Journal of Systems Engineering and Electronics
Vol. 37, No. 3, June 2026, pp.767 – 778

RF-IRSynNet: cross-modal radio frequency-infrared fusion for robust UAV recognition

DUAN Yongsheng, ZHANG Junning*, XUE Lei, and XU Ying
College of Electronic Engineering, National University of Defence Technology, Hefei 230037, China

Abstract: The rapid proliferation of unmanned aerial vehicles (UAVs) has increasingly posed significant challenges for airspace security, particularly under long-range and visually degraded conditions. Effective UAV recognition is thus critical, yet current methodologies typically depend on single-sensor inputs, such as infrared (IR) imaging and radio frequency (RF) analysis, which suffer inherent limitations in complex environments. Although multimodal sensing has been explored in UAV detection, the joint exploitation of IR imagery and RF signals for UAV type recognition remains largely underexplored. The structural heterogeneity between IR and RF features presents challenges for joint representation and decision-making, which remains underexplored in previous work. To address this gap, this paper proposes RF-IRSynNet, a multimodal UAV classification framework that integrates IR imagery and in-flight RF emissions to enhance recognition performance. In RF-IRSynNet, IR images are processed using YOLOv11 to detect UAV candidates and extract structured semantic features. Meanwhile, RF signals are modeled using reservoir computing, which efficiently encodes temporal and spectral dynamics via feature sequences. These modalities are fused through an adaptive confidence-weighted soft-voting strategy, dynamically balancing their contributions based on specific tasks. Experimental results demonstrate that RF-IRSynNet outperforms both unimodal baselines and existing multimodal approaches, achieving robust classification at long ranges. The framework maintains high accuracy even with reduced training data, indicating high efficiency for real-world UAV monitoring.

Keywords: unmanned aerial vehicle (UAV) detection, multimodal, reservoir computing, YOLO, kernel canonical correlation analysis.

DOI: 10.23919/JSEE.2026.000066

1. Introduction

Unmanned aerial vehicles (UAVs) have seen a tremendous surge in civilian and commercial use, finding applications in aerial imaging [1], logistics [2], agriculture [3], and surveillance [4]. The accessibility, low cost, and rapid technological advances of UAVs have democratized their deployment for hobbyists and professionals. However, this widespread proliferation has also raised serious security and privacy concerns. Reports of unauthorized or malicious UAV activities, from incursions into no-fly zones and airport airspace to privacy invasion and contraband smuggling, are on the rise, underscoring the need for effective UAV detection [5].

Researchers have developed a variety of UAV detection and classification technologies leveraging different sensing modalities [6,7]. Each modality presents its own trade-offs [8]. Visual methods, such as electro-optical (EO) and infrared (IR) imaging, provide intuitive classification and are effective at short range, but their performance tends to degrade sharply at longer distances due to resolution loss, occlusion, and environmental interference [9]. In contrast, radio frequency (RF) sensing does not rely on line-of-sight and can passively detect and classify UAVs based on their control or telemetry emissions. This highlights its value in scenarios where visual sensing is impaired or unreliable, such as when UAVs are occluded, beyond the camera’s field of view, or flying at long distances. However, RF signals are susceptible to environmental noise and interference, and studies [10] have shown drops in classification accuracy when moving from controlled laboratory settings to real-world outdoor environments.

Recently, researchers have increasingly explored multimodal UAV detection frameworks that integrate complementary sensors [11,12]. Some studies have explored practical multi-sensor integration to reduce false positives and enhance target classification under real-world constraints [13−15]. For instance, audio-visual fusion for

Manuscript received December 25, 2025.
*Corresponding author.
This work was supported by the National Natural Science Foundation of China (62201602).

temporal reasoning has been proposed to enhance UAV detection under low-signal-to-noise ratio (SNR) conditions [16], and radar-optical combinations for trajectory tracking [17].

However, existing multimodal research has predominantly focused on UAV trajectory tracking and detection, with limited attention given to combining IR imagery and RF signals for UAV type recognition at long distances. This gap is not due to a lack of relevance, but rather to the practical difficulty of integrating these two modalities. The heterogeneity between IR and RF data is a primary challenge: IR imagery encodes spatial and semantic object-level cues [18], while RF signals capture temporal and spectral dynamics of waveform emissions [19]. This heterogeneity introduces challenges in data alignment, joint feature modeling, and dataset construction [20]. As a result, existing work avoids deep fusion of these two modalities, despite their clear complementarity, particularly in long-range or degraded conditions where either modality alone is unreliable.

To address the challenge of heterogeneous feature representation between IR imagery and RF signals, this paper proposes RF-IRSynNet, a multimodal classification framework designed for robust UAV identification by extracting and integrating discriminative UAV-specific features from each modality.

In this framework, YOLOv11 is employed to process IR images, generating bounding boxes, class labels, and confidence scores that serve as visual features. As a one-stage, end-to-end detection framework, YOLOv11 is known for its efficiency and strong performance on small targets, and has been widely adopted in UAV and aerial target recognition studies. However, vision-only approaches remain sensitive to long-range degradation and low-texture infrared imagery, which motivates the integration of RF sensing in this paper.

For the RF modality, a structured dynamic modeling pipeline is designed to capture UAV-specific signal characteristics across both time and frequency domains. Raw RF waveforms are segmented using a sliding window, and each segment is characterized by three parts: power spectral density (PSD), root-mean-square (RMS) energy, and zero-crossing rate (ZCR). These sequential features are individually modeled using reservoir computing, which efficiently encodes their temporal evolution into compact dynamic representations. To integrate heterogeneous dynamics across domains, reservoir kernel canonical correlation analysis (RKCCA) module is introduced to align and fuse the dynamic features into a unified semantic space. This design enables the RF branch to provide high-level dynamic semantics for UAV classification while maintaining low computational cost.

To integrate both modalities, a soft-voting decision layer is introduced to dynamically adjust IR and RF modality weights based on training performance, producing calibrated confidence scores and improved robustness in real-world UAV classification, especially in long-distance or partially observable scenarios where unimodal methods typically degrade.

The main contributions of this paper are summarized as follows:

(i) Multimodal framework targeting long-range UAV classification: RF-IRSynNet is proposed, a multimodal classification framework that integrates IR imagery and RF signals through a decision-level fusion strategy to achieve robust UAV type recognition, particularly under long-range and visually degraded conditions, which is underexplored in previous work.

(ii) Fusion of temporal and spectral RF features: both temporal and spectral patterns are efficiently extracted and modeled from RF signals using reservoir computing, and these extracted features are fused into a unified dynamic-feature representation. This enhances the discriminability of the dynamics of the UAV signal and supports effective cross-modal integration.

(iii) Adaptive confidence-based decision fusion: a lightweight soft-voting mechanism is implemented that adaptively adjusts modality weights based on validation performance. This strategy improves decision robustness across varying conditions by dynamically reflecting each modality's reliability.

(iv) Self-built multimodal UAV dataset: a synchronized IR-RF dataset is constructed featuring multiple UAV types and flight conditions, filling a gap in existing benchmarks and supporting standardized evaluation for multimodal UAV recognition.

The remainder of the paper is organized as follows. Section 2 reviews related work on UAV classification and model-space-based methods. Section 3 presents the proposed RF-IRSynNet framework, including the visual detection module, the RF signal modeling, and the multimodal fusion strategy. Section 4 presents experimental results and analysis. Section 5 concludes the paper and discusses future work.

2. Related works

2.1 UAV detection and classification

Recently, UAV detection becomes a critical research focus. Traditional single-modal methods, such as vision-only detection [21,22], radio frequency-based classification [23,24], or audio-only UAV tracking [25], often suffer from environmental interference such as occlusion, low light, or ambient noise. To this end, multimodal fusion has gained traction. Jovanoska et al. [17] combined radar and bearing data through multiple hypothesis tracking, offering a practical solution for integrating asynchronous and heterogeneous sensor inputs for UAV detection. Diamantidou et al. [26] presented a deep learning framework that fuses high-level features extracted from infrared, optro, and two-dimensional (2D) radar streams, resulting in improved performance compared to single-modality approaches. More recent systems integrate lightweight detectors with fiducial markers [27], or hybrid networks like fusion multimodal deep neural network (FMDNN) [28] that align heterogeneous sensor streams via multi-layer perception (MLP) and convolutional neural network (CNN) branches. Liu et al. [29] introduced GL-YOLO, enhancing YOLO with temporal consistency to capture small UAV targets, and Deng et al. [16] explored joint audio-visual alignment under weak SNR conditions. In addition, ensemble approaches have also emerged. Mscoy et al. [30] integrated acoustic signals, visible-light imagery, and wireless RF data to detect unauthorized or malicious UAVs. Later, the method is optimized with transfer learning and specialized CNN backbones to reduce training data requirements and enhance generalization across operational scenarios [31].

While these approaches demonstrate the potential of multimodal learning for UAV detection, most existing studies remain focused on UAV presence detection or trajectory tracking [32,33], rather than on fine-grained UAV type classification, particularly at extended ranges. Although some methods explore combinations of audio-visual or radar-optical modalities, the joint exploitation of RF and IR sensing, which offers complementary advantages in challenging environments, remains largely under-explored. This limits the effectiveness of current systems in long-range classification, where the performance of visible-spectrum and audio sensors degrades due to low resolution, occlusion, and environmental noise.

2.2 Model-space-based dynamic feature extraction

In recent years, model-space-based representation learning [34] has emerged as an effective framework for time-series analysis, particularly in time-sensitive or data-limited scenarios.

It involves fitting each data sample with an appropriate generative model, resulting in a suitable fitted model that captures data-inherent changing information. These fitted models then serve as more stable and parsimonious dynamic feature representations of the corresponding samples. Afterwards, classification or learning methods could be performed on these features instead of the original data.

Early studies have used auto-regressive moving average [35] and hidden Markov model [36] to connect neighboring elements in sequences. However, these approaches, designed for linear systems, struggle to capture nonlinear changing information. Addressing this, Chen et al. [34] utilized the echo state network (ESN) to fit sequential data. Results demonstrated that when applied to a "next point prediction" task, ESN exhibits considerable performance in capturing non-linear dynamics (i.e., the changing information).

Subsequently, ESN-based implementations have found application in diverse areas, including fault diagnosis of the Barcelona water network [37], ground penetrating radar (GPR) data [38,39], and various types of time-series data [40]. Recent work has further explored kernel-based similarity metrics among the extracted dynamic features [41], which has been applied to bridge time and frequency representations in signal classification tasks.

Notably, model-space-based representation learning focuses on the intrinsic dynamics of data and typically requires less training data and lower computational cost than many deep learning methods, particularly when equipped with well-designed models [42].

3. Methodology

To enhance the accuracy and robustness of UAV recognition in complex environments, RF-IRSynNet, a UAV recognition framework based on IR imagery and RF signals is proposed, integrating the spatial resolution capabilities of IR vision with the penetration advantages of RF signals, as shown in Fig. 1, where $\boldsymbol{A}_{\mathrm{PSD}}$, $\boldsymbol{A}_{\mathrm{RMS}}$, and $\boldsymbol{A}_{\mathrm{ZCR}}$ denote the high-dimensional features of the PSD sequence, RMS sequence, and ZCR sequence, respectively, all extracted using an ESN. $\boldsymbol{A}_{\mathrm{freq}}$ and $\boldsymbol{A}_{\mathrm{time}}$ are the fitted frequency domain and time domain features. The symbols Φ_f, Φ_t, Φ_R, and Φ_Z refer to the transformation functions applied to the frequency, time, RMS, and ZCR features. a_1, a_2, b_1, b_2 are the weights assigned to these differently transformed features for adaptive fusion. The proposed framework comprises three key components.

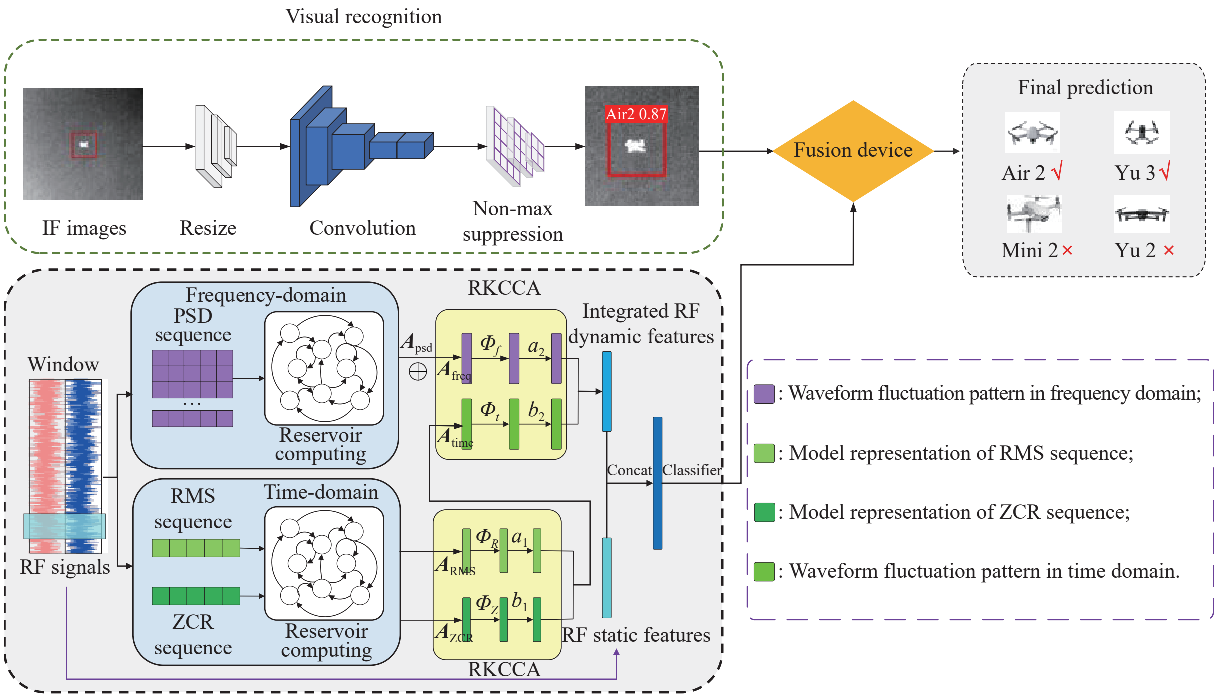


Fig. 1 Overview of the proposed RF-IRSynNet

(i) IR visual recognition: IR images are used as the primary visual modality. YOLOv11, a real-time object detection network, is employed to localize UAV candidates and output structured predictions, including bounding boxes, class labels, and confidence scores, as visual semantic features.

(ii) RF feature processing: to leverage the modal-specific patterns embedded in UAV-emitted RF signals, a reservoir computing network ESN is employed to extract dynamic features in both time and frequency domains. These features are further fused and serve as a complement to the visual modality, particularly when visual data is degraded or low-resolution.

(iii) Multimodal decision-making: to integrate features from visual and RF modalities, a confidence-weighted soft voting strategy is introduced. Modality-specific weights, derived from training-set performance, adjust each modality's confidence, and the weighted sum supports the final classification.

3.1 IR visual recognition

In the RF-IRSynNet framework, IR imagery serves as the primary visual modality for UAV detection. Compared with visible-spectrum imaging, IR is more robust to lighting variations and better suited for low-contrast scenarios, where visual degradation is common. To detect UAVs in IR images efficiently and accurately, we adopt YOLO, a unified neural network architecture capable of performing both bounding box regression and object classification within a single, end-to-end trainable model.

This design enables low-latency inference and makes it feasible for deployment on resource-constrained platforms. Among its versions, YOLOv11 is adopted in the proposed framework for its superior performance in real-time small object detection [43,44], particularly under low-resolution or noisy conditions.

Given an input IR image, YOLOv11 outputs a set of structured predictions, each represented as a tuple (b, c, s), where b is the bounding box coordinates for localizing objects in the image, c is the predicted class label, and s is the class scores for classifying the detected object. Detection performance is evaluated using standard metrics such as intersection over union (IoU) for localization accuracy and mean average precision (mAP) for overall detection quality. Detection quality is assessed using standard object detection metrics. The IoU between a predicted box b_p and a ground truth box b_{gt} is defined as

$$\mathrm{IoU}(b_p, b_{gt}) = \frac{b_p \cap b_{gt}}{b_p \cup b_{gt}}. \tag{1}$$

A prediction is considered correct if $\mathrm{IoU} > \tau$, where τ is set to 0.5. The mAP is computed by averaging precision across recall levels for each class:

$$\mathrm{mAP} = \frac{1}{C} \sum_{c \in C} \int_0^1 \mathrm{Precision}_c(r)\,\mathrm{d}r \tag{2}$$

where C is the collection of all categories, and r is the integration variable.

In the proposed framework, YOLOv11 is trained on IR images to learn discriminative visual patterns associated with different UAV targets under diverse environmental conditions. Once trained, the model is deployed to perform real-time inference on incoming infrared imagery.

For each detection, the category semantics c and detection confidence score s are retained for downstream multimodal decision-making.

3.2 Dynamic feature extraction and fusion from RF

Effectively distinguishing UAV types based on their RF emissions involves extracting discriminative features. As shown in Fig. 2, this section presents an RF dynamic feature extraction in both the time and frequency domains via ESN, followed by cross-domain feature fusion using RKCCA.

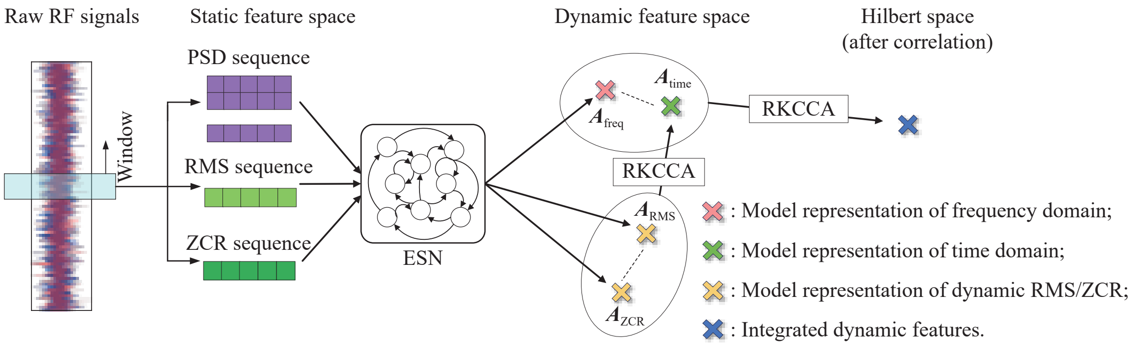


Fig. 2 Dynamic feature extraction and association mapping within the model space

(i) Dynamic feature extraction

As shown in Fig. 3, RF signals used in UAV detection typically present as high-rate, non-stationary sequences characterized by transient fluctuations and local energy bursts. Such "dynamic variations" are difficult to capture through conventional global modeling approaches. Moreover, directly feeding long RF sequences into deep learning models results in high computational cost, overfitting risk, and reduced sensitivity to short-term signal transitions.

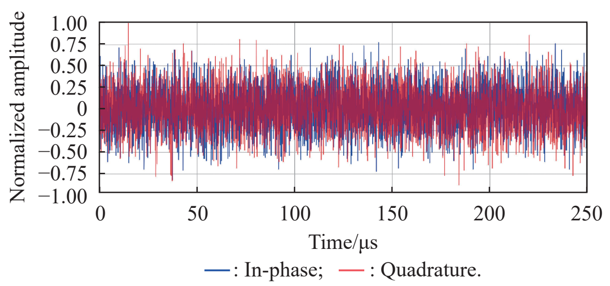


Fig. 3 Temporal transmission patterns of RF signal I/Q components

To address this problem, a fixed-size sliding window is applied to segment the raw in-phase and quadrature (I/Q) waveform into overlapping frames. Each frame retains localized spectral and temporal characteristics. For each frame, three representative features (Fig. 4) are extracted to characterize the signal's behavior across both domains: PSD for frequency-domain energy distribution, and RMS energy along with ZCR for time-domain dynamics. Specifically, PSD captures the spectral energy profile and its temporal variations, RMS reflects the signal's amplitude envelope and overall power, while ZCR quantifies local temporal variability by measuring the rate of sign changes in the waveform. These three features jointly provide a comprehensive description of the RF signal, encompassing both frequency-domain and time-domain information.

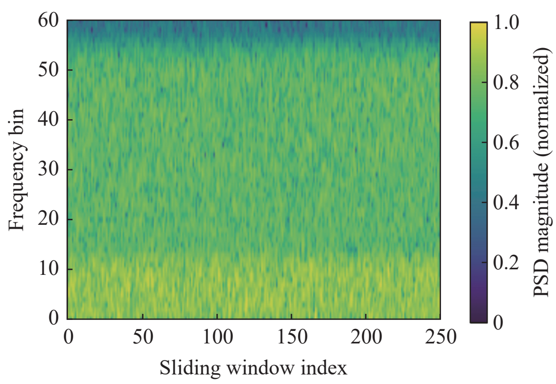


Fig. 4 Sliding-window extracted PSD features of UAV RF signals

To model the temporal evolution, i.e., the dynamics of these feature sequences, we employ the ESN, a reservoir computing network proven effective in capturing the dynamic characteristics inherent in signals. ESNs are well-suited for RF signal modeling in real-time UAV detection tasks due to the following advantages: ESNs can capture nonlinear temporal dependencies, making them suitable for representing the complex dynamics of RF time-series data; during fitting, only the output layer of the ESN needs to be fitted via ridge regression, avoiding costly gradient-based optimization and ensuring rapid application.

An ESN consists of an input layer, a randomly connected recurrent reservoir, and a linear output layer. The reservoir state $\boldsymbol{x}(t) \in \mathbf{R}^d$ is updated iteratively as

$$\boldsymbol{x}(t+1) = \tanh(\boldsymbol{W}_{\text{in}}\boldsymbol{u}(t+1) + \boldsymbol{W}\boldsymbol{x}(t)) \tag{3}$$

where $\boldsymbol{u}(t)$ is the input value at time t, $\boldsymbol{W}_{\text{in}}$ is the input weight, and $\boldsymbol{W}$ is the reservoir weight initialized to satisfy the echo state property (ESP) [45].

The output of the ESN is computed as

$$\hat{\boldsymbol{y}}(t) = \boldsymbol{W}_{\text{out}}\boldsymbol{x}(t) \tag{4}$$

where $\boldsymbol{W}_{\text{out}}$ is the output weight that maps reservoir states to outputs.

Once the entire input sequence is processed, the collected reservoir states $\boldsymbol{R} = [\boldsymbol{x}(1), \boldsymbol{x}(2), \cdots, \boldsymbol{x}(T-1)]^{\text{T}}$ are used to fit the readout model via ridge regression:

$$\boldsymbol{W}_{\text{out}} = \left(\boldsymbol{R}^{\text{T}}\boldsymbol{R} + \theta\boldsymbol{I}\right)^{-1}\boldsymbol{R}^{\text{T}}\boldsymbol{Y} \tag{5}$$

where $\boldsymbol{Y}$ is the target output $[\boldsymbol{u}(2), \boldsymbol{u}(3), \cdots, \boldsymbol{u}(T)]^{\text{T}}$, θ is the regularization parameter, and $\boldsymbol{I}$ is the identity matrix.

Through the fitting process, the readout model captures the temporal evolution of each input sequence, and the resulting output weights encode the learned dynamic behavior. Therefore, these fitted weights serve as compact and informative dynamic features of the corresponding signal.

(ii) RKCCA

Although ESNs provide compact dynamic features of individual sequences (e.g., PSD, RMS, ZCR), effectively integrating these features from both time and frequency domains remains challenging. To address this, RKCCA is introduced. Unlike traditional kernel canonical correlation analysis, which directly operates on raw features, RKCCA leverages the dynamic features extracted above to define a similarity measure in the dynamic feature space. Let the extracted dynamic feature for the sequence s_i be represented as $\boldsymbol{W}_{\text{out}}$, which captures the temporal dynamics of the sequence s_i.

To quantify the similarity between two such dynamic features, $\boldsymbol{W}_{\text{out}_i}$ and $\boldsymbol{W}_{\text{out}}$, the Euclidean distance L_2 is adopted as the dissimilarity metric in the dynamic feature space:

$$L_2\left(\boldsymbol{W}_{\text{out}_i}, \boldsymbol{W}_{\text{out}_j}\right) = \left\|\boldsymbol{W}_{\text{out}_i} - \boldsymbol{W}_{\text{out}_j}\right\|. \tag{6}$$

Based on the distance, a Gaussian kernel is constructed to form the reservoir kernel matrix:

$$\boldsymbol{K}_{ij}^{R} = \exp\left(-\gamma L_2^2(\boldsymbol{W}_{\text{out}}, \boldsymbol{W}_{\text{out}})\right) \tag{7}$$

where γ is a scaling parameter that controls the sensitivity of the kernel to differences between model representations.

Given the reservoir kernel matrices constructed from time-domain and frequency-domain dynamic features, denoted as $\boldsymbol{K}^{R,t}$ and $\boldsymbol{K}^{R,f} \in \mathbf{R}^{N\times N}$ respectively, RKCCA aims to find two projection vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, which maximize the correlation between the two modalities in the kernel-induced feature space. This can be formulated as the following regularized kernel canonical correlation analysis problem:

$$\max_{\alpha,\beta} \frac{\boldsymbol{\alpha}^{\text{T}} K^{R,t} K^{R,f} \boldsymbol{\beta}}{\sqrt{\boldsymbol{\alpha}^{\text{T}}(K^{R,t} + \lambda\boldsymbol{I})^2\boldsymbol{\alpha}}\sqrt{\boldsymbol{\beta}^{\text{T}}(K^{R,f} + \lambda\boldsymbol{I})^2\boldsymbol{\beta}}} \tag{8}$$

where λ is a regularization parameter introduced to ensure numerical stability and to alleviate overfitting, and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ represent the projection directions. In this work, λ is empirically set to 10^{-3}. The above optimization leads to a generalized eigenvalue problem. The canonical components corresponding to the largest eigenvalues are retained, and the resulting projections are used to obtain correlated dynamic representations from the time-domain and frequency-domain views.

The final fused dynamic feature is then concatenated with static spectral features and fed into a Random Forest classifier, further enhancing classification accuracy at long distances. The classifier outputs the predicted UAV category and also produces class-wise probability estimates, which are used as confidence scores to quantify the model's predictive certainty.

3.3 Multimodal decision-making

After obtaining the results from the IR visual detection and RF signal analysis, a multimodal decision-making module is proposed to enhance the robustness and reliability of UAV classification under diverse environmental conditions. By integrating classification confidences from two complementary sensing modalities, this module leverages their respective strengths to compensate for individual limitations and improve overall decision accuracy.

To make a unified decision, this module performs decision-level fusion of confidence scores from both modalities. A weighted aggregation scheme is adopted, where the contribution of each modality is determined by its validated classification accuracy. This ensures that more reliable modalities exert greater influence on the final decision, enabling the system to dynamically adapt to various operational contexts.

The fused confidence score $C_f(i)$ for each UAV class i is computed as

$$C_f(i) = \omega_v \cdot C_v(i) + \omega_{rf} \cdot C_{rf}(i) \tag{9}$$

where $C_f(i)$ and $C_{rf}(i)$ denote the confidence scores from the visual and RF modalities, respectively. The weights ω_v and ω_{rf} are derived from each modality's classification accuracy on a validation dataset and can be periodically updated based on online performance monitoring.

By integrating multimodal information in a performance-weighted manner, the proposed framework enables robust and adaptive UAV classification across varying conditions. The final decision is obtained by selecting the class with the highest fused confidence score, ensuring that the most reliable sources contribute most to the outcome. This multimodal strategy improves classification resilience in real-world scenarios, where single-modality systems may fall short.

4. Experiments

4.1 Experimental settings

The dataset used in this study is constructed through the synchronized acquisition of RF signals and IR imaging data, enabling a multimodal approach to UAV detection and recognition. It encompasses four types of widely used commercial drones (Fig. 5): DJI Air 2, DJI Mini 2, DJI Yu 2, and DJI Yu 3. Data collection is conducted across various daytime and nighttime scenarios. All drones are flown at a fixed altitude of 40 m, with two different horizontal distances: 50 m and 200 m, simulating common real-world surveillance settings.

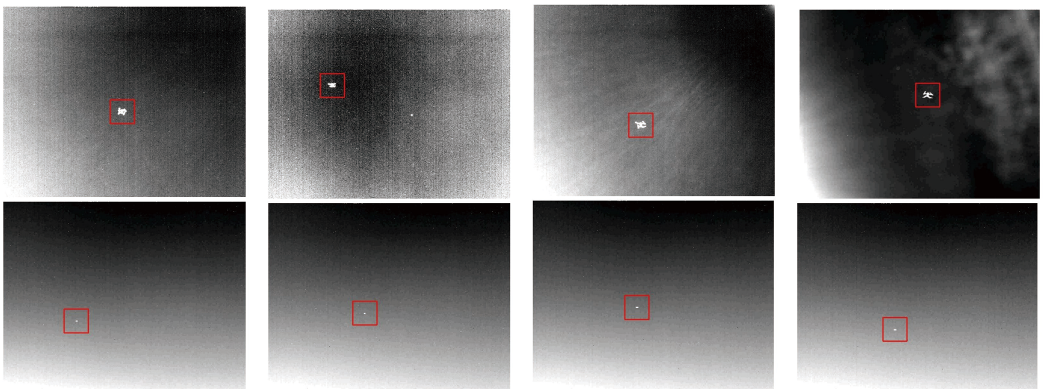


Fig. 5 UAV image dataset showing visual degradation at 50 m and 200 m

To facilitate efficient data organization and multimodal correlation, a five-tuple encoding scheme is adopted, including the device identifier, UAV type, distance, UAV action, and the session or repetition index. This systematic labeling supports fine-grained filtering and retrieval for multimodal analysis.

The RF dataset contains UAV signal recordings captured by high-sampling-rate devices. Each RF sample is stored in IQ format with a sampling rate of 40 MHz. These signals provide valuable temporal and spectral characteristics for each UAV type and scenario. Each RF sample corresponds to 200 ms of signal duration. The signal is segmented using a sliding window of 10 ms with an overlap ratio of 75% to extract time- and frequency-domain features for subsequent dynamic modeling.

The IR dataset consists of thermal images captured using a FLIR SC7000 IR camera, operating in the 3−5 μm mid-wave infrared (MWIR) range. The camera records at a resolution of 640 × 512 pixels. Images are stored in JPG format and contain thermal signatures of UAVs under diverse lighting and environmental conditions.

4.2 Experimental setup

All experiments are conducted with an AMD 7700 CPU and an NVIDIA 3090 GPU (16 GB VRAM). The software environment includes Python 3.9, PyTorch 2.0, CUDA 11.8 support for deep learning, and custom signal processing modules developed using NumPy. To facilitate the classification of UAVs using IR image data, a custom dataset is constructed and annotated.

The dataset is organized in accordance with the YOLO format, including a configuration Yet Another Markup Language (YAML) file specifying the class names and dataset paths. The dataset comprises four UAV categories, and images are resized to a uniform resolution of 512 × 512 pixels during training to ensure consistency and computational efficiency.

The model is trained for 30 epochs using stochastic gradient descent with a learning rate of 0.001, a batch size of 16, and a weight decay of 5e-4. The training loss consisted of objectness loss, classification loss, and bounding box regression loss.

An ESN is used for RF signals classification, configured with a reservoir of 200 neurons, a spectral radius of 0.9, and an input scaling factor of 0.1. The readout layer is derived via ridge regression.

4.3 Experimental results

(i) Classification performance evaluation

To comprehensively assess the effectiveness of the proposed multimodal classification framework, we perform comparative experiments against several representative baselines, including vision-only, RF-only, and multimodal approaches. All evaluations are conducted using five-fold cross-validation to ensure statistical robustness and generalizability.

Table 1 summarizes the UAV classification accuracy of different methods across varying train-test splits (50:50 to 80:20) and two distances 50 m (short-range) and 200 m (long-range), which the results are reported as mean ± standard deviation over five-fold cross-validation. The p-values are evaluated using paired t-tests on the five-fold results. The compared methods include visual-only baselines (e.g., fast region-based convolutional neural network (Faster R-CNN) [46], YOLOv11 [43,44], real-time detection transformer (RT-DETR) [47]), RF based methods (e.g., PSD + support vector machine (SVM)[48], residual network50 V2 (ResNet50 V2) [10] and RF-ESN), which serves as the RF modality branch in RF-IRSynNet, modeling the temporal dynamics of sliding-windowed RF sequences via an ESN and using the fitted reservoir readout model as a feature representation for downstream classifier), and multimodal approaches (e.g., 2DCNN+1DCNN [30], and the proposed method). For the proposed method, results are reported as mean ± standard deviation over five-fold cross-validation. Paired t-tests are conducted on the five-fold results to assess statistical significance. RT-DETR and RF-ESN are chosen as comparison baselines at 50 m and 200 m, respectively, as they achieve the best performance under the corresponding conditions.

Table 1 Detection accuracy of different methods under various conditions

Algorithm	Modality	Distance/m	Train: Test Split/%			
			50:50	60:40	70:30	80:20
Faster R-CNN [46]	Visual	50	89.9	90.2	91.7	92.5
		200	47.5	63.8	69.1	71.3
YOLOv11 [43]	Visual	50	89.3	91.8	93.3	94.5
		200	62.6	66.7	73.3	79.5
RT-DETR [47]	Visual	50	88.9	86.1	92.7	95.3
		200	60.4	68.5	74.8	81.7
PSD+SVM [48]	RF	50	55.7	56.1	58.5	59.7
		200	51.6	52.4	53.7	55.2
ResNet50 [10]	RF	50	61.6	67.8	73.9	76.0
		200	58.9	61.1	72.8	75.4
RF-ESN	RF	50	81.1	83.6	86.9	89.5
		200	79.3	81.9	84.8	88.5
2DCNN+1DCNN [30]	Visual+RF	50	84.5	86.3	88.3	92.7
		200	49.5	55.0	58.3	64.1
Proposed method	Visual+RF	50	92.6±1.3	93.8±1.1	96.7±0.8	98.0±0.5
		200	90.4±1.6	92.0±1.3	93.6±1.0	94.5±0.9
p-value (Proposed method vs. RT-DETR)		50	3.15e-3	5.33e-4	1.72e-3	5.32e-3
p-value (Proposed method vs. RF-ESN)		200	6.64e-6	1.51e-5	3.97e-5	1.19e-4

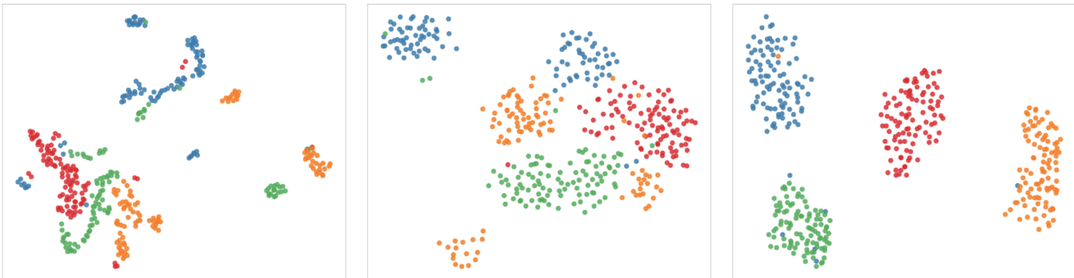
At 50 m, most visual-based models achieve high accuracy due to clear IR imagery. In comparison, the proposed method achieves the highest accuracy (92.6%−94.5%) across all splits, slightly outperforming YOLOv11 and significantly outperforming the 2DCNN+1DCNN baseline. This demonstrates the benefit of incorporating complementary RF features, even at short ranges.

At 200 m, where visual degradation becomes prominent, the gap between methods widens. Visual-only models suffer a sharp performance drop (e.g., Faster R-CNN drops to 47.5% with 50% training data), while RF-only methods remain more stable. Notably, the proposed method significantly outperforms all others at long range, reaching up to 94.5% accuracy with an 80:20 split. Even with only 50% of training data, it remains a high accuracy of 90.4%, outperforming the second-best (RF-ESN) by more than 10%.

These results validate the robustness of RF-IRSynNet under challenging long-range conditions and limited training data. The adaptive fusion of IR and RF modalities not only improves generalization but also ensures consistent performance across different data availability scenarios.

(ii) Qualitative analysis via feature visualization

Besides quantitative comparisons, qualitative analysis is further conducted to assess the discriminative power of different RF feature representations. The feature embeddings is visualized using UMAP in Fig. 6, comparing PSD-based features, ResNet50 V2 outputs, and ESN-extracted dynamic features.



(a) PSD (b) Visual embeddings (ResNet50 V2) (c) Reservoir state output (ESN)

Fig. 6 UMAP visualization of feature representations

As shown in Fig. 6(a), the PSD features obtained through sliding windows exhibit noticeable class overlap, indicating limited separability in the frequency domain. ResNet50 V2 (Fig. 6(b)) improves the global clustering structure, but certain class boundaries remain unclear. In contrast, the ESN-extracted dynamic features (Fig. 6(c)) produce compact and well-separated clusters, highlighting their effectiveness in capturing class-specific temporal dynamics of RF signals.

These findings underscore the advantage of reservoir computing for RF-based UAV recognition. Unlike static or CNN-based methods, the ESN preserves essential temporal dependencies that improve class separability in downstream tasks.

4.4 Ablation study

(i) Single-modality and multimodal integration

To validate the effectiveness of our proposed fusion strategy, three configurations are compared: (i) vision-only classification, (ii) RF-only classification, and (iii) multimodal classification using RF-IRSynNet, which receives prediction inputs from both the vision and RF modules. All experiments are conducted on a mixed dataset covering both 50 m and 200 m ranges to ensure comprehensive performance assessment across varying conditions.

As summarized in Table 2, RF-IRSynNet outperforms the unimodal baselines by integrating information from both modalities, achieving the highest accuracy and F1-score. This performance gain highlights the complementary nature of the two modalities, validating the effectiveness of the proposed dual-modality framework in improving classification accuracy and discrimination across UAV class boundaries.

Table 2 Classification performance across modalities

Model	Accuracy	Precision	Recall	F1-score
Vision only	0.8700	0.8700	0.8700	0.8712
RF only	0.8900	0.8900	0.8875	0.8800
RF-IRSynNet	0.9625	0.9626	0.9625	0.9623

To further assess robustness, we conduct five-fold cross-validation. Table 3 shows that while the vision-only and RF-only models exhibit higher variance across folds, indicating instability under different test splits. In contrast, RF-IRSynNet consistently yields the highest accuracy across all folds, with minimal variance, demonstrating strong generalization and resilience to input variations.

Table 3 Detection accuracy of different methods under various conditions

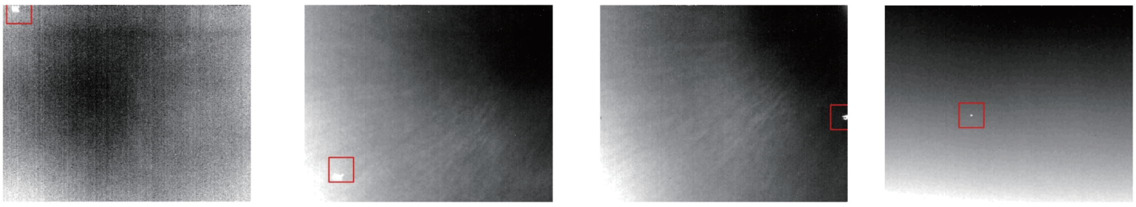
Fold	Vision only	RF only	RF-IRSynNet
Fold 1	0.8500	0.9625	0.9625
Fold 2	0.9375	0.8875	0.9750
Fold 3	0.8750	0.8625	0.9500
Fold 4	0.8375	0.9000	0.9750
Fold 5	0.8500	0.8375	0.9500
Average (Std.)	0.8700 (0.0350)	0.8900 (0.0421)	0.9625 (0.0112)

The proposed fusion framework is designed to exploit synergy between visual and RF modality through independent feature extraction pipelines and a calibrated decision-level fusion mechanism, enabling the model to mitigate the limitations of individual modalities and sustain high accuracy even when one input stream degrades.

Overall, these results confirm that the proposed multimodal design not only improves peak classification performance but also enhances robustness across scenarios.

The improvement validates the architectural choices behind RF-IRSynNet and underscores the benefit of leveraging both visual and RF modalities in long-range UAV recognition tasks.

Fig. 7 illustrates the detection performance of different models under four challenging corner cases: (i) motion blur, (ii) background similarity, (iii) incomplete target, and (iv) small target at a long distance. For each case, the ground truth label is provided. The results show that the YOLO model fails to correctly identify the target in all four scenarios (indicated by “×”), while the proposed RF model successfully detects the target in all cases (indicated by “√”). The final voting result consistently aligns with the RF prediction, demonstrating the robustness of the proposed method in handling complex real-world conditions.



(a) Motion blur
Ground truth: YU3
YOLO prediction: YU2(0.384) → ✗
RF prediction: YU3(0.857) → ☑
Voting result: YU3 → ☑

(b) Background similarity
Ground truth: Air2
YOLO prediction: Mini2(0.268) → ✗
RF prediction: Air2(0.742) → ☑
Voting result: Air2 → ☑

(c) Incomplete target
Ground truth: YU2
YOLO prediction: Air2(0.561) → ✗
RF prediction: YU2(0.805) → ☑
Voting result: YU2 → ☑

(d) Small target at a long distance
Ground truth: Air2
YOLO prediction: YU2(0.539) → ✗
RF prediction: Air2(0.723) → ☑
Voting result: Air2 → ☑

Fig. 7 Multimodal fusion resolves visual-only model failures in challenging cases

(ii) Decision-making strategy effectiveness

To assess the impact of different strategies on overall classification performance, we compare several common decision methods. The goal is to determine how the choice of decision-making mechanism influences the integration of heterogeneous modalities and the final decision quality.

Table 4 summarizes the average classification accuracy of three different strategies across five folds.

Table 4 Classification performance across modalities

Decision-making strategy	Accuracy	Precision	Recall	F1-score
Max confidence	0.9425	0.9429	0.9425	0.9425
Soft confidence	0.9575	0.9580	0.9575	0.9574
Weighted sum	0.9625	0.9626	0.9625	0.9623

Max confidence: For each input sample, the prediction from the modality with the higher confidence score is chosen as the final decision. This strategy prioritizes whichever modality expresses higher certainty, without considering historical or contextual reliability.

Soft confidence: For each input sample, the final prediction is obtained by averaging the confidence scores of both modalities. This strategy assumes both modalities contribute equally and combines their outputs in a balanced way, without favoring one over the other based solely on confidence level.

Weighted sum (the proposed method): Class confidence scores from each modality are linearly combined using fixed weights, where the weights are derived from the classification accuracy of each modality on the training set.

From the results, it is observed that the weighted sum decision-making strategy achieves the highest performance across all metrics, including accuracy, precision, recall, and F1-score, indicating not only improved correctness but also better balance between true positive coverage and false positive control. The max confidence method provides a simple and intuitive mechanism, but it can be misled by overly confident yet incorrect predictions. Soft confidence strategy improves robustness by leveraging each prediction, but lacks sample-specific adaptivity.

The proposed method goes further by evaluating the actual confidence distribution of each prediction and assigning influence accordingly. This approach enhances the model’s robustness in cases where one modality is unreliable due to occlusion, interference, or sensor noise.

5. Conclusions

This paper proposes RF-IRSynNet, a multimodal framework for long-range UAV classification by integrating infrared image and RF signal dynamics. RF-IRSynNet combines YOLOv11-based infrared detection with reservoir computing-based RF modeling to leverage the complementary strengths of both modalities, thereby addressing the limitations of unimodal methods under challenging conditions such as extended distances and partial observability. An adaptive soft-voting mechanism is introduced to dynamically calibrate modality contributions based on environmental reliability, further enhancing system robustness. Experimental results show that the proposed method outperforms both unimodal and existing multimodal approaches, achieving robust classification performance at a 200-meter range. Additionally, the model maintains high accuracy even when trained on only 50% of the available data, indicating strong efficiency and generalization capabilities. For future work, we plan to extend our approach beyond decision-level fusion by exploring feature-level integration strategies to enhance robustness and applicability in complex UAV detection scenarios.

References

[1] ZHANG H, GAO Y, WANG X, et al. UAV swarm path planning for sensor data collection via double prepartitioned deep Q-network. IEEE Sensors Journal, 2025, 25(17): 34091–34103.

[2] DUAN Y S, ZHANG J N, XUE L, et al. Embedded RF fingerprint interpretation: multi-channel complex residual networks with adaptive sphere space decision boundaries. Journal of Systems Engineering and Electronics, 2026, 37(1): 137−147.

[3] LIU J, XIANG J J, JIN Y J, et al. Boost precision agriculture with unmanned aerial vehicle remote sensing and edge intelligence: a survey. Remote Sensing, 2021, 13(21): 4387.

[4] SUN F K, NA Z Y, PEI J F. A hybrid multitask learning approach for efficient UAV signal identification. IEEE Sensors Journal, 2025, 25(17): 33064–33073.

[5] WANG B S, LI Q, MAO Q C, et al. A survey on vision-based anti-unmanned aerial vehicles methods. Drones, 2024, 8(9): 518.

[6] HUANG M G, WANG T, JING T, et al. Case-based reasoning of operation strategies recommendation for UAV swarm. Journal of Systems Engineering and Electronics, 2025, 36(6): 1548–1561.

[7] WANG P F, HU J F, HU W, et al. Anti-swarm UAV radar system based on detection data fusion. Journal of Systems Engineering and Electronics, 2024, 35(5): 1167–1176.

[8] ZHANG J, NA X, XI R, et al. mmHawkeye: passive UAV detection with a COTS mmWave radar. Proc. of the 20th Annual IEEE International Conference on Sensing, Communication, and Networking, 2023: 267−275.

[9] LAI Y C, HUANG Z Y. Detection of a moving UAV based on deep learning-based distance estimation. Remote Sensing, 2020, 12(18): 3035.

[10] PODDER P, ZAMWODNIOK M, MADRIA S. Deep learning for UAV detection and classification via radio frequency signal analysis. Proc. of the 25th IEEE International Conference on Mobile Data Manage, 2024: 165−174.

[11] ZHOU W Q, SONG C, LIU C L, et al. A prediction model of maize field yield based on the fusion of multitemporal and multimodal UAV data: a case study in Northeast China. Remote Sensing, 2023, 15(14): 3483.

[12] LIU J, XU Q Y, SU M. UAV swarm target identification and quantification based on radar signal independency characterization. Remote Sensing, 2024, 16(18): 3512.

[13] DONG Y F, WU F Y, ZHANG S J, et al. Securing the skies: a comprehensive survey on anti-UAV methods, benchmarking, and future directions. https://arxiv.org/abs/2504.11967.

[14] SEMENYUK V, KURMASHEV I, LUPIDI A, et al. Advances in UAV detection: integrating multisensor systems and AI for enhanced accuracy and efficiency. International Journal of Critical Infrastructure Protection, 2025, 49: 100744.

[15] WANG H Y, SHEN Q, DENG Z L. A diverse knowledge perception and fusion network for detecting targets and key parts in UAV images. Neurocomputing, 2025, 612: 128748.

[16] DENG T C, ZHOU Y, WU W H, et al. Multi-modal UAV detection, classification, and tracking algorithm-technical report for CVPR 2024 UG2 challenge. https://arxiv.org/abs/2405.16464.

[17] JOVANOSKA S, BROTJE M, KOCH W. Multisensor data fusion for UAV detection and tracking. Proc. of the 19th International Radar Symposium, 2018. DOI: 10.23919/IRS.2018.8447971.

[18] AASEN H, HONKAVAARA E, LUCIEER A, et al. Quantitative remote sensing at ultra-high resolution with UAV spectroscopy: a review of sensor technology, measurement procedures, and data correction workflows. Remote Sensing, 2018, 10(7): 1091.

[19] JIANG W, LIU Z, WANG Y P, et al. Realizing small UAV targets recognition via multi-dimensional feature fusion of high-resolution radar. Remote Sensing, 2024, 16(15): 2710.

[20] GUI S X, SONG S, QIN R J, et al. Remote sensing object detection in the deep learning era−a review. Remote Sensing, 2024, 16(2): 327.

[21] MUELLER M, SMITH N, GHANEM B. A benchmark and simulator for UAV tracking. Proc. of the 14th European Conference on Computer Vision, 2016: 445–461.

[22] KHAN M U, DIL M, MISBAH M, et al. Translearn-YOLOX: improved-YOLO with transfer learning for fast and accurate multiclass UAV detection. Proc of the International Conference on Communication, Computing and Digital Systems, 2023. DOI: 10.1109/C-CODE58145.2023.10139896.

[23] HE Z C, HUANG J J, QIAN G B. UAV detection and identification based on radio frequency using transfer learning. Proc. of the 8th International Conference on Computer and Communications, 2022: 1812–1817.

[24] ZHANG D L, DING G R, ZHANG J N, et al. Enhancing data augmentation diversity: a diffusion model-based approach for few-shot specific emitter identification. IEEE Trans. on Information Forensics and Security, 2025, 21: 2769–2784.

[25] DUMITRESCU C, MINEA M, COSTEA I M, et al. Development of an acoustic system for UAV detection. Sensors, 2020, 20(17): 4870.

[26] DIAMANTIDOU E, LALAS A, VOTIS K, et al. Multimodal deep learning framework for enhanced accuracy of UAV detection. Proc. of the Computer Vision Systems: 12th International Conference, 2019: 768–777.

[27] NEVES F S, CLARO R M, PINTO A M. End-to-end detection of a landing platform for offshore UAVs based on a multimodal early fusion approach. Sensors, 2023, 23(5): 2434.

[28] YANB T, LU Y, DENG H L, et al. A hybrid multimodal neural network-based anomaly detection model for UAVs. IEEE Trans. on Aerospace and Electronic Systems, 2025, 61(4): 10273−10291.

[29] LIU J Q, PLOTEGHER L, ROURA E, et al. Real-time detection for small UAVs: combining YOLO and multi-frame

motion analysis. IEEE Trans. on Aerospace and Electronic Systems, 2025, 60(5): 13419–13433.

[30] MSCOY J, RAWAL A, RAWAT D B, et al. Ensemble deep learning for sustainable multimodal UAV classification. IEEE Trans. on Intelligent Transportation Systems, 2022, 24(12): 15425–15434.

[31] MCCOY J, RAWAT D B. Optimized machine learning based multimodal UAV detection using ensemble stacking. Proc. of the IEEE 6th International Conference on Cognitive Machine Intelligence, 2024: 40–49.

[32] DU G C, ZHOU P Y, YADIKAR N, et al. Toward a dynamic tree-Mamba encoder for UAV tracking with vision-language. Knowledge-Based Systems, 2025, 323: 113731.

[33] XU Q, XU Z M, WANG H B, et al. Online learning discriminative sparse convolution networks for robust UAV object tracking. Knowledge-Based Systems, 2025, 308: 112742.

[34] CHEN H H, TINO P, RODAN A, et al. Learning in the model space for cognitive fault diagnosis. IEEE Trans. on Neural Networks and Learning Systems, 2013, 25(1): 124–136.

[35] XIONG Y M, YEUNG D Y. Mixtures of ARMA models for model-based time series clustering. Proc. of the IEEE International Conference on Data Mining, 2002: 717–720.

[36] SRIVASTAVA P K, DESAI D K, NANDI S, et al. HMM-ModE−improved classification using profile hidden Markov models by optimising the discrimination threshold and modifying emission probabilities with negative training sequences. BMC Bioinformatics, 2007, 8(1): 1–17.

[37] QUEVEDO J, CHEN H H, CUGUEROM A, et al. Combining learning in model space fault diagnosis with data validation/reconstruction: application to the Barcelona water network. Engineering Applications of Artificial Intelligence, 2014, 30: 18–29.

[38] CHEN A, ZHOU X R, FAN Y, et al. Underground diagnosis based on GPR and learning in the model space. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2023, 46(5): 3832–3844.

[39] ZHOU X R, LIU S, CHEN A, et al. Learning in CubeRes model space for anomaly detection in 3D GPR data. Proc. of the 33rd International Joint Conference on Artificial Intelligence, 2024: 5662–5670.

[40] LIU S K, ZHOU X R, CHEN H H. Multiscale temporal dynamic learning for time series classification. IEEE Trans. on Knowledge and Data Engineering, 2025, 37(6): 3543–3555.

[41] CHEN L, ZHOU X R, CHEN H H. Audio scanning network: bridging time and frequency domains for audio classification. Proc. of the AAAI Conference on Artificial Intelligence, 2024: 11355–11363.

[42] MA Q L, LI S, ZHUANG W Q, et al. Self-supervised time series clustering with model-based dynamics. IEEE Trans. on Neural Networks and Learning Systems, 2020, 32(9): 3942–3955.

[43] KHANAM R, HUSSAIN M. Yolov11: an overview of the key architectural enhancements. https://www.arxiv.org/abs/2410.17725.

[44] HAN J M, SHANG Z H, DONG Y, et al. Research on YOLO-DBO object detection method based on 1064-nm laser irradiation locking. IEEE Sensors Journal, 2025, 25(17): 32793–32803.

[45] JAEGER H. The ‘echo state’ approach to analysing and training recurrent neural networks-with an erratum note. Bonn: German National Research Center for Information Technology GMD Technical Report, 2001: 148.

[46] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards realtime object detection with region proposal networks. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137–1149.

[47] ZHAO Y A, LV W Y, XU S L, et al. DETRs beat YOLOs on real-time object detection. Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 16965–16974.

[48] KI R, KUMBASAR N, ORAL E A, et al. Drone classification using RF signal-based spectral features. Engineering Science and Technology, an International Journal, 2022, 28: 101028.