 4.9% and 5.2%, respectively. Operating at 235.2 FPS with only 23.74 M parameters, SPA-DETR outperforms contemporary detectors such as YOLOv10m, as well as heavier models like YOLOv8m and RT-DETR-R50, highlighting its efficiency and practical value for real-time low-altitude security applications.

Keywords: unmanned aerial vehicle (UAV); radio frequency (RF) spectrogram; object detection; RT-DETR; spatial-preserving attention; focal loss



Academic Editor: Yun Zhang

Received: 19 May 2026

Revised: 29 June 2026

Accepted: 27 July 2026

Published: 1 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.

1. Introduction

The rapid development of unmanned aerial vehicle (UAV) technology has transformed civilian logistics and modern electronic warfare. However, their widespread use has increased the risk of “low–slow–small” aerial threats, making unauthorized incursions a persistent concern for low-altitude security [1,2]. Traditional radar systems often struggle to distinguish small UAVs from birds or urban obstacles [3]. Passive radio frequency (RF) sensing provides a practical alternative by transforming intercepted time-domain signals into 2D spectrograms [4]. This enables the use of computer vision (CV) algorithms to analyze the temporal-frequency patterns of UAV communication links. Nonetheless, the sparse and transient nature of frequency-hopping signals requires detection systems with high sensitivity and real-time processing capabilities.

While mainstream object detection frameworks, such as the YOLO series and DETR variants, perform well on natural images, applying them directly to RF spectrograms yields suboptimal results. The main issue is the small physical footprint of frequency-hopping signals. During their brief dwell time, these pulses occupy very few pixels. Standard architectures extract semantic features through successive downsampling stages, such as max-pooling or strided convolutions. This spatial compression benefits typical large objects, but it acts as a lossy filter for weak UAV signals, diluting critical spatial details before deeper layers can process them [5]. Moreover, RF spectrograms suffer from a severe foreground–background imbalance, with ambient noise dominating the image. This causes a high false-alarm rate, as standard loss functions struggle to prioritize rare signal pulses over background clutter [6].

To address these structural issues, we propose SPA-DETR, a custom architecture based on the Real-Time Detection Transformer (RT-DETR) [7] for weak UAV signal detection. To preserve the spatial footprint of hopping signals, we fundamentally redesign the network’s downsampling paradigm. Rather than relying on conventional lossy operations, we introduce a Spatial-Preserving Attention (SPA) block, which integrates Space-to-Depth Convolution (SPDConv) with a Parallel Patch-Aware Attention (PPA) module. We deploy this block across both the early backbone and the multi-scale fusion neck. The SPDConv operation retains the geometric details of transient signals without spatial loss, while the subsequent PPA mechanism filters out the accompanying background noise. In addition to these structural changes, we introduce an Adaptive Threshold Focal Loss (ATFL) to handle the extreme sample disparity. This objective function dynamically shifts the learning focus toward hard-to-detect signal patches. This direct approach reduces false alarms without adding computational overhead during inference.

The primary contributions of this work are summarized as follows:

- (1) We design the SPA block to replace conventional downsampling. By combining the lossless dimension reduction of SPDConv with the noise-filtering of PPA across the backbone and neck stages, this module maintains the spatial integrity of weak frequency-hopping signals and isolates them from ambient interference.
- (2) We propose the ATFL to address the severe foreground–background imbalance in RF spectrograms. It dynamically adjusts the optimization focus based on target confidence, improving the network’s sensitivity to small targets in noisy environments.
- (3) We conduct extensive experiments on the public RFUAV dataset. With a compact size of 23.74 M parameters, SPA-DETR achieves an $mAP_{50:95}$ of 86.8% and an AP_S of 85.6% at 235.2 FPS. It consistently outperforms contemporary detectors like YOLOv8m, YOLOv10m, and the RT-DETR-R50 baseline.

2. Related Work

RF-based UAV detection has transitioned from manual feature engineering to data-driven deep learning methodologies. Early approaches primarily relied on statistical signal processing to identify UAV signatures. For instance, energy detection serves as a foundational technique by evaluating received signal power against predefined thresholds. Recently, Sova et al. proposed a burst-aware cascade framework that leverages low-cost energy detection as a preliminary filter [8]. Similarly, Magiera introduced a non-parametric lightweight classification paradigm based on the physical statistical characteristics of RF transmissions, extracting features such as bandwidth and pulse duration [9]. However, while computationally efficient, these threshold-based models remain highly susceptible to fluctuations in low signal-to-noise ratio (SNR) environments, often suffering from severe false-alarm rates when facing channel multipath fading. To improve noise resilience, researchers have extensively investigated cyclostationary feature analysis and RF finger-

printing, which exploit the inherent periodicity and transient behaviors of modulated signals. For example, Morge-Rollet et al. explored physical-layer protocol statistical fingerprints (PLPSF), utilizing power spectral entropy and packet time-sequence features to distinguish specific UAV models [10]. Furthermore, Ezuma et al. transformed 1D time-domain signals into energy–time–frequency trajectories to extract transient energy features for high-precision classification [11]. Despite demonstrating superior robustness against ambient interference, these methods demand complex mathematical transformations (e.g., cyclic autocorrelation), incurring significant computational overhead that inherently restricts their viability for real-time deployment [12].

To bypass the limitations of handcrafted features and computational bottlenecks, research shifted towards transforming 1D time-series data into 2D time–frequency spectrograms, enabling the direct application of deep learning. Initially, researchers employed standard CNN-based classifiers to categorize the entire spectrogram. For instance, Xin et al. introduced a Complex-Valued CNN (CV-CNN) that jointly processes power spectral density and Sobel edge features to significantly enhance anti-interference capabilities in low SNR regimes [13]. While effective for simple presence detection or global classification, these image-level classifiers inherently lack spatial awareness. They fail to precisely localize the exact time and frequency boundaries of multiple transient pulses within heavily cluttered backgrounds. Consequently, recent studies have adapted mainstream object detectors to not only classify but also physically bound UAV signals within the spectrogram. For example, Zhu et al. designed a high-resolution passive system that utilizes the YOLO framework to explicitly anchor frequency-hopping signal regions before applying a secondary CNN classification. Nevertheless, directly applying these standard bounding-box architectures to highly transient frequency-hopping signals remains suboptimal. Unlike continuous communication links that form solid shapes, frequency-hopping pulses appear as isolated, microscopic energy patches with minimal spatial footprints. Conventional convolutional architectures rely on lossy downsampling, which inadvertently discards these discrete footprints amid background clutter, explicitly requiring the targeted structural modifications we propose.

Detecting small objects remains a fundamental challenge in computer vision. Under standard metrics like MS COCO, targets smaller than 32×32 pixels lose critical information during the successive downsampling stages of conventional Convolutional Neural Networks (CNNs). To address this spatial degradation, SPDConv [14] was introduced to achieve lossless downsampling by reorganizing spatial dimensions into the channel axis. Recent architectures have successfully applied this concept; for example, DBYOLOv8 utilizes SPDConv for complex drone imagery [15], and SPT-YOLO applies it to infrared small target detection to preserve fine-grained geometric features [16]. However, retaining all spatial information also preserves the background noise inherent in complex environments. Consequently, researchers often combine spatial preservation techniques with attention mechanisms. For the extreme sparsity of frequency-hopping signals, standard attention modules are insufficient. A more aggressive, patch-aware filtering approach is necessary to suppress ambient clutter without attenuating the weak signal.

Transformer-based architectures, such as DETR [17], have simplified object detection by framing it as an end-to-end set prediction problem. This eliminates post-processing steps like Non-Maximum Suppression (NMS), which are required by YOLO-family models and introduce variable inference latency. Building on this, RT-DETR bridged the gap between Transformer accuracy and CNN inference speeds. Nevertheless, the baseline RT-DETR relies heavily on deep semantic features from the backbone, often neglecting the low-level textural details necessary for small-object localization. Recent adaptations for aerial imagery, including SO-RTDETR [18] and SDU-RTDETR [19], address this by enhancing

multi-scale feature fusion and utilizing shallower detection layers (e.g., the P2 layer). While these enhancements improve spatial awareness, they increase computational overhead and are not optimized for the severe class imbalance of RF spectrograms. Our proposed SPA-DETR addresses this gap by unifying the SPA block and ATFL. This architecture achieves a balance between sensitivity to micro-signals and real-time efficiency without significantly increasing the parameter count.

3. Proposed Method

3.1. Overall Architecture of SPA-DETR

To address the challenges of detecting weak frequency-hopping signals, we introduce SPA-DETR, a custom architecture based on the RT-DETR-R18 framework, as illustrated in Figure 1. While it retains the standard three-part structure—a feature extraction backbone, a multi-scale hybrid encoder, and a Transformer-based decoder—it fundamentally changes how spatial information is processed.

The input spectrogram first passes through the ResNet-18 backbone [20]. However, instead of using the conventional max-pooling operation at the early P2 stage, SPA-DETR integrates our proposed SPA block. This replacement prevents the loss of the tiny geometric footprints of UAV signals during the initial stages of feature extraction. After this initial extraction, the hierarchical features enter the hybrid encoder for multi-scale fusion. Here, we embed additional SPA blocks within the bottom-up pathways of the feature pyramid, replacing the standard strided convolutions typically used for downsampling. This cross-stage design ensures that the fine spatial details preserved in the early layers are successfully propagated to higher levels without accumulating background noise. Finally, the decoder uses these filtered, multi-scale feature maps for bounding box regression and classification.

During training, the entire network is optimized using the ATFL. This loss function dynamically adjusts the learning gradients, forcing the network to prioritize rare signal patches over the dominant background clutter.

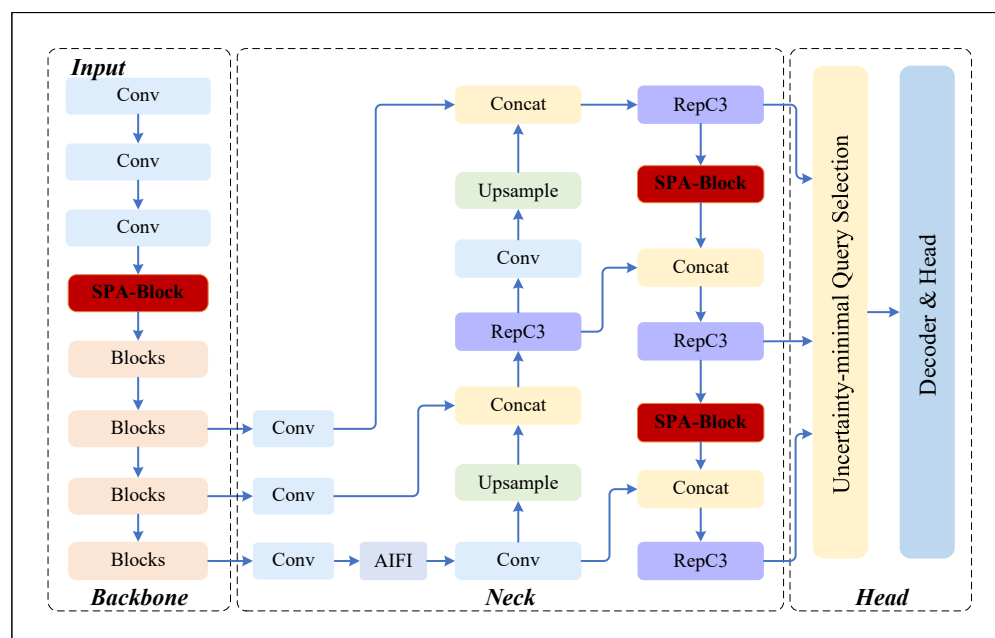


Figure 1. Overall architecture of the proposed SPA-DETR. The SPA blocks are strategically integrated into the early stage of the ResNet-18 backbone and the bottom-up pathways of the hybrid encoder to prevent the loss of microscopic signal footprints.

3.2. SPA-Block

To circumvent the information loss dictated by conventional spatial compression, we engineer the SPA, the detailed internal structure of which is depicted in Figure 2. In standard convolutional architectures, downsampling operations—such as max-pooling or strided convolutions—achieve an expanded receptive field by discarding intermediate pixels. For microscopic frequency-hopping signals that span only a minimal pixel region, such aggressive decimation often results in the complete erasure of the target’s geometric footprint.

Addressing this structural vulnerability, the primary phase of our SPA-block adopts SPDConv. Rather than discarding spatial data, SPDConv downsizes the feature map by physically reorganizing spatial blocks into the channel dimension. Given an intermediate feature map $X \in \mathbb{R}^{C \times H \times W}$, a scaling factor of $s = 2$ partitions the spatial dimensions into sub-tensors. These sub-tensors are subsequently concatenated along the channel axis to yield a newly structured tensor $X' \in \mathbb{R}^{4C \times \frac{H}{2} \times \frac{W}{2}}$, followed by a non-strided convolution to regulate the channel depth. This lossless transformation guarantees that every raw pixel of the weak UAV signal is explicitly retained.

However, the indiscriminate preservation of spatial data introduces a critical secondary challenge: the complex background noise, which dominates the RF spectrogram, is similarly retained and folded into the expanded channel dimensions. If left unmitigated, this amplified ambient clutter can severely overwhelm the network’s semantic reasoning.

To decouple the weak signal features from this retained noise, we sequentially integrate the PPA module [21]. This specific pairing is not an arbitrary combination, but a carefully designed architectural synergy. Because the preceding SPD operation physically embeds spatial patterns into the channel depth, the traditional spatial structure is fundamentally altered. Consequently, the resulting tensor strictly requires an advanced mechanism capable of jointly decoding these newly formed cross-channel correlations and the localized spatial dependencies.

Operating directly on the preserved features X_{spd} , the PPA module provides a multi-branch parallel architecture perfectly matched to this reorganized representation. It incorporates a cascade of 3×3 convolutions for multi-scale local feature extraction, alongside a 1×1 skip connection (X_{skip}) for identity preservation. Crucially, it deploys two Local-Global Attention (LGA) branches operating at distinct patch sizes (2×2 and 4×4). These patch-aware branches are mathematically designed to align with the sub-tensors partitioned by the SPDConv. This precise alignment allows the network to meticulously evaluate local patch dependencies while establishing a global receptive field. Let $Conv_{3 \times 3}^{(i)}$ denote the i -th independent convolutional layer in the cascade, where each layer maintains its own distinct set of learnable parameters without weight sharing. The intermediate fused tensor X_{fuse} is then aggregated via element-wise addition:

$$X_{fuse} = \sum_{i=1}^3 Conv_{3 \times 3}^{(i)}(X_{spd}) + LGA_2(X_{skip}) + LGA_4(X_{skip}) + X_{skip}. \quad (1)$$

To further distill this enriched feature representation, X_{fuse} passes through a dual-attention mechanism. It initially enters an Efficient Channel Attention (ECA) module [22]. Unlike traditional channel attention, ECA directly recalibrates the expanded channel weights—which now encode rich spatial footprints—without dimensionality reduction. This is immediately succeeded by a Spatial Attention (SA) module, exploiting max-pooling and average-pooling across the channel axis to generate a robust spatial mask that represses irrelevant ambient clutter. The final output is formulated as:

$$F_{out} = SA(ECA(X_{fuse})). \quad (2)$$

Ultimately, the synergy between SPDCnv and PPA establishes an integrated feature refinement mechanism. SPDCnv secures detection sensitivity by ensuring microscopic UAV pixels survive the decimation process, whereas PPA provides the necessary precision by explicitly decoding the folded patch dependencies and stripping away co-located noise. This complementary mechanism empowers the SPA-block to act as a highly selective filter, isolating transient signals in extreme noise-dominant environments.

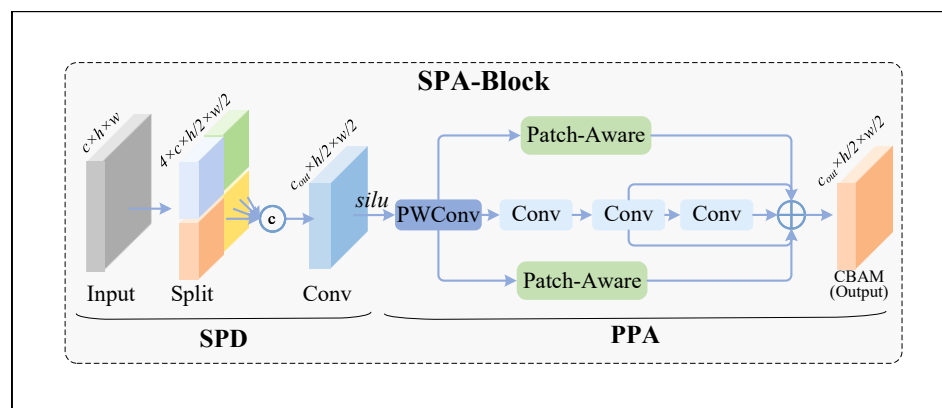


Figure 2. Detailed internal structure of the SPA-block, comprising the SPD for spatial restructuring and the PPA module for feature purification.

3.3. ATFL

While the SPA block successfully extracts features for microscopic targets, architectural improvements alone cannot resolve the optimization bottleneck in RF spectrograms. Even with spatial preservation, background noise still dominates the image tensor. During backpropagation, the small gradients generated by rare frequency-hopping signals are often overwhelmed by the large gradients accumulated from easily classified background pixels. Due to this extreme foreground–background imbalance, standard Binary Cross-Entropy (BCE) loss and conventional Focal Loss with a static focusing parameter are inadequate [23–25].

To address this optimization gap, we introduce the ATFL. ATFL replaces static weight penalization with a batch-aware, dynamic modulation factor $M(p_t)$ to adjust gradient flows based on the network’s training status [26]. Let p_t denote the estimated probability for the ground-truth class. To assess the overall learning difficulty of the current training batch, ATFL calculates the exponential moving average (EMA) of the batch’s mean prediction probability, denoted as $\hat{p}_c$. It then dynamically derives the focusing parameter as $\gamma = -\ln(\hat{p}_c)$. This adaptive mechanism allows the model to adjust its penalization strength in response to fluctuating signal-to-noise ratios across different batches [27].

To complement the filtered features from the SPA block, ATFL applies a piecewise asymmetric modulation strategy pivoting on a confidence threshold of 0.5. For easily classified, high-confidence samples (typically the background noise suppressed by the PPA module, where $p_t > 0.5$), the modulation factor attenuates their gradient contributions using the dynamic γ . Conversely, for hard-to-detect samples (the microscopic UAV signals preserved by SPDCnv, where $p_t \leq 0.5$), the mechanism introduces a regulatory hyperparameter λ (set to 1.5 in our experiments) to elevate the compensation baseline to $\lambda - p_t$. It also applies a dynamic exponent of $-\ln(p_t)$. This piecewise modulation factor $M(p_t)$ is defined as:

$$M(p_t) = \begin{cases} (1 - p_t)^\gamma, & \text{if } p_t > 0.5 \\ (\lambda - p_t)^{-\ln(p_t)}, & \text{if } p_t \leq 0.5 \end{cases} \quad (3)$$

From a mathematical perspective, the exponent $-\ln(p_t)$ in the lower branch theoretically diverges towards infinity as $p_t \rightarrow 0$. To guarantee numerical stability during the actual forward and backward passes, p_t is strictly clamped with a small lower-bound epsilon ($\epsilon = 10^{-6}$) before the logarithmic operation. Furthermore, our training pipeline inherently employs gradient clipping with a maximum gradient norm set to 0.1 to prevent any anomalous gradient explosions caused by extreme sample predictions, ensuring stable convergence.

The final ATFL objective incorporates this modulation factor directly into the standard BCE formulation:

$$\mathcal{L}_{ATFL} = M(p_t) \cdot \mathcal{L}_{BCE}. \quad (4)$$

Ultimately, ATFL and the SPA block form a complete detection ecosystem. During the forward pass, the SPA block isolates the physical footprints of transient signals. During the backward pass, ATFL amplifies the corresponding gradients, ensuring that these weak features receive sufficient weight updates [28]. By combining spatial feature preservation with mathematical gradient scaling, this approach significantly reduces missed detections without adding computational overhead during inference.

4. Experiments

4.1. Datasets and Evaluation Metrics

To validate the proposed SPA-DETR architecture, we conduct experiments on a custom object detection dataset [29]. This dataset is entirely generated by converting the raw In-phase and Quadrature (I/Q) frequency data from the RFUAV benchmark [30] into 2D time–frequency spectrograms. Unlike simulated collections, the original RFUAV benchmark stands as a highly comprehensive dataset in this domain. Specifically, it comprises approximately 1.3 TB of raw 1D I/Q streams collected from 37 distinct UAV models using Universal Software Radio Peripheral (USRP) devices in complex real-world environments. To adapt this massive raw data for our visual detection framework, we transform it into 2D time–frequency spectrograms using the Short-Time Fourier Transform (STFT) [31]. This preprocessing step visualizes the energy distribution of the transient pulses, forming the direct visual input for our network.

We partition the dataset into 3340 training samples, 946 validation samples, and 477 testing samples. For quantitative evaluation, we follow the standard MS COCO protocol [32]. The primary performance metrics are the mean Average Precision (mAP) computed over multiple Intersection over Union (IoU) thresholds, specifically mAP_{50} and the more stringent $mAP_{50:95}$. Because our core focus is capturing microscopic frequency-hopping footprints, we include the Average Precision for small objects (AP_S) to specifically measure detection sensitivity for tiny targets. Finally, to assess the model's viability for deployment on constrained edge devices, we report the parameter count (Params), floating-point operations (FLOPs), and real-time inference speed (FPS).

4.2. Implementation Details

We implement the detection pipeline using PyTorch (v2.0.0) [33] and train all models on a single NVIDIA GeForce RTX 4090 GPU (NVIDIA, Santa Clara, CA, USA) with 24GB of VRAM. To balance computational cost and feature resolution, we resize all input spectrograms to 640×640 pixels before training. We train SPA-DETR and all baseline models to full convergence. Specifically, we train the proposed SPA-DETR for 300 epochs with a batch size of 16. We optimize our architecture end-to-end using the AdamW optimizer [34], setting the initial learning rate to 1×10^{-4} and the weight decay to 1×10^{-4} . To ensure early training stability, we apply a warmup schedule for the first 3 epochs. For a fair comparison,

we train the benchmark models (e.g., the YOLO series [35,36]) according to their officially recommended hyperparameters and training schedules.

For data augmentation, we apply a domain-specific strategy tailored for time-frequency representations. Unlike natural image processing, we exclude spatial-destructive techniques such as Mosaic [37] and Mixup [38] to preserve the physical structure and coherence of the UAV RF pulses. Instead, we apply controlled geometric transformations, including translation and scaling, to simulate physical time delays and multipath fading. We also use random erasing [39] to mimic burst channel interference and localized occlusions. This domain-aware augmentation strategy ensures a realistic evaluation for all subsequent benchmark comparisons and ablation studies.

To ensure a fair and reproducible assessment of inference speed, all FPS measurements reported in this study were conducted under identical hardware and software conditions. Specifically, we measured the inference throughput on a single RTX 4090 GPU using a fixed batch size of 16. All models, including the comparative baselines, were executed using their native PyTorch weights (.pt) in standard evaluation mode with Convolution Batch Normalization (Conv-BN) fusion applied. We explicitly excluded external inference engines or hardware-specific optimizations, such as TensorRT or TorchScript, to reflect the architectural efficiency natively. For each model, the final FPS was calculated by averaging the execution latency over 1000 test iterations, immediately following a 200-iteration GPU warm-up phase.

4.3. Ablation Study

To evaluate the individual contributions and interactions of our proposed modules, we conduct ablation experiments on the RFUAV dataset (Table 1). The baseline RT-DETR-R18 struggles to locate transient frequency-hopping signals due to their small physical footprints, yielding an $mAP_{50:95}$ of 81.9% and an AP_5 of 80.4%.

Table 1. Ablation study of the proposed components on the RFUAV dataset. The SPA block integrates both SPDConv and PPA.

Model	SPDConv	PPA	ATFL	$mAP_{50:95}$	mAP_{50}	AP_5	Params (M)	GFLOPs	FPS
Baseline (RT-DETR-R18)				81.9	99.7	80.4	19.90	57.2	235.6
+SPDConv	✓			83.2	99.7	81.7	19.26	56.8	340.0
+PPA		✓		84.2	99.7	82.7	24.38	71.6	265.5
+SPA-Block (SPD+PPA)	✓	✓		85.7	99.8	84.3	23.74	71.1	235.9
+ATFL (on Baseline)			✓	82.8	99.7	81.4	19.90	57.2	234.8
Ours (SPA-DETR)	✓	✓	✓	86.8	99.8	85.6	23.74	71.1	235.2

Analyzing the structural modifications reveals that adding modules in isolation provides steady gains. As indicated in Table 1, integrating SPDConv alone improves the $mAP_{50:95}$ to 83.2% (a 1.3% increase). While it retains more spatial pixels, the simultaneous retention of background noise limits the overall accuracy. Similarly, deploying only the PPA module yields an $mAP_{50:95}$ of 84.2%. However, combining both modules into the complete SPA block produces a strong synergistic effect, increasing the $mAP_{50:95}$ to 85.7% and the small-object AP_5 to 84.3%. This performance growth effectively validates our theoretical inference in Section 3.2: the spatial feature reorganization and attention decoding mechanisms act as a synergistic signal isolation framework.

Beyond structural changes, modifying the optimization function addresses the optimization bottleneck. Replacing the standard BCE loss with ATFL alone improves the baseline's $mAP_{50:95}$ to 82.8% (a 0.9% increase). While the isolated gain is modest, ATFL proves its value when integrated into the full system. Importantly, this optimization occurs

entirely during the backward pass, improving performance without adding network parameters or increasing inference computational cost (remaining at 19.90 M and 57.2 GFLOPs for the baseline version).

Furthermore, regarding the regulatory hyperparameter λ in ATFL, its general sensitivity and threshold stability have been systematically investigated in its foundational literature. Following a similar evaluation paradigm, we specifically explored its behavior across the extreme imbalance regimes unique to RF spectrograms. While the original literature suggests higher values (e.g., $\lambda = 3.5$) for certain infrared imagery, empirical observations on our RFUAV dataset indicate that setting λ too high ($\lambda \geq 2.0$) over-amplifies the loss for uncertain background noise, causing training instability and degrading convergence. Conversely, a lower setting ($\lambda = 1.0$) fails to provide sufficient gradient compensation for weak signals. The selected setting of $\lambda = 1.5$ acts as an optimal threshold tailored for RF characteristics, dynamically maintaining training stability while ensuring high sensitivity to microscopic transient targets.

When integrating both the SPA block and ATFL, the complete SPA-DETR architecture achieves 86.8% in $mAP_{50:95}$ and 85.6% in AP_5 . Compared to the native RT-DETR-R18 baseline, this represents a substantial absolute improvement of 4.9% and 5.2% in $mAP_{50:95}$ and AP_5 , respectively. Simultaneously, the model maintains a real-time inference speed of 235.2 FPS, confirming that the proposed approach successfully balances high-precision detection with the practical constraints of real-world deployment [40].

Consistent with our rigorous evaluation protocol, all ablation results reported in Table 1 represent the average outcomes of strictly three independent runs ($3\times$) using different random seeds. To explicitly confirm the statistical significance of our architectural enhancements, we detail the standard deviations for the primary $mAP_{50:95}$ metric across different configurations: Baseline ($\pm 0.16\%$), +SPDConv ($\pm 0.17\%$), +PPA ($\pm 0.14\%$), +SPA-Block ($\pm 0.15\%$), and +ATFL ($\pm 0.18\%$). As demonstrated, the progressive performance gains achieved by each proposed module (e.g., an absolute +1.3% from SPDConv and an additional +1.5% from the integrated SPA-Block) significantly exceed these marginal fluctuations. This confirms that the effectiveness of our proposed components is statistically robust and not an artifact of random seed selection.

4.4. Comparison with State-of-the-Art

To evaluate the detection performance of SPA-DETR, we benchmark it against mainstream object detectors on the RFUAV dataset. The comparative models include the widely used YOLO series (YOLOv8s, YOLOv8m [41], and YOLOv10m), the native RT-DETR-R50 equipped with a deeper backbone, and a recent specialized small-object variant, EABI-DETRv2 [42]. Table 2 summarizes the quantitative results.

Table 2. Quantitative comparison with representative object detectors on the RFUAV dataset.

Model	$mAP_{50:95}$	mAP_{50}	AP_5	Params (M)	GFLOPs	FPS
YOLOv8s	82.8	95.3	82.0	11.13	28.5	997.2
YOLOv8m	85.5	95.4	84.7	25.85	78.7	421.5
YOLOv10m	80.6	95.9	79.0	16.47	63.5	400.4
RT-DETR-R18	81.9	99.7	80.4	19.90	57.2	235.6
RT-DETR-R50	86.0	99.8	84.8	41.97	125.7	158.3
EABI-DETRv2	84.7	99.7	83.3	13.56	85.4	224.6
Ours (SPA-DETR)	86.8	99.8	85.6	23.74	71.1	235.2

Analyzing the quantitative results reveals that SPA-DETR achieves consistent improvements across all accuracy metrics. Compared to YOLOv8m, which obtains high accuracy

among the comparative baselines (yielding an $mAP_{50:95}$ of 85.5%), our model achieves a superior performance of 86.8%. For the small-object-specific AP_5 , SPA-DETR outperforms all comparative models with a score of 85.6%. These results confirm that the synergistic design, combining spatial feature preservation with asymmetric gradient scaling, effectively handles microscopic RF footprints.

It is worth noting that all evaluated models achieve a near-saturated mAP_{50} (exceeding 95%). This phenomenon is intrinsically tied to the visual characteristics of RF spectrograms. Unlike natural images with complex semantic textures, RF signals appear as distinct high-energy discrete regions with relatively simple morphologies. Consequently, achieving a loose localization ($\text{IoU} \geq 0.5$) is a straightforward task for modern deep learning detectors. However, the true challenge in practical electronic countermeasures is not merely finding the general location of a signal, but precisely delineating the boundaries of microscopic, transient pulses amid severe ambient noise. This difficulty is accurately reflected in the more stringent $mAP_{50:95}$ and AP_5 metrics, where models demonstrate significant performance divergence. This observation explicitly justifies our architectural focus on optimizing for these stricter criteria rather than the saturated mAP_{50} .

Furthermore, the comparative experiments highlight a critical observation regarding model scaling and efficiency. When expanding the native RT-DETR from a ResNet-18 to a ResNet-50 backbone, the parameter count more than doubles (from 19.90 M to 41.97 M), and the $mAP_{50:95}$ increases to 86.0%. Although the deeper backbone provides a 4.1% gain over the R18 baseline, the proposed SPA-DETR (86.8%) still surpasses this much heavier RT-DETR-R50 by 0.8% in $mAP_{50:95}$ using only 23.74 M parameters. This comparison demonstrates that targeted architectural interventions, such as the SPA block, are significantly more parameter-efficient than conventional model scaling for decoding time–frequency targets.

Beyond standard baselines, to comprehensively evaluate our method against specialized architectures, we included EABI-DETRv2, a recent state-of-the-art RT-DETR variant explicitly designed for detecting microscopic aerial objects. It incorporates efficient multi-scale attention, bidirectional feature fusion, and the Focaler-MPDIoU loss function to optimize hard sample localization. As shown in Table 2, despite these specialized structural optimizations and advanced regression loss, our proposed SPA-DETR consistently achieves superior performance. For instance, while EABI-DETRv2 yields an impressive $mAP_{50:95}$ of 84.7% and an AP_5 of 83.3%, SPA-DETR maintains a distinct advantage. This empirically demonstrates that the integration of the SPA block and our proposed ATFL provides a more robust and accurate solution for microscopic target detection in complex electromagnetic environments than existing domain-specific alternatives.

It is worth noting that all performance metrics reported in Table 2 represent the average outcomes of strictly three independent runs ($3\times$) using different random seeds. To provide a comprehensive variance analysis and avoid point-estimate reporting, we detail the standard deviations for the primary $mAP_{50:95}$ metric across all comparative baselines: YOLOv8s ($\pm 0.14\%$), YOLOv8m ($\pm 0.17\%$), YOLOv10m ($\pm 0.22\%$), RT-DETR-R18 ($\pm 0.16\%$), EABI-DETRv2 ($\pm 0.19\%$), and the strongest baseline RT-DETR-R50 ($\pm 0.18\%$). In comparison, our proposed SPA-DETR exhibits a variance of $\pm 0.15\%$. Crucially, our absolute improvement margins significantly exceed these fluctuations. For instance, the 0.8% absolute margin over RT-DETR-R50 is substantially larger than its $\pm 0.18\%$ variance. This confirms that our broader superiority over all evaluated models is statistically robust and originates directly from our architectural enhancements rather than experimental variance.

Finally, while substantially improving detection accuracy, SPA-DETR maintains a practical balance between computational complexity and inference speed. Its computational requirement of 71.1 GFLOPs is lower than that of YOLOv8m (78.7 GFLOPs). Although the lightweight YOLO variants operate at higher absolute frame rates (e.g., YOLOv10m

at 400.4 FPS), the 235.2 FPS throughput of SPA-DETR far exceeds the real-time processing threshold required by practical electronic countermeasure systems. Overall, SPA-DETR establishes a highly favorable trade-off between speed, parameter efficiency, and accuracy for microscopic signal detection.

4.5. Qualitative Evaluation and Visualization

To provide an intuitive understanding of the practical detection performance, Figure 3 presents a qualitative comparison between the baseline RT-DETR-R18 and the proposed SPA-DETR. We specifically select RT-DETR-R18 to serve as the native baseline, aiming to visually demonstrate the direct architectural improvements brought by our method. The selected test samples primarily contain microscopic frequency-hopping signals embedded within complex background noise, representing the most challenging scenarios in practical electronic countermeasures.

Visual inspection of these scenarios reveals a significant difference in detection reliability. As highlighted in the left column of Figure 3, RT-DETR-R18 is prone to missed detections (false negatives) when encountering transient frequency-hopping pulses with minimal physical footprints. This phenomenon directly reflects the spatial information loss typical of traditional convolutional networks during deep feature extraction [43]. In contrast, the right column demonstrates that SPA-DETR consistently captures these extremely small targets. This visual evidence corroborates the effectiveness of the SPA block in preserving spatial details and isolating weak signals from severe ambient interference.

The improved performance can be attributed to the synergy between architectural preservation and optimization. While standard detectors often struggle to distinguish microscopic signals from background clutter, SPA-DETR establishes a clear separation between the target and noise. This is consistent with our quantitative findings, proving that the integration of the SPA block and ATFL effectively enhances the network's sensitivity to transient, low-energy targets. These visualization results further confirm the robustness of the SPA-DETR architecture when processing complex time–frequency features in real-world environments.

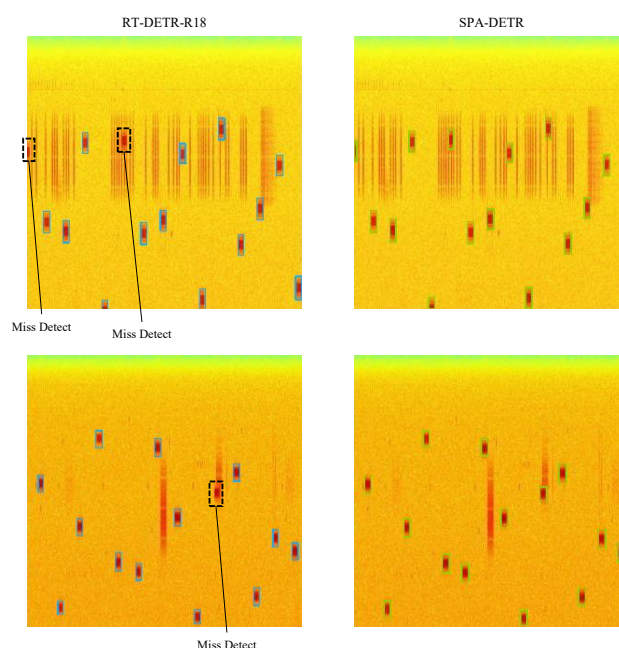


Figure 3. *Cont.*

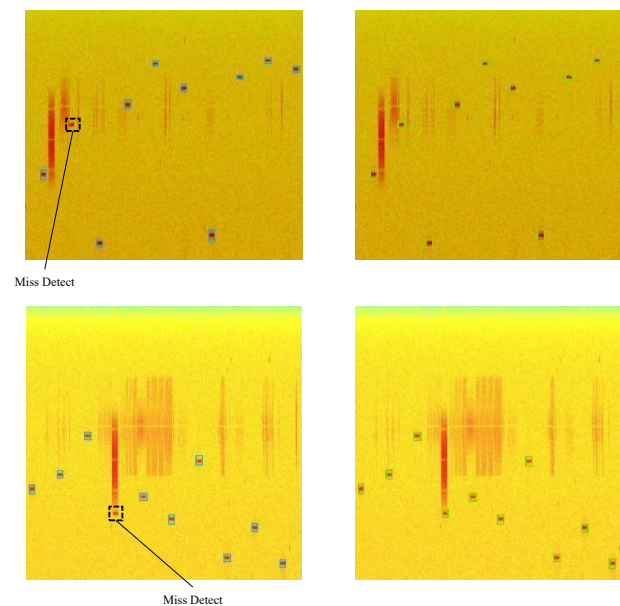


Figure 3. Qualitative detection results on the RFUAV dataset. The left column shows the detection results of the baseline RT-DETR-R18, where dashed boxes indicate missed detections (false negatives). The right column demonstrates the successful localizations achieved by the proposed SPA-DETR in the identical spectrograms.

5. Conclusions

This study presents SPA-DETR, a custom object detection architecture designed to localize transient UAV signals within complex RF spectrograms. To overcome the spatial information loss inherent in standard convolutional downsampling, we introduce the SPA block. By replacing conventional pooling operations across the backbone and multi-scale neck, the SPA block maintains global spatial fidelity throughout the feature hierarchy. Integrating SPDConv with a PPA module, this design establishes a closed-loop feature purification system. It successfully preserves the microscopic footprints of frequency-hopping pulses while filtering out ambient background noise.

Furthermore, to address the optimization bottleneck caused by severe foreground–background imbalance, we propose the ATFL. Operating entirely during backpropagation, ATFL mitigates gradient domination through asymmetric gradient scaling. This mechanism directs the network’s focus toward hard-to-detect microscopic targets without adding computational overhead during inference. Extensive evaluations on the RFUAV dataset validate these contributions. SPA-DETR achieves an $mAP_{50:95}$ of 86.8% and an AP_5 of 85.6%, improving upon the baseline RT-DETR-R18 by 4.9% and 5.2%, respectively. It substantially outperforms mainstream models like RT-DETR-R50 and YOLOv8m while maintaining a real-time inference speed of 235.2 FPS. Ultimately, this research demonstrates that targeted architectural modifications guided by underlying signal physics are significantly more effective for processing time–frequency targets than conventional model scaling. Future work will focus on the quantized deployment of this architecture on resource-constrained edge devices [44] and evaluate its generalizability across a wider range of electromagnetic signal morphologies [45].

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

Funding: This research was funded by the National Natural Science Foundation of China (NSFC), grant number 62301493.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw RFUAV I/Q data is available at <https://github.com/kitoweknd/RFUAV/> (accessed on 8 November 2025), reference [30]. The processed time–frequency spectrogram dataset used in this study is openly available at https://universe.roboflow.com/mmdrone-fj6em/final_2-qqpzd (accessed on 20 January 2026), reference [29].

Acknowledgments: The authors would like to thank the creators of the RFUAV dataset for making their data publicly available.

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

References

- Li, Q.; Wang, F.; Zhang, Y.; Li, R.; Shi, S.; Li, Y.; Li, X. Novel Micro-UAV Identification Approach Assisted by Combined RF Fingerprint. *IEEE Sens. J.* **2024**, *24*, 26802–26813. <https://doi.org/10.1109/JSEN.2024.3411106>.
- Zhu, G.; Briso, C.; Liu, Y.; Lin, Z.; Mao, K.; Li, S.; He, Y.; Zhu, Q. An Intelligent Passive System for UAV Detection and Identification in Complex Electromagnetic Environments via Deep Learning. *Drones* **2025**, *9*, 702. <https://doi.org/10.3390/drones9100702>.
- He, T.; Hou, J.; Chen, D. Multimodal UAV Target Detection Method Based on Acousto-Optical Hybridization. *Drones* **2025**, *9*, 627. <https://doi.org/10.3390/drones9090627>.
- Mi, Y.; Wen, R.; Wang, J. MaRNet: A UAV Radio Frequency Recognition Method Combining Mamba With ResNet. *IEEE Access* **2026**, *14*, 11648–11659. <https://doi.org/10.1109/ACCESS.2026.3656331>.
- Hua, W.; Chen, Q. A survey of small object detection based on deep learning in aerial images. *Artif. Intell. Rev.* **2025**, *58*, 1–67. <https://doi.org/10.1007/s10462-025-11150-9>.
- Oksuz, K.; Cam, B.C.; Kalkan, S.; Akbas, E. Imbalance problems in object detection: A review. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3388–3415. <https://doi.org/10.1109/TPAMI.2020.2981890>.
- Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 17–18 June 2024; pp. 16965–16974. <https://doi.org/10.1109/CVPR52733.2024.01605>.
- Sova, I.; Kozlov, O.; Kondratenko, Y.; Atamanyuk, I.; Aleksieieva, A. Burst-Aware Cascade Detection of UAV Radio-Frequency Signals Using Energy and Cyclostationary Analysis. *Appl. Sci.* **2026**, *16*, 5618. <https://doi.org/10.3390/app16115618>.
- Magiera, J. UAV Classification Based on Statistics of RF Transmissions. *Appl. Sci.* **2026**, *16*, 499. <https://doi.org/10.3390/app16010499>.
- Morge-Rollet, L.; Le Jeune, D.; Le Roy, F.; Le Roy, F.; Canaff, C.; Gautier, R. Drone Detection and Classification Using Physical-Layer Protocol Statistical Fingerprint. *Sensors* **2022**, *22*, 6701. <https://doi.org/10.3390/s22176701>.
- Ezuma, M.; Erden, F.; Anjinappa, C.K.; Ozdemir, O.; Guvenc, I. Micro-UAV detection and classification from RF fingerprints using machine learning techniques. In Proceedings of the 2019 IEEE Aerospace Conference; IEEE: Piscataway, NJ, USA, 2019. <https://doi.org/10.1109/AERO.2019.8741970>.
- Al-Sa'd, M.F.; Al-Ali, A.; Mohamed, A.; Khattab, T.; Erbad, A. RF-based drone detection and identification using deep learning approaches: An initiative towards a large open source drone database. *Future Gener. Comput. Syst.* **2019**, *100*, 86–97. <https://doi.org/10.1016/j.future.2019.05.007>.
- Xin, Y.; Mu, J.; Jing, X.; Liu, W. Radio Frequency Signal Recognition of Unmanned Aerial Vehicle Based on Complex-Valued Convolutional Neural Network. *Sensors* **2026**, *26*, 620. <https://doi.org/10.3390/s26020620>.
- Sunkara, R.; Luo, T. No More Strided Convolutions or Pooling: A New CNN Building Block for Low-Resolution Images and Small Objects. In Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD); Springer: Berlin/Heidelberg, Germany, 2022; pp. 443–459. https://doi.org/10.1007/978-3-031-26409-2_27.
- Tan, Y.; Xu, B.; Sun, J.; Xu, C.; Pan, W.; Dai, S.; Liu, H. DBYOLOv8: Dual-Branch YOLOv8 Network for Small Object Detection on Drone Image. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 124. <https://doi.org/10.14569/IJACSA.2025.01601124>.
- Qi, Y.; Yang, S.; Jia, Z.; Song, Y.; Zhu, J.; Liu, X.; Zheng, H. An Investigation of Infrared Small Target Detection by Using the SPT-YOLO Technique. *Technologies* **2025**, *13*, 40. <https://doi.org/10.3390/technologies13010040>.
- 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); Springer: Berlin/Heidelberg, Germany, 2020; pp. 213–229. https://doi.org/10.1007/978-3-030-58452-8_13.

18. Liu, J.; Cao, Y.; Wang, Y.; Dong, C. SO-RTDETR for Small Object Detection in Aerial Images. *Preprints* **2024**, 2024101980. <https://doi.org/10.20944/preprints202410.1980.v1>.
19. Zeng, P.; Yang, W. SDU-RTDETR: A small object detection algorithm for UAV perspective based on RT-DETR. In Proceedings of the Fourth International Conference on Electronics Technology and Artificial Intelligence (ETAI 2025); SPIE: Bellingham, WA, USA, 2025; Volume 13692, p. 136922P. <https://doi.org/10.1117/12.3068237>.
20. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>.
21. Xu, S.; Zheng, S.; Xu, W.; Xu, R.; Wang, C.; Zhang, J.; Guo, L. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME), Falls, ON, Canada, 15–19 July 2024. <https://doi.org/10.1109/ICME57554.2024.10687431>.
22. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 11534–11542. <https://doi.org/10.1109/CVPR42600.2020.01155>.
23. 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 (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988. <https://doi.org/10.1109/ICCV.2017.324>.
24. Li, X.; Wang, W.; Wu, L.; Chen, S.; Hu, X.; Li, J.; Tang, J.; Yang, J. Generalized Focal Loss: Learning Qualified and Distributed Bounding Boxes for Dense Object Detection. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), virtual, 6–12 December 2020; Volume 33, pp. 21002–21012. <https://doi.org/10.48550/arXiv.2006.04388>.
25. Zhang, H.; Wang, Y.; Dayoub, F.; Sunderhauf, N. VarifocalNet: An IoU-Aware Dense Object Detector. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), virtual, 19–25 June 2021; pp. 8514–8523. <https://doi.org/10.1109/CVPR46437.2021.00841>.
26. Chen, Y.; Sun, H.; Tian, L.; Yang, Y.; Wang, S.; Wang, T. Detecting infrared UAVs on edge devices through lightweight instance segmentation. *PLoS ONE* **2025**, *20*, e0330074. <https://doi.org/10.1371/journal.pone.0330074>.
27. Yang, L.; Wang, Y.; Miao, D.; Duan, X. Damage Detection Method for High-Strength Aramid Conveyor Belts. *IEEE Access* **2025**, *13*, 2671–2678. <https://doi.org/10.1109/ACCESS.2024.3520204>.
28. Yang, B.; Zhang, X.; Zhang, J.; Luo, J.; Zhou, M.; Pi, Y. EFLNet: Enhancing Feature Learning Network for Infrared Small Target Detection. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–11. <https://doi.org/10.1109/TGRS.2024.3365677>.
29. Mmdrone. Final_2 Dataset. 2025. Available online: https://universe.roboflow.com/mmdrone-f6em/final_2-qqpzd (accessed on 20 January 2026).
30. Shi, R.; Yu, X.; Wang, S.; Zhang, Y.; Xu, L.; Pan, P.; Ma, C. RFUAV: A Benchmark Dataset for Unmanned Aerial Vehicle Detection and Identification. *arXiv* **2025**, arXiv:cs:2503.09033.
31. Xu, L.; Shi, R.; Zhang, Y. A Radio-Frequency Sensor Based UAV Detection and Identification System Using Improved Vision Transformer Based Model. *IEEE Sens. J.* **2025**, *25*, 23437–23450. <https://doi.org/10.1109/JSEN.2025.3530937>.
32. 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 (ECCV)*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 740–755. https://doi.org/10.1007/978-3-319-10602-1_48.
33. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An imperative style, high-performance deep learning library. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 8–14 December 2019; Volume 32. <https://doi.org/10.48550/arXiv.1912.01703>.
34. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019. <https://doi.org/10.48550/arXiv.1711.05101>.
35. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. 2023. Available online: <https://docs.ultralytics.com/models/yolov8#overview> (accessed on 15 November 2025).
36. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), Vancouver, BC, Canada, 10–15 December 2024; Volume 37. <https://doi.org/10.52202/079017-3429>.
37. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. <https://doi.org/10.48550/arXiv.2004.10934>.
38. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
39. Zhong, Z.; Zheng, L.; Kang, G.; Li, S.; Yang, Y. Random Erasing Data Augmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 13001–13008. <https://doi.org/10.1609/aaai.v34i07.7000>.

40. Seidaliyeva, U.; Ilipbayeva, L.; Taissariyeva, K.; Smailov, N.; Matson, E.T. Advances and Challenges in Drone Detection and Classification Techniques: A State-of-the-Art Review. *Sensors* **2024**, *24*, 14248220.
41. Yaseen, M. What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector. *arXiv* **2024**, arXiv:2408.15857. <https://doi.org/10.48550/arXiv.2408.15857>.
42. Li, F.; Zhang, Y.; Fan, Y. EABI-DETR: An Efficient Aerial Small Object Detection Network. *Biomimetics* **2025**, *10*, 770. <https://doi.org/10.3390/biomimetics10110770>.
43. Feng, Q.; Xu, X.; Wang, Z. Deep learning-based small object detection: A survey. *Math. Biosci. Eng.* **2023**, *20*, 6551–6590. <https://doi.org/10.3934/mbe.2023282>.
44. Tang, S.; Zhang, L.; Liu, X.; Lv, R.; Qin, R. RFHS-RTDETR: Multi-Domain Collaborative Network With Hierarchical Feature Integration for UAV-Based Object Detection. *IEEE Access* **2025**, *13*, 121684–121695. <https://doi.org/10.1109/ACCESS.2025.3586649>.
45. Gao, N.; Zeng, T.; Chen, B.; Cai, D.; Jin, S.; Matthaiou, M. Multi-Domain Supervised Contrastive Learning for UAV Radio-Frequency Open-Set Recognition. *IEEE J. Sel. Areas Commun.* **2026**, *44*, 4083–4098. <https://doi.org/10.1109/JSAC.2026.3669136>.

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.