SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2026, Vol. 69, Iss. 11, 210305:1–210305:20
https://doi.org/10.1007/s11432-026-5213-1

Special Topic: Communication, Sensing, and Target Recognition for Low-Altitude Economy

RF-based multi-drone detection and classification under open-set setting

Jiajun WU[1], Yihan MAO[1], Shiyang HUA[1], Zongyu ZHANG[2*], Hangguan SHAN[1] & Zhiguo SHI[1]

[1]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China*
[2]*College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, China*

Received 15 March 2026/Revised 28 August 2026/Accepted 12 October 2026/Published online 26 October 2026

Abstract Detecting unauthorized drone intrusions via radio frequency (RF) signals is critical for mitigating severe threats to public safety. However, existing RF-based multi-drone monitoring systems predominantly rely on closed-set assumptions, inevitably misclassifying unknown drone signals as known classes or background noise, thereby failing to secure airspace against novel intruders. To address this issue, we propose a novel coarse-to-fine framework designed specifically for open-set multi-drone flight control signal (FCS) detection and classification. First, by exploiting the intrinsic morphological characteristics of FCS hopping bursts in the time-frequency domain, we design a class-agnostic proposal generation module to capture all potential signal bursts in the spectrogram, ensuring high recall for both known and unknown FCSs. Second, we construct semantics for each FCS hopping burst by fusing the texture feature, the geometry feature and the position feature. These features are refined via deformable attention modules to extract precise FCS details, followed by a global self-attention mechanism explicitly modeling the intra-frame correlations among disjoint FCS bursts. Third, we leverage a multi-objective optimization strategy to regularize the global semantic distribution and enforce intra-frame semantic consistency. Building upon this semantic distribution, the framework adaptively rejects unknown FCSs through outlier analysis and subsequently refines their categories via clustering. Extensive experiments on real-world datasets demonstrate that the proposed method significantly outperforms existing methods in terms of detection accuracy and open-set generalizability, effectively classifying both known and unknown drones.

Keywords anti-drone, multi-target recognition, open-set recognition, signal detection and classification, signal semantic learning

Citation Wu J J, Mao Y H, Hua S Y, et al. RF-based multi-drone detection and classification under open-set setting. Sci China Inf Sci, 2026, 69(11): 210305, https://doi.org/10.1007/s11432-026-5213-1

1 Introduction

Recently, the explosive growth of the drone industry in commercial and civilian sectors has unveiled the immense potential of the low-altitude economy [1–4]. While the widespread use of drones empowers various industries and offers significant convenience, illegal or unauthorized drone operations pose serious security risks to the public, including privacy infringement, aviation disruption, etc. [5,6]. To effectively govern low-altitude airspace, regulatory authorities have enforced source management measures, such as mandatory remote identification (RID) broadcasting and geo-fencing systems. However, these preventive measures are inherently ineffective against non-cooperative drones that intentionally evade supervision, necessitating the deployment of effective anti-drone systems to monitor and manage the low-altitude airspace.

To detect such threats, various sensing technologies have been developed, including radar, audio, video, and radio frequency (RF) analysis [7]. Among these methods, passive RF-based approaches are particularly promising due to their long detection range, friendliness to urban electromagnetic environments, and insensitivity to non-line-of-sight (NLoS) scenarios. Typically, the drone RF signals are composed of the video transmission signal (VTS) and the flight control signal (FCS). Since malicious operators can consciously disable the video feed to evade surveillance, existing methods based on VTS detection face the risk of stealth attacks [8]. In contrast, as the indispensable communication link continuously carries flight logs and control instructions between the drone and the ground controller, the FCS provides a persistent and reliable signature that cannot be turned off during flight. Therefore, this paper specifically focuses on the detection and classification of the FCS.

Existing RF-based open-set recognition (OSR) methodologies predominantly operate under a strict single-target assumption [9–12]. Under this premise, algorithms extract features from the entire spectrogram to construct a global

* Corresponding author (email: zongyu_zhang@zju.edu.cn)

© Science China Press 2026

semantic vector. However, in real-world deployments involving multiple concurrent drones, these methods fail to generate precise semantics for all localized FCS hopping bursts. To detect multiple FCS hopping bursts within the spectrogram, recent multi-target RF frameworks leverage general-purpose visual object detectors to detect and classify individual FCS hopping bursts [13–15]. Nevertheless, these frameworks are constrained to closed-set assumptions. They forcefully categorize every detected hopping burst and its corresponding bounding box into predefined training classes, thereby failing to detect unknown FCSs. Consequently, achieving multi-target OSR necessitates a tailored unified paradigm: the system must first explicitly detect every individual hopping burst and its spatial boundary in the spectrogram, and subsequently perform fine-grained open-set classification on these localized hopping bursts.

Compared to existing RF-based single-target OSR methods and general-purpose visual object detectors, achieving OSR in multi-target RF scenarios presents three primary challenges. (1) Generalizing valid detections to unknown FCS hopping bursts. Existing closed-set detectors rely on learnable filters optimized solely for the specific signal textures of the training set. These filters tend to exhibit negligible response to unknown FCS hopping bursts, causing an initial detection failure that fundamentally precludes any subsequent open-set classification. (2) Constructing fine-grained semantics for tiny hopping bursts. Unlike visual objects with rich textures, the drone FCS appears as small, scattered energy bursts on spectrograms. Conventional detectors operating on the entire image struggle to capture the fine-grained internal textures and precise frequency positions, leading to ambiguous semantic representations that are insufficient for OSR. (3) Refined grouping of hopping bursts for multiple unknown drones. When confronting multiple concurrent unknown drones, refined open-set classification demands the effective grouping of all scattered FCS bursts within the spectrogram. Existing methods fail in this regard by treating each FCS burst as an isolated entity and ignoring the intra-frame correlations between them. Consequently, FCS bursts from a single drone may be erroneously clustered into separate unknown classes.

To address the above-mentioned challenges, in this work we propose a two-stage spectrogram monitoring network designed for open-set multi-drone detection and classification. First, to overcome the dependency on known patterns, we introduce an energy-based region proposal generation module based on the property of the FCS, ensuring high recall for both known and unknown FCS hopping bursts. Then, we generate initial semantics for each proposal. These initial semantics are refined through a deformable transformer decoder, which sparsely interacts with the spectrogram to capture precise signal details. Furthermore, a global self-attention mechanism is employed to model the relationships between different signal bursts, which effectively groups intra-frame components. After the semantic construction, the output semantics are optimized via prototype learning to form a discriminative semantic space. Finally, the system applies adaptive thresholds to classify known signal semantics and isolate unknown outliers. These rejected unknown semantics are then directly clustered for fine-grained open-set recognition. To sum up, the main contributions of this paper are summarized as follows.

(1) An energy-based region proposal generation module is designed, which leverages the localized time-frequency energy distributions of frequency hopping spread spectrum (FHSS) bursts. This method ensures the generation of high-recall, class-agnostic region proposals, effectively capturing potential unknown signal bursts without requiring prior training on their specific patterns.

(2) A dual-source semantic initialization and refinement mechanism is proposed. By jointly modeling the texture, geometry, and position features of each proposal, and further refining these initial representations via a deformable attention mechanism, our method captures precise fine-grained FCS details, yielding discriminative semantics for accurate FCS classification.

(3) We introduce a global self-attention mechanism coupled with a novel intra-frame consistency loss to explicitly model the spatial correlations among the intra-frame discrete hopping bursts. In this way, the system groups disjoint signals emitted from the same drone, enabling robust clustering and precise classification in the presence of multiple unknown drones.

The rest of this paper is organized as follows. In Section 2, the related work is reviewed. In Section 3, we analyze the signal model for FCS. Section 4 elaborates on the proposed method. Section 5 presents the experimental results and analysis. Finally, we conclude the paper in Section 6.

2 Related work

In this section, we survey the literature relevant to RF-based drone detection and classification. We first review the existing research on RF-based drone detection and classification in closed-set scenarios, ranging from early feature-based methods to advanced deep learning frameworks designed for complex closed-set scenarios. Subsequently, we discuss the transition towards open-set learning, specifically highlighting the limitations of the existing approach

when applied to the RF-based multi-drone detection task.

2.1 RF-based drone detection and classification in closed-set scenarios

RF-based drone detection and classification primarily focus on passively monitoring the in-phase/quadrature (I/Q) RF signals emitted by the drone or its remote controller. With the rapid advancement of deep learning, research in this field has gradually shifted from manual feature engineering [16] to learning-based approaches [17–24]. Typically, raw signals are transformed into time-frequency representations using techniques such as short time Fourier transform (STFT) [18] or wavelet transform [20], from which neural networks such as convolutional neural networks (CNNs) [23] or long short-term memory (LSTM) [24] automatically extract discriminative features to perform classification. Notably, Xue et al. [25] indicated that the combination of spectrograms with real-valued CNNs is often considered an optimal choice in terms of classification accuracy, training time, and parameter efficiency. However, these early feasibility studies predominantly assume ideal channel conditions and single-target scenarios, which significantly limit their direct applicability in complex, real-world airspace.

To bridge the gap between ideal assumptions and practical deployments, subsequent research primarily focused on mitigating environmental signal degradation. Medaiyese et al. [26] adopted denoising autoencoders to enhance robustness against signal-to-noise ratio (SNR) variations. To solve the data shortage in complex scenarios, Zhang et al. [27] augmented training data by mixing pure drone signals with different environmental background signals. Furthermore, Wang et al. [28] proposed a generalized denoising and interference-removal framework, where the received signal is explicitly decomposed into the clean signal, interference, and noise components. Beyond pure data-driven denoising, some studies focus on protocol-aware analysis. By analyzing communication protocols to obtain a priori knowledge of frame structures and modulation parameters, Li et al. [29] derived more discriminative Zadoff-Chu sequence features, which allow seamless integration with signal filtering or image reconstruction algorithms to obtain pure signals. While these denoising and protocol-aware techniques significantly improve robustness against environmental interference, they still inherently operate under a single-target assumption. Consequently, they fail to scale when applied to complex scenarios involving multiple drone signals.

To overcome this single-target bottleneck and manage real-world scenarios with simultaneous multi-drone operations, the research focus has expanded from single-label classification to multi-target object detection, where RF-based drone detection is regarded as a combined task of object detection and classification on spectrograms. Basak et al. [13] were the first to propose a unified framework for multi-signal detection and localization, demonstrating that a YOLO-lite-based framework outperforms traditional methods. Addressing the issues of anchor mismatch and cross-domain adaptation, Zhao et al. [14] proposed a keypoint-based anchor-free detector integrated with a spectrogram domain adaptation generative adversarial network for robust practical deployments. Despite these architectural advancements, a major limitation is that the aforementioned frameworks are implicitly tailored for continuous, wide-band VTS. Consequently, they tend to overlook the highly dynamic and elusive FCSs, which are equally critical for drone detection. To fill this gap, the SMNet framework [15] is proposed as a specialized FCS-oriented solution. It integrates noise-insensitive edge and corner features to robustly localize the faint FCS hopping bursts, and employs mixup-based data augmentation to enable multi-drone detection while only single-drone data is available for learning. Chen et al. [30] further developed a four-stage optimized local-maximum method that combines multilevel peak extraction, dynamic frequency-block updating, and kurtosis filtering to extract multiple FCSs under low SNR, multisignal coexistence, and spectral overlap, with an efficient field-programmable gate array (FPGA) implementation. Nevertheless, this method focuses on signal extraction rather than category recognition and does not address unknown-class discrimination.

Despite these significant improvements in robustness and multi-target capability, the majority of existing work operates under a closed-set assumption, assuming that all testing samples belong to the classes seen during training. In reality, drone models are iterated and updated rapidly, making it difficult for datasets and algorithms to keep pace with the new emerging drones. Consequently, the inability to detect and classify unknown FCSs poses a security risk, necessitating the exploration of OSR techniques.

2.2 Open-set learning for drone detection and open-world object detection

Different from the closed-set assumption, OSR requires the system to not only accurately classify known classes but also effectively reject unknown classes that are unseen during training. In closed-set classification networks, the Softmax function inherently forces the output into a closed probability distribution, assigning high confidence even to unknown inputs. To mitigate this, Soltani et al. [31] employed thresholds derived from the SoftMax probabilities and classification ratios of multiple neural networks. The OpenMax layer [10] fits Weibull distributions to activation vectors to estimate the probability of an input being unknown. Similarly, the uncertainty-inspired

open-set model [12] parameterizes the output feature evidence into a Dirichlet distribution for uncertainty quantification. To further enhance open-set adaptability, Li et al. [32] proposed a multi-modal marginal prototype framework combined with a generative adversarial network (GAN)-based generation strategy to synthesize virtual unknown samples. Despite the efficiency in rejecting unknown samples, these methods fail to achieve fine-grained discrimination for multiple unknown classes, yielding an incomplete perception of the monitored airspace.

For refined classification of specific unknown classes, SR2CNN [9] utilizes auto-encoders to map signals into a feature space, classifying them based on Mahalanobis distance thresholds. Han et al. [33] improved this by employing the Euclidean distance to better recognize unknown jamming patterns. However, many of these methods rely on validation data from the open-set to determine thresholds, which is impractical when unknown data is unavailable. To tackle this issue, the S3R framework [11] introduces an outlier analysis-based semantic classifier to determine self-adaptive bounding thresholds solely from known semantics, featuring dual-path feature extraction for spectrograms. Furthermore, it proposes a semantic augmentation method that leverages intermediate features to sparsely represent unknown semantics, thereby significantly improving the refined classification of unknown drones. Additionally, Mao et al. [34] proposed a multi-domain prototype network to address robust OSR under few-shot settings. Guo et al. [35] proposed LOSDR, which jointly extracts local FCS-block details and global frequency-hopping patterns, aligns RF semantics with CLIP-derived text embeddings, and uses class decorrelation and adaptive entropy thresholding to improve robustness under attenuation and frequency-hopping interference with only 3.7M parameters. Gao et al. [36] fused ResNet-based texture features with Transformer-based time-frequency position features through supervised contrastive learning and constructed Open-RFNet using an improved generative OpenMax strategy; by freezing the feature extractor and retraining only the classification head with generated unknown samples, the method balances known- and unknown-class recognition with limited additional training overhead. However, a fundamental limitation shared by all the aforementioned RF-based methods is the single-target assumption. They process the input signal as a single instance, failing to function in realistic multi-target scenarios where multiple drone signals overlap in the time-frequency domain.

To address multi-target detection and classification under an open-set setting, Joseph et al. [37] proposed the concept of OWOD together with a learning framework based on contrastive clustering and an energy-based unknown detector. Subsequently, Gupta et al. [38] proposed OW-DETR, a transformer-based end-to-end detector that utilizes attention-driven pseudo-labeling to learn a separation between foreground objects and background. More recently, Zhang et al. [39] modeled the joint distribution of latent objectness and category information and introduced a dynamic Gaussian prior to stabilize training in low-data regimes, thereby reducing unknown detection uncertainty and misclassification rates. Nevertheless, directly applying these OWOD frameworks to RF-based multi-drone detection poses a significant challenge. The aforementioned OWOD methods rely heavily on the existence of unlabeled latent objects in the training images to generate pseudo-labels for unknown classes. In contrast, in RF-based drone detection tasks, the training data only consists of signals from known classes, and signals from unknown drones are entirely unavailable during the training phase. This fundamental data discrepancy makes existing work not directly applicable, highlighting the urgent need for a tailored OSR framework specifically designed for RF signals.

3 Drone RF signal analysis

Typically, when a drone operates in the monitored airspace, the raw signal $s(t)$ over a time interval T can be expressed as

$$s(t) = s_{\mathrm{FCS}}(t) + s_{\mathrm{VTS}}(t) + s_i(t) + n(t), \tag{1}$$

where $s_{\mathrm{FCS}}(t)$ represents the FCS component, $s_{\mathrm{VTS}}(t)$ represents the VTS component, $s_i(t)$ refers to the signals from other interfering sources such as Wi-Fi and Bluetooth, and $n(t)$ refers to the inherent noise of the RF receiver. The FCS $s_{\mathrm{FCS}}(t)$ predominantly utilizes FHSS to ensure robustness against interference, which can be modeled as a collection of K discrete, short-duration RF bursts $b_k(t)$ hopping across the frequency band, denoted as

$$s_{\mathrm{FCS}}(t) = \sum_{k=1}^{K} b_k(t) = \sum_{k=1}^{K} A \cdot x_k(t) \cdot W_\tau(t - t_k) \cdot \cos(2\pi f_k t + \phi_k), \tag{2}$$

where A is the amplitude of the signal, $x_k(t)$ denotes the baseband modulated signal for the k-th hopping burst. The modulation method for $x_k(t)$ differs, such as DJI Ocusync, DJI Lightbridge, etc., $W_\tau(\cdot)$ is the rectangular window function with dwell time τ, t_k is the precise start time of the k-th hopping burst, and f_k and ϕ_k represent the center

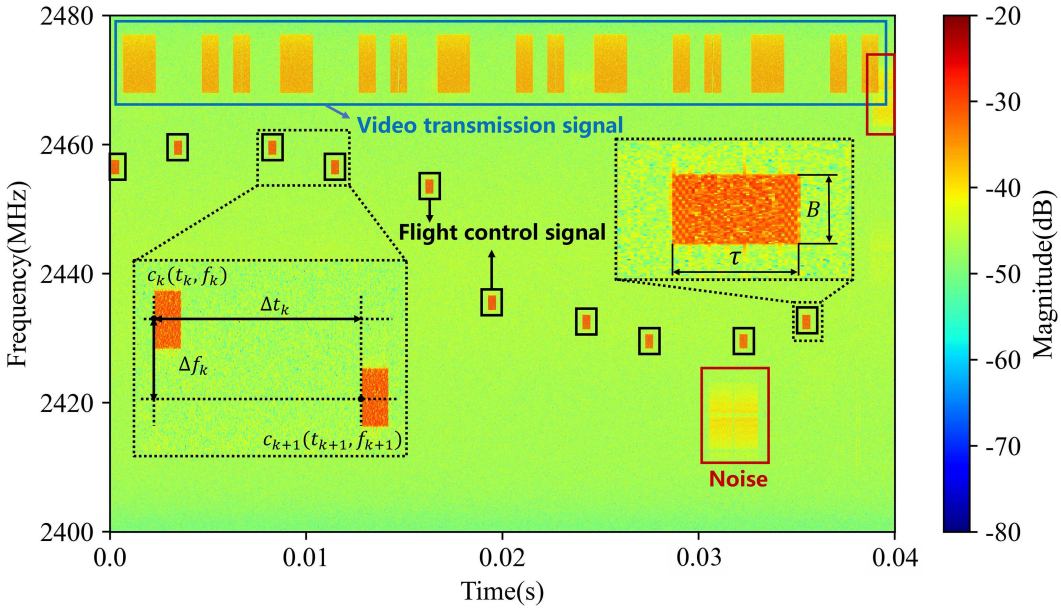

Figure 1 Typical time-frequency spectrogram of drone signals, illustrating the distinct patterns of the VTS and the FCS.

frequency and the initial phase for the k-th hopping burst. The continuous-time signal $s(t)$ is then captured by the RF receiver to obtain the discrete-time raw I/Q sequence with length N, denoted as $s[n]$.

To decouple the drone RF signal from the received raw signal in the time domain, we convert $s[n]$ into a time-frequency spectrogram $\boldsymbol{X}$ using STFT. By sliding an analysis window of length Q across $s[n]$ and computing the Q-point discrete Fourier transform at each step, the spectrogram $\boldsymbol{X}$ is calculated as

$$[\boldsymbol{X}]_{p,q} = 20\log_{10}\left|\sum_{i=0}^{Q-1} s[i+pR]\cdot g_Q[i]\cdot \mathrm{e}^{-\mathrm{j}2\pi\frac{q}{Q}i}\right|, \tag{3}$$

where $p = 0, 1, 2, \ldots, P-1$ is the time index with P being the total number of sliding steps, $q = 0, 1, 2, \ldots, Q-1$ is the frequency index, R represents the sliding length of the short-time analysis window, g_Q is the Hamming window function with length Q, and $|\cdot|$ calculates the magnitude of the complex number.

Figure 1 illustrates a typical drone spectrogram, in which the FCS appears as narrow-band, short-duration, discrete bursts hopping across the time period and the frequency band. We characterize the FCS through three key types of features, namely the texture feature, the geometric feature, and the position feature. Specifically, the texture feature refers to the internal energy distribution within a single hopping burst, which is primarily derived from the baseband modulated component $x(t)$. Different types of drones employ various modulation settings for the baseband signal $x(t)$, which manifest as unique internal signal textures within this single FCS burst. The geometric feature is the 2-dimensional (2D) shape of each burst in the spectrogram, determined by the instantaneous bandwidth B of $x(t)$ and the dwell time τ defined by the window function $W_\tau(\cdot)$. In addition, the position feature encodes the distinct frequency-hopping patterns, i.e., the global distribution of all hopping bursts in the spectrogram. These unique patterns can be jointly governed by the absolute time-frequency position of all hopping bursts, denoted as $\{(t_k, f_k)\}$, and the relative position correlations between adjacent bursts, represented by the hopping interval $\{\Delta t_k\}$ and the frequency stride $\{\Delta f_k\}$.

4 Methods

The overview of the proposed method is illustrated in Figure 2. To tackle the challenges of open-set FCS detection and classification, we propose a coarse-to-fine two-stage strategy. The first stage integrates an energy-based module to generate class-agnostic proposals with high recall, and subsequently constructs initial semantics for each proposal. The second stage utilizes a Transformer-based architecture to refine the semantics and the bounding box for each signal. The refined semantics are then optimized via prototype learning to form a discriminative semantic space, which is used for the classification of known and unknown drones.

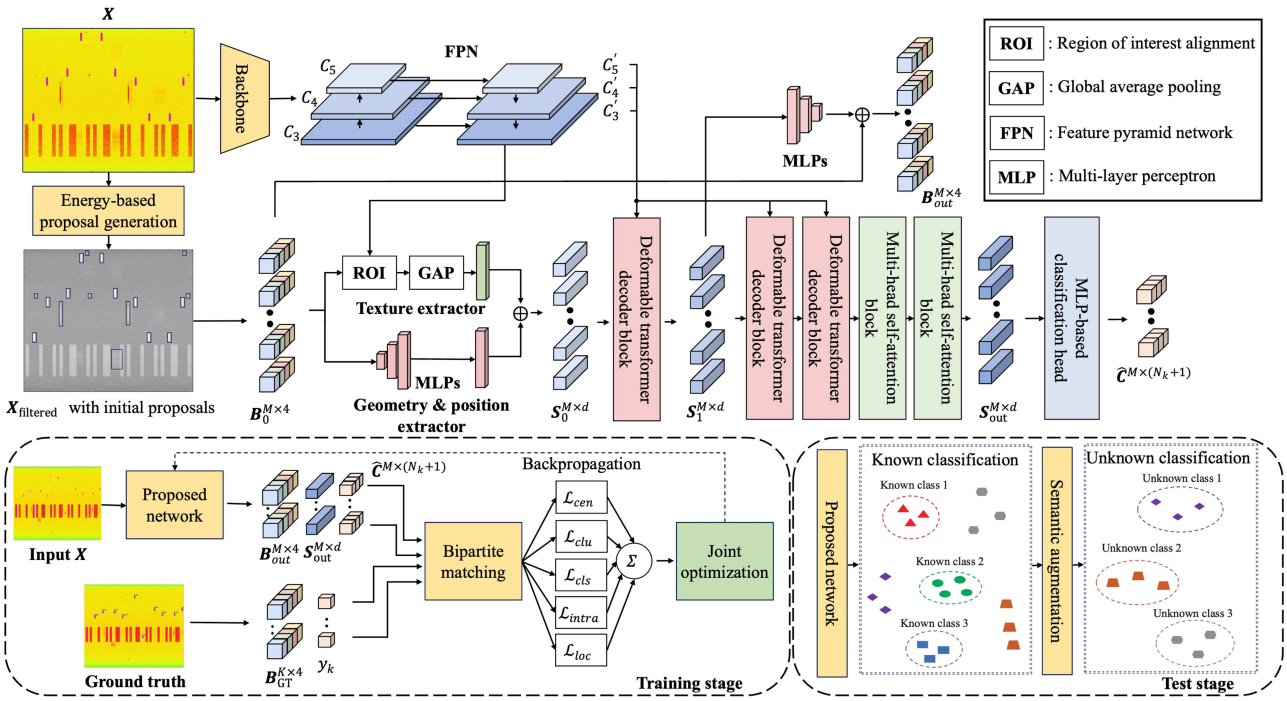

Figure 2 Overview of the proposed framework.

4.1 Stage I: signal semantic generation

To guarantee high recall for all FCS hopping bursts, this stage aims to detect all candidate signal locations and instantiate their initial semantics based on the backbone features. It integrates the energy-based proposal generation for physical localization, the backbone for feature encoding, and the semantic construction module for semantic initialization.

4.1.1 *Energy-based proposal generation*

In the time-frequency domain, individual RF signal bursts are inherently bounded by their transmission time duration and bandwidth, naturally manifesting as localized rectangular energy blocks in the spectrogram. However, raw spectrograms are susceptible to high-frequency background noise and internal signal texture variations, which complicates precise boundary extraction. To mitigate this, we first employ morphological filtering on the spectrogram, which can be expressed as

$$\boldsymbol{X}_{\text{filtered}} = (\boldsymbol{X} \bullet B) \circ B = ((\boldsymbol{X} \oplus B) \ominus B) \ominus B \oplus B, \tag{4}$$

where B represents the rectangular structuring kernel, and $\bullet$, $\circ$, $\oplus$, and $\ominus$ denote the closing, opening, dilation, and erosion operators, respectively.

Physically, the closing step bridges micro-gaps within a single signal burst, and the opening step acts as a spatial filter to remove isolated impulse noise points. Leveraging the robustness of edge features against noise demonstrated in our previous work [15], we utilize the Canny operator to extract the boundaries of $\boldsymbol{X}_{\text{filtered}}$. The contours of these boundaries are then converted into bounding boxes. Bounding boxes that are excessively large (e.g., wideband Wi-Fi signals) or extremely small (e.g., residual impulse noise) are subsequently discarded. Let M denote the number of boxes retained after filtering. These valid bounding boxes then serve as the initial class-agnostic proposals, formulated as $\boldsymbol{B}_0 = \{\boldsymbol{b}_m^{(0)}\}_{m=1}^{M}$.

4.1.2 *Backbone*

To transform the input spectrogram into deep feature representations suitable for semantic extraction, we employ a composite backbone network consisting of a ResNet-18 [40] and a feature pyramid network (FPN) [41]. Specifically, the ResNet-18 extracts features from the spectrogram $\boldsymbol{X}$ through a series of convolutional layers with residual connections, yielding a hierarchy of feature maps $\boldsymbol{C} = \{C_3, C_4, C_5\}$, where the spatial resolution decreases as the channel depth increases. C_3, C_4, and C_5 correspond to the outputs of layer1, layer2, and layer3, with channel

dimensions $(64, 128, 256)$ and downsampling strides $(4, 8, 16)$, respectively. Given that FCS bursts occupy a small proportion of the spectrogram, we introduce FPN to mitigate the potential information loss of FCS bursts in deep layers. It upsamples the deeper, semantically stronger feature maps and fuses them with the shallower, spatially precise feature maps via lateral connections. This process produces a set of multi-scale enhanced feature maps, denoted as $\boldsymbol{C'} = \{C'_3, C'_4, C'_5\}$.

Specifically, the FPN uses 1×1 lateral projections, two-fold nearest-neighbor upsampling, and 3×3 smoothing convolutions:

$$\begin{aligned} &T_5 = \mathrm{Conv}_{1\times1}(C_5), \\ &T_i = \mathrm{Conv}_{1\times1}(C_i) + \mathcal{U}_2(T_{i+1}), \quad C'_i = \mathrm{Conv}_{3\times3}(T_i), \end{aligned} \tag{5}$$

where $i \in \{4, 3\}$ for the top-down recursion and $C'_5 = \mathrm{Conv}_{3\times3}(T_5)$. The highest-resolution feature C'_3 is used for ROIAlign-based semantic initialization, while $\{C'_3, C'_4, C'_5\}$ are used by the deformable decoder for subsequent localization and classification.

4.1.3 *Semantic initialization*

Given the feature maps $\boldsymbol{C'}$ and the initial proposals $\boldsymbol{B}_0$, we construct the initial semantic $\boldsymbol{S}_0$ by fusing texture features and geometric positions. First, we utilize the region-of-interest (ROI) align operator to extract texture features $\boldsymbol{s}_t^{(m)} \in \mathbb{R}^d$ from the multi-scale FPN feature map C'_3 based on the initial bounding boxes $\boldsymbol{b}_m$. These features are flattened and projected to dimension d. Second, the proposals $\boldsymbol{b}_m$ are encoded via a multi-layer perceptron (MLP) to extract the geometry feature and the position feature, denoted as $\boldsymbol{s}_p^{(m)} \in \mathbb{R}^d$. The initial semantic $\boldsymbol{s}_0^{(m)} \in \mathbb{R}^d$ for the m-th proposal is computed as

$$\boldsymbol{s}_0^{(m)} = \boldsymbol{s}_t^{(m)} + \boldsymbol{s}_p^{(m)} = \underbrace{\mathrm{GAP}(\mathrm{ROIAlign}(C'_3, \boldsymbol{b}_m))}_{\text{texture part}} + \underbrace{\mathrm{MLP}(\boldsymbol{b}_m)}_{\text{position part}}, m = 1, 2, \ldots, M, \tag{6}$$

where $\mathrm{GAP}(\cdot)$ refers to the global average pooling operation. The aggregate initial semantic set is denoted as $\boldsymbol{S}_0 = \{\boldsymbol{s}_0^{(1)}, \ldots, \boldsymbol{s}_0^{(M)}\}$, where d is the dimension of each semantic.

4.2 Stage II: semantic interaction and refinement

Building upon the initial signal semantic $\boldsymbol{S}_0$ and corresponding bounding boxes $\boldsymbol{B}_0$ generated in stage I, stage II aims to refine $\boldsymbol{B}_0$ and transform $\boldsymbol{S}_0$ into a more discriminative semantic representation $\boldsymbol{S}_{\mathrm{out}}$. This transition operation is achieved through a cascaded architecture comprising the deformable transformer decoder layers and global self-attention (SA) layers.

The initial semantic representations $\boldsymbol{S}_0$ are processed by L_{dec} deformable decoder layers. Let $\Phi_l(\cdot)$ denote the operation of the l-th layer. The semantic representations are updated recursively as

$$\boldsymbol{S}_l = \begin{cases} \Phi_1(\boldsymbol{S}_0, \boldsymbol{B}_0), & l = 1, \\ \Phi_l(\boldsymbol{S}_{l-1}, \boldsymbol{B}_{\mathrm{out}}), & 2 \leqslant l \leqslant L_{\mathrm{dec}}, \end{cases} \tag{7}$$

where $\boldsymbol{B}_{\mathrm{out}}$ denotes the refined boxes obtained after the first decoder layer. We define the output of the final layer $\boldsymbol{S}_{L_{\mathrm{dec}}}$ by $\boldsymbol{S}_{\mathrm{dec}}$.

Specifically, each layer Φ_l is composed of an SA module, a deformable cross-attention (DCA) module [42], and a feed-forward network (FFN). In the DCA module, the network adaptively samples N_{point} key sampling points around the center of the reference box $\boldsymbol{B}$ on the feature maps, aggregating local visual contexts efficiently. The SA module is then employed to model the global interactions among different proposal queries. Subsequently, the FFN, which is implemented as an MLP, is used for feature projection.

To construct robust signal semantics that are invariant to initial localization errors, we operate box refinement after the first layer. For the first layer Φ_1, the operation interacts with the initial proposals $\boldsymbol{B}_0$ to generate the intermediate semantic $\boldsymbol{S}_1$ as

$$\boldsymbol{S}_1 = \Phi_1(\boldsymbol{S}_0, \boldsymbol{B}_0) = \mathrm{MLP}\Big(\mathrm{Norm}\big(\mathrm{SA}(\mathrm{DCA}(\boldsymbol{S}_0, \boldsymbol{B}_0, \boldsymbol{C'}))\big)\Big), \tag{8}$$

where $\mathrm{Norm}(\cdot)$ denotes the layer normalization operation. Then, $\boldsymbol{S}_1$ is fed into an MLP to predict the coordinate offsets $\Delta\boldsymbol{B}$. The refined bounding boxes $\boldsymbol{B}_{\mathrm{out}} = \{\hat{\boldsymbol{b}}_m\}_{m=1}^M$, which serve as the aligned spatial anchors for all subsequent layers, are updated as

$$\boldsymbol{B}_{\mathrm{out}} = \boldsymbol{B}_0 + \Delta\boldsymbol{B} = \boldsymbol{B}_0 + \mathrm{MLP}(\boldsymbol{S}_1). \tag{9}$$

With the aligned anchors, the remaining layers Φ_l $(l = 2, \ldots, L_{\text{dec}})$ process the input query $\boldsymbol{S}_{l-1}$ as

$$\boldsymbol{S}_l = \Phi_l(\boldsymbol{S}_{l-1}, \boldsymbol{B}_{\text{out}}) = \text{MLP}\Big(\text{Norm}\big(\text{SA}(\text{DCA}(\boldsymbol{S}_{l-1}, \boldsymbol{B}_{\text{out}}, \boldsymbol{C}'))\big)\Big). \tag{10}$$

The final output $\boldsymbol{S}_{\text{dec}}$ represents the locally refined semantics that have fully absorbed the visual information from the spectrogram based on the precise signal localization.

To extract position features from the intra-frame proposals and model the spatial dependencies between disjoint signal bursts, we append L_{sa} layers of global SA blocks to $\boldsymbol{S}_{\text{dec}}$. Specifically, we first generate high-dimensional spatial position embeddings $\boldsymbol{P}$ based on the refined bounding boxes $\boldsymbol{B}_{\text{out}}$ to emphatically model the position features. For each proposal with normalized center coordinates (c_x, c_y), the encoding is computed independently for each coordinate component $u \in \{c_x, c_y\}$ across varying frequency bands, denoted as

$$\text{PE}(u, \omega) = \begin{cases} \sin(u \cdot 2\pi \cdot \Omega_v), & \text{if } \omega = 2v, \\ \cos(u \cdot 2\pi \cdot \Omega_v), & \text{if } \omega = 2v + 1, \end{cases} \tag{11}$$

where $\Omega_v = 10000^{\frac{-2v}{d/2}}$ represents the frequency scale. The final position embedding $\boldsymbol{p}_m$ for each bounding box $\boldsymbol{b}_m$ is obtained by

$$\boldsymbol{p}_m = \text{cat}(\text{PE}(c_x), \text{PE}(c_y)), m = 1, 2, ..., M, \tag{12}$$

where $\text{cat}(\cdot)$ denotes the concatenation operation. By injecting $\boldsymbol{P}$ into the semantic features, the subsequent SA blocks aggregate positional correlations among disjoint signal bursts, identifying whether they belong to the same hopping sequence. Let $\Psi_i(\cdot) = \text{MLP}(\text{Norm}(\text{SA}(\cdot)))$ denote the operation of the i-th SA block. The final high-level semantic representation $\boldsymbol{S}_{\text{out}}$ is obtained by

$$\boldsymbol{S}_{\text{out}} = \Psi_{L_{\text{sa}}}\Big(\Psi_{L_{\text{sa}}-1}\Big(\ldots \Psi_1(\boldsymbol{S}_{\text{dec}}, \boldsymbol{P}), \boldsymbol{P}\Big) \ldots, \boldsymbol{P}\Big). \tag{13}$$

To facilitate effective model convergence during training, we employ an MLP-based classifier followed by a Softmax activation layer to project the high-level semantics $\boldsymbol{S}_{\text{out}}$ into class probability outputs $\hat{\boldsymbol{C}}$, which can be denoted as

$$\hat{\boldsymbol{C}} = \text{Softmax}\big(\text{MLP}(\boldsymbol{S}_{\text{out}})\big), \tag{14}$$

where $\hat{\boldsymbol{C}} = \{\hat{\boldsymbol{c}}_m\}_{m=1}^{M}$ with $\hat{\boldsymbol{c}}_m \in \mathbb{R}^{N_k+1}$ representing the predicted probability distribution for the m-th signal semantic, N_k denotes the number of known drone protocols, and the extra dimension of $\hat{\boldsymbol{c}}_m$ corresponds to the probability of noise or interference signal.

4.3 Semantic optimization

Aiming to extract robust and distinguishable semantics $\boldsymbol{s}_{\text{out}}$ for different drone types, we propose a two-step optimization pipeline. Specifically, we first perform bipartite matching [43] to establish an optimal one-to-one assignment between the predicted candidates and the ground truth. Based on this assignment, we then apply various loss functions to jointly optimize the semantic space and regress the bounding boxes.

4.3.1 *Bipartite matching*

Let $\mathcal{Y} = \{\boldsymbol{y}_k\}_{k=1}^{K}$ denote the ground truth set of K signal bursts in a spectrogram, where $\boldsymbol{y}_k = (y_k, \boldsymbol{b}_k)$ with y_k being the class label and $\boldsymbol{b}_k$ being the bounding box of the k-th signal burst. The network predicts $\hat{\mathcal{Y}} = \{\hat{\boldsymbol{y}}_m\}_{m=1}^{M}$, where $\hat{\boldsymbol{y}}_m = (\hat{\boldsymbol{c}}_m, \hat{\boldsymbol{b}}_m, \boldsymbol{s}_{\text{out}}^{(m)})$ with $\hat{\boldsymbol{c}}_m$ being the class probability distribution, $\hat{\boldsymbol{b}}_m$ being the predicted box, and $\boldsymbol{s}_{\text{out}}^{(m)}$ being the output semantic. Since the M predicted proposals consist of both drone signal bursts and other noise or interference, the first step is to establish a one-to-one assignment that uniquely pairs each ground-truth instance $\boldsymbol{y}_k$ with a predicted candidate $\hat{\boldsymbol{y}}_m$. Specifically, we search for an assignment σ that minimizes the matching cost using the Hungarian algorithm. The optimal assignment $\hat{\sigma}$ is computed as

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{k=1}^{K} \mathcal{L}_{\text{match}}(\boldsymbol{y}_k, \hat{\boldsymbol{y}}_{\sigma(k)}), \tag{15}$$

$$\mathcal{L}_{\text{match}}(\boldsymbol{y}_k, \hat{\boldsymbol{y}}_{\sigma(k)}) = \gamma_{\text{cls}} \cdot \mathcal{L}_{\text{cls}}(y_k, \hat{\boldsymbol{c}}_{\sigma(k)}) + \gamma_{\text{L1}} ||\boldsymbol{b}_k - \hat{\boldsymbol{b}}_{\sigma(k)}||_1 + \gamma_{\text{giou}} \cdot \mathcal{L}_{\text{giou}}(\boldsymbol{b}_k, \hat{\boldsymbol{b}}_{\sigma(k)}), \tag{16}$$

where $\mathcal{L}_{\text{cls}}(\cdot)$ refers to the focal loss [44], $||\cdot||_1$ represents the operation of calculating the L1-norm, and $\mathcal{L}_{\text{giou}}(\cdot)$ refers to the generalized intersection over union (IoU) loss, with γ_{cls}, γ_{L1}, and γ_{giou} being the respective weighting coefficients. Based on the optimal assignment $\hat{\sigma}$, the predictions corresponding to the indices $\{\hat{\sigma}(k)\}_{k=1}^{K}$ are identified as positive predictions with ground truth $\mathcal{Y}$. Conversely, the remaining predictions that are not matched to any ground truth are labelled as negative predictions and are treated as noise or interference class during training.

4.3.2 *Loss function*

With the optimal assignment $\hat{\sigma}$ determined, the correspondence between the predicted signal semantics and the ground truth categories is established. Let $\boldsymbol{\theta}$ denote the learnable parameters of the network. The training objective is to optimize $\boldsymbol{\theta}$ such that the predicted semantics $\boldsymbol{S}_{\text{out}}$ form a discriminative feature space. Inspired by [9, 11], we introduce the center loss, the cluster loss, and the classification loss to jointly optimize the semantic space.

Firstly, the center loss $\mathcal{L}_{\text{cen}}$ minimizes the intra-class variance by penalizing the distance between the output semantic $\boldsymbol{s}_{\text{out}}^{(m)}$ and its corresponding class center $\boldsymbol{\mu}_{y_m}$, formulated as

$$\mathcal{L}_{\text{cen}}(\boldsymbol{\theta}; \boldsymbol{S}_{\text{out}}) = \frac{1}{M}\sum_{m=1}^{M} ||\boldsymbol{s}_{\text{out}}^{(m)} - \boldsymbol{\mu}_{y_m}||_2, \tag{17}$$

where y_m is the ground truth (GT) class label of the m-th prediction, $\boldsymbol{\mu}_{y_m}$ is the learnable center for class y_m, and $||\cdot||_2$ represents the operation of calculating the L2-norm.

Secondly, the cluster loss $\mathcal{L}_{\text{clu}}$ ensures inter-class separability by introducing a distance margin ϑ to penalize $\boldsymbol{s}_{\text{out}}^{(m)}$ if it falls close to any incorrect class center $\boldsymbol{\mu}_k$, formulated as

$$\mathcal{L}_{\text{clu}}(\boldsymbol{\theta}; \boldsymbol{S}_{\text{out}}) = \frac{1}{M}\sum_{m=1}^{M}\sum_{y\in\mathcal{K}, y\neq y_m} \max\{0, \vartheta - ||\boldsymbol{s}_{\text{out}}^{(m)} - \boldsymbol{\mu}_y||_2\}, \tag{18}$$

where $\mathcal{K}$ denotes the set of all known class labels, and ϑ is a preset margin threshold.

Thirdly, since the predicted probability distribution $\hat{\boldsymbol{C}}$ consists of both positive predictions and negative predictions, the classification loss $\mathcal{L}_{\text{cls}}$ is defined as the focal loss, which is a modified version of the standard cross-entropy loss that down-weights easy negative predictions and focuses more on positive predictions, denoted as

$$\mathcal{L}_{\text{cls}}(\boldsymbol{\theta}; \hat{\boldsymbol{C}}) = -\alpha(1-p_m)^{\gamma_f}\log(p_m), \tag{19}$$

where p_m is the predicted probability of the GT class label y_m in $\hat{\boldsymbol{c}}_m$, α is a balancing parameter, and γ_f is a focusing parameter that down-weights easy predictions.

While the combination of $\mathcal{L}_{\text{cen}}$, $\mathcal{L}_{\text{clu}}$, and $\mathcal{L}_{\text{cls}}$ optimizes the global distribution of the semantic space, these loss functions treat each proposal as an independent individual target, overlooking the physical correlation inherent in the signal data. In reality, FHSS signals manifest as multiple discrete bursts within a single spectrogram frame. Since these bursts originate from the same transmitter, they share identical modulation textures and should possess mutually consistent semantics. To capture this physical property, we introduce the intra-frame consistency loss $\mathcal{L}_{\text{intra}}$. For a specific class c appearing in the current spectrogram $\boldsymbol{X}$, let $\mathcal{S}_c$ be the set of predicted foreground semantics belonging to this class. $\mathcal{L}_{\text{intra}}$ can be denoted as

$$\mathcal{L}_{\text{intra}} = \sum_{c\in\mathcal{C}_{\text{frame}}}\left(\frac{1}{M_c}\sum_{\boldsymbol{s}_c\in\mathcal{S}_c}||\boldsymbol{s}_c - \boldsymbol{\mu}_c'||_2\right), \tag{20}$$

where $\mathcal{C}_{\text{frame}}$ is the set of classes in the current spectrogram frame $\boldsymbol{X}$, M_c is the number of predicted semantics in $\mathcal{S}_c$, and $\boldsymbol{\mu}_c'$ is the mean of all semantics in $\mathcal{S}_c$.

For the localization of each FCS hopping burst, we also employ the standard L1 loss and generalized IoU (GIoU) loss to regress the signal boundary of the positive prediction. Furthermore, without explicit constraints, negative predictions are inherently prone to converging toward degenerate solutions, where the predicted bounding boxes expand to cover nearly the entire spectrogram. To alleviate this localization drift, we introduce an additional regularization term that constrains negative predictions to remain close to their initialized energy-based bounding boxes $\boldsymbol{B}_0$, which are generated in stage I. The localization loss $\mathcal{L}_{\text{loc}}$ is finally denoted as

$$\mathcal{L}_{\text{loc}} = \sum_{k=1}^{K}\left[\lambda_{\text{L1}}||\boldsymbol{b}_k - \hat{\boldsymbol{b}}_{\hat{\sigma}(k)}||_1 + \lambda_{\text{giou}}\mathcal{L}_{\text{giou}}(\boldsymbol{b}_k, \hat{\boldsymbol{b}}_{\hat{\sigma}(k)})\right]$$

$$+\sum_{m\notin\hat{\sigma}}\left[\lambda_{\text{L1_neg}}||\boldsymbol{b}_m^{(0)}-\hat{\boldsymbol{b}}_m||_1+\lambda_{\text{giou_neg}}\mathcal{L}_{\text{giou}}(\boldsymbol{b}_m^{(0)},\hat{\boldsymbol{b}}_m)\right], \tag{21}$$

where $m\notin\hat{\sigma}$ denotes the indices of the unmatched negative predictions, λ_{L1}, λ_{giou}, $\lambda_{\text{L1_neg}}$, and $\lambda_{\text{giou_neg}}$ are weighting coefficients. This constraint ensures the detector maintains the physical priors derived from stage I for all objects.

Finally, we optimize the network parameters $\boldsymbol{\theta}$ using the stochastic gradient-based method to minimize the total loss, defined as

$$\mathcal{L}_{\text{total}}=\lambda_{\text{cen}}\mathcal{L}_{\text{cen}}+\lambda_{\text{clu}}\mathcal{L}_{\text{clu}}+\lambda_{\text{cls}}\mathcal{L}_{\text{cls}}+\lambda_{\text{intra}}\mathcal{L}_{\text{intra}}+\mathcal{L}_{\text{loc}}, \tag{22}$$

where λ_{cen}, λ_{clu}, λ_{cls}, and λ_{intra} are the weighting coefficients.

4.4 Semantic classification

After optimizing the network parameters $\boldsymbol{\theta}$, the semantics of known categories are gathered in the semantic space $\mathbb{R}^d$. The semantic domain for known class c can be denoted as

$$\mathcal{R}_c=\{\boldsymbol{s}\in\mathbb{R}^d\mid d_{\mathcal{M}}(\boldsymbol{s},\boldsymbol{\mu}_c)<\Theta_c^*\}, \tag{23}$$

where $\boldsymbol{\mu}_c$ is the semantic center of class c, $d_{\mathcal{M}}(\cdot,\cdot)$ calculates the Mahalanobis distance, and Θ_c^* is the adaptive boundary threshold for class c. Driven by the optimization of $\mathcal{L}_{\text{cen}}$, the semantics $\mathcal{S}_c$ for each known class c tend to approximate a Gaussian distribution [9]. This statistical property enables the construction of class-specific decision boundaries Θ_c^* based on the closed-set semantics. Specifically, let $\{\boldsymbol{s}_c^{(n)}\}_{n=1}^{N_c}$ denote the closed-set semantic representations of class c used for boundary estimation. Define the distance multiset of class c as

$$\mathcal{D}_c=\left\{d_{\mathcal{M}}\left(\boldsymbol{s}_c^{(n)},\boldsymbol{\mu}_c\right)\right\}_{n=1}^{N_c}, \tag{24}$$

where N_c is the total number of semantics in class c. By sorting the distance set $\mathcal{D}_c$ in descending order and applying the 3σ principle, Θ_c^* is identified as the first distance value that falls within the statistical 3σ threshold, denoted as

$$\Theta_c^*=\delta_c^{\hat{n}},\quad \hat{n}=\min\left\{n_c\in\{1,\ldots,N_c\}\mid\delta_c^{n_c}\leqslant 3\sigma_c\right\}, \tag{25}$$

where δ_c^n is the n-th largest distance in $\mathcal{D}_c$, and σ_c is the standard deviation of $\mathcal{D}_c$.

With the adaptive boundaries $\{\Theta_c^*\}_{c=1}^{N_k+1}$ established for all known classes, each test semantic $\boldsymbol{s}_{\text{test}}$ is assigned to its nearest class c if its Mahalanobis distance to $\boldsymbol{\mu}_c$ falls within the corresponding boundary Θ_c^*. Otherwise, if the semantic representation lies outside the decision boundaries of all known classes, it is rejected as unknown and stored in the candidate buffer $\mathcal{B}_u$ for further analysis. This decision rule can be formally expressed as

$$\hat{y}=\begin{cases}\hat{c}, & \text{if } d_{\mathcal{M}}(\boldsymbol{s}_{\text{test}},\boldsymbol{\mu}_{\hat{c}})<\Theta_{\hat{c}}^*,\\ c_{\text{u}}, & \text{otherwise,}\end{cases} \tag{26}$$

where $\hat{c}=\arg\min_{c\in\mathcal{K}}d_{\mathcal{M}}(\boldsymbol{s}_{\text{test}},\boldsymbol{\mu}_c)$ represents the nearest known class, and c_{u} denotes the designated label for the unknown class.

Once the outliers are rejected into the candidate buffer $\mathcal{B}_u$, we aim to discover their latent categories. However, the semantic output $\boldsymbol{s}_{\text{out}}$ has been deeply optimized to fit the distribution of known classes, potentially suppressing generalized details necessary for distinguishing novel patterns. To compensate for this, we implement a semantic augmentation strategy. The augmented semantic $\boldsymbol{s}_{\text{aug}}$ is constructed by concatenating the refined semantics $\boldsymbol{s}_{\text{out}}$ with the preserved raw texture features $\boldsymbol{s}_t$ generated in Subsection 4.1, denoted as

$$\boldsymbol{s}_{\text{aug}}=\text{cat}(\boldsymbol{s}_{\text{out}},\boldsymbol{s}_t). \tag{27}$$

Let $\mathcal{S}_{\text{aug}}=\{\boldsymbol{s}_{\text{aug}}^{(v)}\}_{v=1}^{N_u}$ denote the set of augmented semantics corresponding to all rejected semantics in $\mathcal{B}_u$, where N_u is the total number of candidates. We first verify the hypothesis that all semantics in $\mathcal{S}_{\text{aug}}$ belong to a single unknown class. If this hypothesis holds, the spatial distribution of $\mathcal{S}_{\text{aug}}$ should exhibit a degree of compactness comparable to that of known classes. Therefore, the adaptive boundary Θ_U^* derived from $\mathcal{S}_{\text{aug}}$ must satisfy the condition that it falls within the maximum boundary range of known classes, formulated as

$$\Theta_U^*\leqslant\max_{c\in\mathcal{K}}\Theta_c^*. \tag{28}$$

Table 1 Introduction to the experimental dataset: classes, number of samples, and box sizes. Due to different settings of sampling bandwidth, for classes A–J, height: 1 pixel = 0.098 MHz, while for classes M–Q, height: 1 pixel = 0.078 MHz. The sampling segment is unified as 50 ms, so for all classes, width: 1 pixel = 0.041 ms.

Index	Class	Average size of boxes		Number of frames	Number of annotations
		Width (pixel)	Height (pixel)		
A	DJI Mini 3 Pro	14	26	201	1993
B	DJI Mavic 3	13	23	201	1658
C	DJI MATRICE 300	13	24	201	1981
D	DJI MATRICE 30T	13	24	201	1661
E	DJI Mavic Pro	12	13	201	1997
F	DJI Mini 2	13	14	201	2002
G	Futaba T14SG	47	21	201	1398
H	DJI Phantom 3	46	27	200	1391
I	VBar	45	10	283	1364
J	Futaba T16IZ	35	26	300	952
M	DJI Mini 2 SE	13	17	201	1986
N	DJI Mavic 3 Pro	13	29	195	1947
O	DJI Mini 4 Pro	14	57	201	1956
P	DJI Mavic 3 Pro + DJI Mini 2 SE	–	–	100	1996
Q	DJI Mavic 3 Pro + DJI Mini 4 Pro	–	–	100	1978

Conversely, if the condition above is violated, it suggests the presence of multiple unknown categories. In this case, we determine the optimal number of unknown clusters u^* via an iterative evaluation strategy. Specifically, we vary the potential cluster number u from 2 to a preset maximum $N_{\max}$. For each u, we apply the K-means [45] algorithm to partition $\mathcal{S}_{\text{aug}}$ and calculate the Silhouette Coefficient to evaluate the clustering quality. The optimal number of categories u^* is determined by identifying the peak value of the Silhouette Coefficient curve. Finally, with u^* fixed, we execute K-means again to obtain the final partition of the unknown categories.

5 Experiment

5.1 Experimental setup

5.1.1 *Dataset and implementation settings*

We construct the experimental dataset based on DroneRFa [46] and DroneRFb-DIR [47], selecting 13 representative drone models to ensure diversity. The RF signals are processed via STFT with a window size of $Q = 2048$ applied to 50 ms signal segments. Subsequently, the generated spectrograms are downsampled to a resolution of 1220×1024 to balance feature resolution and computational efficiency. Detailed statistics of the selected classes and annotations are provided in Table 1.

To comprehensively evaluate the model's performance, we design six distinct experimental scenarios with different partitions of known and unknown categories, as outlined in Table 2. For the known classes in scenarios I through VI, samples are randomly split into training and testing sets at a ratio of 8:2, while samples belonging to the unknown classes are excluded during training and appear exclusively in the testing set.

The proposed model is implemented in PyTorch and trained on a single NVIDIA GeForce RTX 3090 GPU. The model is trained for up to 300 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4}. The learning rate is multiplied by 0.95 every 10 epochs using the StepLR scheduler.

For all experimental scenarios, we set the number of deformable decoder layers to $L_{\text{dec}} = 3$, the number of global SA blocks to $L_{\text{sa}} = 1$, and the semantic dimension to $d = 128$. The batch size is 16, and the loss weights are set to $\lambda_{\text{cen}} = 0.1$, $\lambda_{\text{clu}} = 0.2$, $\lambda_{\text{cls}} = 2$, and $\lambda_{\text{intra}} = 0.05$. These weights were selected to balance the magnitudes of the four loss terms and ensure stable convergence of the network.

5.1.2 *Evaluation metrics*

Following standard protocols for OSR and object detection, we employ a comprehensive set of metrics to evaluate both the detection and classification performance.

Table 2 Scenarios with different dataset splits.

Scenario	Setting	Class indexes
I	Known (12)	A, B, C, D, E, F, G, H, I, J, M, N
	Unknown (1)	O
II	Known (11)	A, B, C, D, E, F, G, H, I, J, M
	Unknown (2)	N, O
III	Known (10)	A, B, C, D, E, F, G, H, I, M
	Unknown (3)	J, N, O
IV	Known (9)	A, B, C, D, E, F, G, H, M
	Unknown (4)	I, J, N, O
V	Known (10)	A, B, C, D, E, F, G, H, I, J
	Unknown (3)	M, N, O
VI	Known (10)	A, B, C, D, E, F, G, H, I, J
	Unknown (2)	P, Q

To evaluate the detection performance, we first introduce two commonly used indicators, namely precision and recall. Specifically, precision reflects the proportion of detected signal bursts that are actual FCS signals, while recall measures the proportion of actual positive instances that the model correctly identifies. To align our evaluation criteria with existing methods, we establish a unified rule for assigning confidence scores when computing these metrics. For targets accepted as known classes according to (26), we utilize the classification confidence score from $\hat{\boldsymbol{c}}$. Conversely, for targets rejected as unknown, which intrinsically lack a defined classification confidence, we uniformly assign a confidence score of 1 while evaluating precision and recall.

To evaluate the model's ability to handle unknown classes, we adopt four specific indicators defined in [9], i.e., true known rate (TKR), true unknown rate (TUR, also known as unknown recall [37]), known precision (KP), and unknown precision (UP). These metrics collectively assess the model's performance in both closed-set classification and OSR. Let TK, TU, FK, and FU denote the number of testing samples recognized as true known, true unknown, false known, and false unknown, respectively. Then, TKR, TUR, KP, and UP can be calculated as

$$\mathrm{TKR}=\frac{\mathrm{TK}}{\mathrm{TK}+\mathrm{FU}},\quad \mathrm{TUR}=\frac{\mathrm{TU}}{\mathrm{TU}+\mathrm{FK}},\quad \mathrm{KP}=\frac{\mathrm{TK}}{\mathrm{TK}+\mathrm{FK}},\quad \mathrm{UP}=\frac{\mathrm{CU}}{\mathrm{TU}+\mathrm{FU}},\tag{29}$$

where correctly unknown (CU) denotes the number of unknown FCS bursts that are correctly matched to their corresponding latent clusters. Thus, KP measures the precision of accepting samples into the known set, whereas the class-wise precision and recall reported below evaluate exact known-class classification. The matching between the predicted clusters and unknown drone classes is obtained via a simple bipartite assignment strategy that maximizes the total number of correctly aligned samples.

5.1.3 *Comparison methods*

To comprehensively evaluate the performance of the proposed framework, we compare it against two categories of methods: state-of-the-art open-world object detection and domain-specific closed-set drone detectors.

• **Open-set detector.** We select OW-DETR [38], a leading OWOD model in computer vision, as the primary open-set competitor. It employs an attention-driven pseudo-labeling mechanism to discover unknown instances from the background. However, a fundamental domain assumption mismatch renders it ineffective for RF signals. OW-DETR is predicated on the assumption that the training images contain latent, unannotated instances, which are exploited to model the conceptual representation of unknown classes. In contrast, the RF training spectrograms employed in this study strictly comprise annotated known signals and pure background noise. Consequently, the pseudo-labeling mechanism forces the model to fit background noise as potential objects, leading to optimization collapse. As a result, the model fails to extract meaningful features, exhibiting negligible response and near-zero recall to both known and unknown FCSs.

• **Closed-set detector.** We compare with SMNet [15], a robust closed-set drone detection network designed for spectrograms. Since SMNet focuses exclusively on classifying known categories, it serves as a performance benchmark for known classes but is expected to misclassify unknown FCSs. To provide a direct comparison of the feature extraction capability in a closed-set setting, we also evaluate a variant of our proposed method using standard SoftMax activation, denoted as Ours (SoftMax).

Table 3 Detailed comparisons of precision and recall in scenario II. For each class, the best precision and recall values are shown in bold, including ties. a) The unknown class detection for SMNet is implemented as a binary classification, where all known drone signals are assigned a unified label to focus solely on detection.

Class (index)		Ours		Ours (SoftMax)		SMNet [15]	
		Precision	Recall	Precision	Recall	Precision	Recall
Known classes	DJI Mini 3 Pro (A)	0.8547	0.8635	**0.8604**	**0.8763**	0.6875	0.7549
	DJI Mavic 3 (B)	0.5578	0.5620	**0.8743**	**0.8892**	0.5962	0.6528
	DJI MATRICE 300 (C)	0.6351	0.7788	**0.8445**	**0.8565**	0.6470	0.7173
	DJI MATRICE 30T (D)	0.7111	0.7930	**0.9318**	**0.9554**	0.6743	0.7495
	DJI Mavic Pro (E)	0.7616	0.7613	**0.8739**	**0.8822**	0.5073	0.5679
	DJI Mini 2 (F)	0.8179	0.8276	**0.9125**	**0.9471**	0.5693	0.6416
	Futaba T14SG (G)	0.7942	0.8035	**0.8717**	0.9083	0.8584	**0.9780**
	DJI Phantom 3 (H)	0.8944	0.9137	**0.9308**	**0.9569**	0.8135	0.9199
	VBar (I)	**0.6674**	0.7055	0.6642	0.7055	0.6030	**0.7163**
	Futaba T16IZ (J)	**0.9901**	0.9939	0.9782	**0.9970**	0.5864	0.6357
	DJI Mini 2 SE (M)	**0.8617**	**0.9078**	0.8431	**0.9078**	0.6934	0.9028
Unknown classes	DJI Mavic 3 Pro (N)	**0.7316**	**0.7807**	–	–	0.6045[a)]	0.6382[a)]
	DJI Mini 4 Pro (O)	**0.4341**	**0.5286**	–	–	0.0000[a)]	0.0000[a)]
Metrics	TKR	0.8044		–		–	
	TUR	0.9134		–		–	
	KP	0.9338		–		–	
	UP	0.4986		–		–	

5.2 Performance evaluation

5.2.1 *Open-set detection and classification*

Selecting scenario II as a typical scenario, we present a detailed comparison of precision and recall for each known and unknown class in Table 3. The unknown class detection for SMNet is implemented as a binary classification, where all known drone signals are assigned a unified label to focus solely on detection. The results demonstrate that our proposed method achieves superior performance in both known and unknown classes.

Specifically, regarding the detection of unknown classes, Table 1 reveals a critical physical discrepancy: the bandwidth of the DJI Mini 4 Pro significantly exceeds that of any drone in the known set, while the DJI Mavic 3 Pro shares similar texture features and geometric scales with the known classes. Consequently, even when SMNet is adapted for binary detection, it only achieves moderate performance on the DJI Mavic 3 Pro while failing completely to respond to the DJI Mini 4 Pro. This empirical evidence validates our earlier analysis regarding the inherent limitations of conventional detectors: by relying on learnable filters that overfit specific training data, these models are rendered incapable of activating when encountering unknown FCSs with vastly different modulation schemes.

In contrast, our method overcomes this limitation by implementing the energy-based proposal generation module. By leveraging the continuous energy structure of FHSS signals to achieve class-agnostic detection, we ensure high recall for signal patterns regardless of their appearance in the training set. As a result, our method successfully captures the challenging DJI Mini 4 Pro, attaining 43.41% precision and 52.86% recall.

Furthermore, the proposed two-stage framework demonstrates superior feature extraction and construction capabilities compared to existing detectors. For known classes, Ours (SoftMax) outperforms SMNet significantly. For instance, on the DJI MATRICE 30T (D), Ours (SoftMax) baseline achieves 0.9318 in precision and 0.9554 in recall, outperforming SMNet by +25.7% and +20.6%, respectively. Meanwhile, we observe that transitioning from the closed-set SoftMax classifier to the open-set distance-based classifier introduces a necessary trade-off. The recall of the proposed method is generally comparable to or slightly lower than Ours (SoftMax), indicating that outlier samples located far from their class centers in the semantic space are rejected by the adaptive boundaries. However, the semantic classification algorithm guarantees robust rejection of unknowns, achieving a TKR of 0.8044 and a TUR of 0.9134, thereby establishing an effective balance between accepting known samples and rejecting unknown outliers.

Finally, although the UP is relatively low at 0.4986 due to the inevitable interference of false alarms from background noise and rejected known samples, the semantic augmentation strategy significantly enhances the distinguishability of unknown FCSs. The subsequent clustering process effectively separates valid unknown signals from noise, yielding a recall of 0.7807 and precision of 0.7316 for the DJI Mavic 3 Pro. Notably, even though SMNet

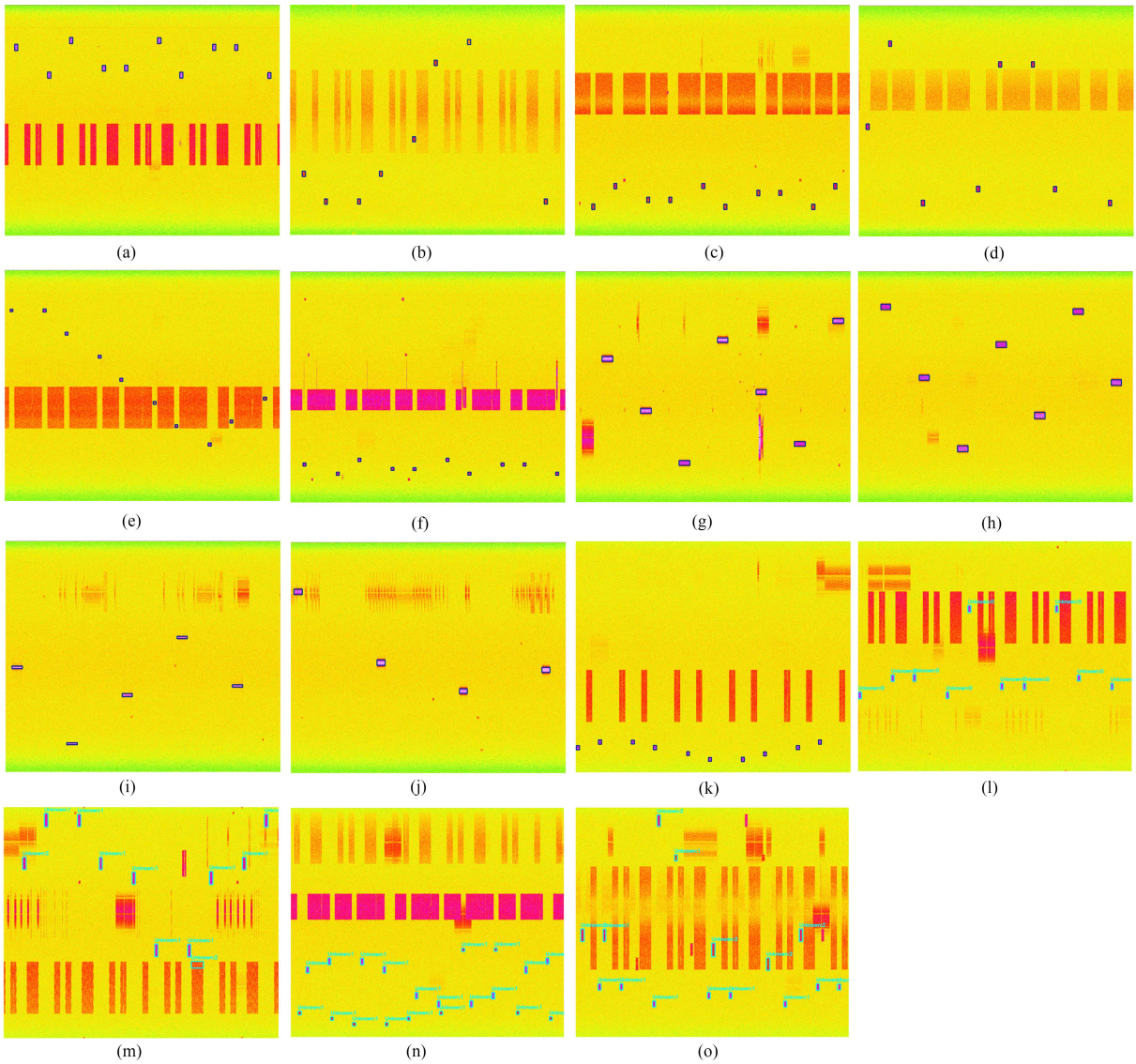

Figure 3 Visualization of detection and classification results under scenarios II and VI. (a) DJI Mini 3 Pro; (b) DJI Mavic 3; (c) DJI MATRICE 300; (d) DJI MATRICE 30T; (e) DJI Mavic Pro; (f) DJI Mini 2; (g) Futaba T14SG; (h) DJI Phantom 3; (i) VBar; (j) Futaba T16IZ; (k) DJI Mini 2 SE; (l) DJI Mavic 3 Pro; (m) DJI Mini 4 Pro; (n) DJI Mavic 3 Pro + DJI Mini 2 SE; (o) DJI Mavic 3 Pro + DJI Mini 4 Pro.

treats this as a simpler binary task, our multi-class open-set framework still outperforms it by +12.7% in precision and +14.2% in recall. While the DJI Mini 4 Pro is also effectively detected, its classification performance is lower than that of the Mavic 3 Pro, which can be attributed to a less compact cluster distribution in the semantic space.

Figure 3 provides a qualitative visualization of the detection and classification results under scenarios II and VI. Panels (a)–(m) present representative single-drone results from scenario II, in which classes A–J and M are known, whereas the DJI Mavic 3 Pro (N) and DJI Mini 4 Pro (O) are unknown. The detected bounding boxes closely follow the localized FCS hopping bursts across drone models with diverse time-frequency patterns, illustrating the effectiveness of the proposed method in localizing known and previously unseen signals. Panels (n) and (o) show the more challenging scenario VI, where two unknown drones operate simultaneously. Despite the coexistence and interleaving of hopping bursts from different emitters, the proposed method detects the concurrent signals and assigns their bursts to the corresponding unknown groups, demonstrating its capability to handle multi-drone open-set scenarios.

5.2.2 *Comparison with classical CFAR detectors*

To compare the proposed class-agnostic proposal generator with classical adaptive-threshold detectors, we replace the morphology-based proposal stage with cell-averaging CFAR (CA-CFAR) and ordered-statistic CFAR (OS-

Table 4 Comparison with classical CFAR proposal generators in scenario II. The highest value for each metric, excluding proposals per frame, is shown in bold.

Method	Proposal recall			Proposals per frame	FPS	TKR	TUR	KP	UP
	IoU = 0.1	IoU = 0.3	IoU = 0.5						
Proposed	0.9375	0.9285	0.8179	14.97	**324.45**	0.8044	**0.9134**	**0.9338**	**0.4986**
CA-CFAR	0.9689	0.9367	**0.8269**	80.12	8.68	0.8439	0.5731	0.7839	0.4517
OS-CFAR	**0.9701**	**0.9421**	0.6941	122.35	5.05	**0.8889**	0.6861	0.6539	0.4066

Table 5 Model complexity and network-forward latency. The lower parameter count and shorter inference latency are shown in bold.

Method	Native input	Parameters	Mean latency (P95) (ms)
SMNet	$3 \times 512 \times 512$	18.77M	**10.01 (10.40)**
Proposed	$1 \times 1024 \times 1220$	**13.45**M	13.59 (14.03)

CFAR) [48, 49], while keeping the downstream semantic network and open-set decision procedure unchanged. The comparison is conducted under scenario II. In addition to the end-to-end open-set metrics, we report proposal recall at three IoU thresholds, the average number of proposals per spectrogram, and proposal-generation throughput. The throughput is measured for the proposal stage only and therefore isolates the computational cost of the three front ends.

As shown in Table 4, CA-CFAR and OS-CFAR obtain slightly higher proposal recall at the loose IoU thresholds of 0.1 and 0.3, but generate 5.35 and 8.17 times as many proposals as the proposed method, respectively. At the stricter IoU threshold of 0.5, the proposed method remains comparable to CA-CFAR and clearly outperforms OS-CFAR. The CFAR variants also achieve higher TKR, indicating that more known targets are retained. However, this gain is accompanied by substantial decreases in TUR, KP, and UP, suggesting that the much larger proposal sets introduce more ambiguous or spurious candidates into the downstream open-set stage. In addition, the proposed proposal generator is 37.4 times faster than CA-CFAR and 64.2 times faster than OS-CFAR. Overall, the proposed morphology-based front end provides a better balance among localization recall, proposal quantity, downstream open-set performance, and computational efficiency.

5.2.3 *Computational efficiency*

To complement the accuracy comparison, Table 5 compares the model complexity and network-forward latency of the proposed method and SMNet. Each model is evaluated at its native input resolution.

As shown in Table 5, the proposed model uses 13.45M parameters, 28.3% fewer than SMNet, while exhibiting a moderately higher mean latency of 13.59 ms compared with 10.01 ms for SMNet. This moderate latency increase should be considered in light of the substantially larger spatial input resolution of the proposed method, which contains approximately 4.77 times as many spatial locations, together with the additional deformable decoding and global self-attention operations used for feature refinement. Despite these differences, the proposed model achieves a mean network-forward latency of 13.59 ms, corresponding to approximately 73.6 spectrograms per second, indicating that the increased input resolution does not result in prohibitive inference latency.

5.2.4 *Robustness and generalization*

• Multi-scenario generalization.

In Figure 4, we evaluate the generalizability of the proposed method across the six scenarios defined in Table 2, analyzing the impact of increasing open-set complexity on model performance.

First, by examining scenarios I–IV, where the number of unknown categories progressively increases from 1 to 4, we observe that the primary detection indicators remain highly stable despite the expanding open-set. Specifically, TKR consistently exceeds 0.8, and TUR maintains a robust level of approximately 0.9. This stability indicates that the proposed adaptive boundary mechanism effectively distinguishes between known and unknown domains regardless of the diversity of unknown signal patterns.

Second, we compare scenarios V and VI to analyze the impact of signal concurrency. While scenario V contains single-target unknown signals, scenario VI involves multi-target scenarios where multiple unknown drones operate simultaneously. TKR remains unchanged at 0.8398, and KP increases from 0.7939 to 0.9371, showing that accepting known signals remains stable. In contrast, TUR decreases from 0.8574 to 0.8143, and UP decreases from 0.5188 to 0.4245. This result suggests that concurrent unknown bursts mainly challenge unknown rejection and fine-grained

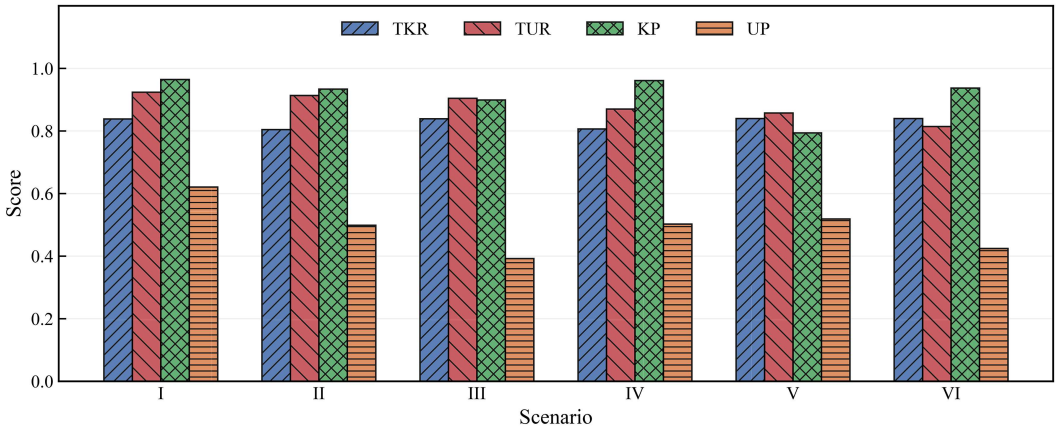

Figure 4 Performance of the proposed method across various scenarios.

Table 6 Robustness under progressively increasing noise levels in scenario II. For each metric, the highest value across noise levels is shown in bold.

Level	σ	Proposal recall IoU = 0.5	TKR	TUR	KP	UP
0	0.0	**0.8179**	**0.8044**	0.9134	0.9338	0.4986
1	0.5	0.8062	0.7544	0.9196	0.9373	**0.5711**
2	1.0	0.8015	0.7572	0.9227	0.9403	0.4849
3	1.5	0.7903	0.7465	0.9250	0.9415	0.5501
4	2.0	0.7723	0.7428	0.9243	**0.9423**	0.5347
5	2.5	0.7535	0.7392	**0.9262**	0.9411	0.5194

grouping rather than known-set acceptance. The remaining gap highlights the difficulty of separating interleaved unknown sources and motivates future improvements in multi-target unsupervised clustering.

• **Noise robustness.**

Inspired by the level-wise robustness evaluation used for FCS monitoring in SMNet [15], we evaluate the proposed framework under progressively increasing noise levels in scenario II without retraining or fine-tuning. Zero-mean additive Gaussian noise is added to the preprocessed spectrograms at standard deviations $\sigma \in \{0.5, 1.0, 1.5, 2.0, 2.5\}$. The original input evaluated within the same robustness pipeline is denoted as Level 0. The adaptive class boundaries estimated from the original training set are kept fixed throughout the test. For each nonzero noise level, the evaluation is repeated with three independently generated noise realizations, and their arithmetic mean is reported. This controlled perturbation is intended as a stability test rather than a specialized denoising scheme.

Table 6 shows gradual rather than abrupt degradation over the tested perturbation range. Proposal recall decreases from 0.8179 to 0.7535, indicating that proposal generation degrades progressively as the injected noise increases. At Level 1, proposal recall decreases by only 1.17 percentage points, whereas TKR decreases by 5.00 percentage points. This suggests that known-sample recognition is more sensitive to the imposed perturbation than proposal localization, with some known proposals crossing the fixed class boundaries as the learned semantics shift. TUR remains stable or increases slightly because this displacement produces more conservative rejection; it should not be interpreted as improved unknown recognition. Similarly, the stable KP should be considered together with the decreasing TKR: known-set acceptance remains precise, but a larger fraction of known proposals is rejected before classification. UP varies non-monotonically from 0.4849 to 0.5711 across the nonzero levels, suggesting greater variability in the downstream unknown-class discovery stage. Overall, the framework exhibits gradual degradation rather than noise immunity.

To determine whether these observations depend on a precisely tuned rejection threshold, we further scale every class-specific boundary according to

$$\Theta_c^{(\rho)} = \rho \Theta_c^*, \qquad \rho \in \{0.50, 0.55, \ldots, 1.50\}, \tag{30}$$

where the boundary scale ρ is introduced only for this sensitivity analysis and $\rho = 1$ recovers the training-calibrated boundary in (26). The experiment uses the original input and one fixed Level-3 input realization ($\sigma = 1.5$) under scenario II. For each input, the checkpoint, class prototypes, covariance matrices, proposals, semantic representations, and proposal-to-class Mahalanobis distances are kept fixed. Consequently, only the known/unknown decisions

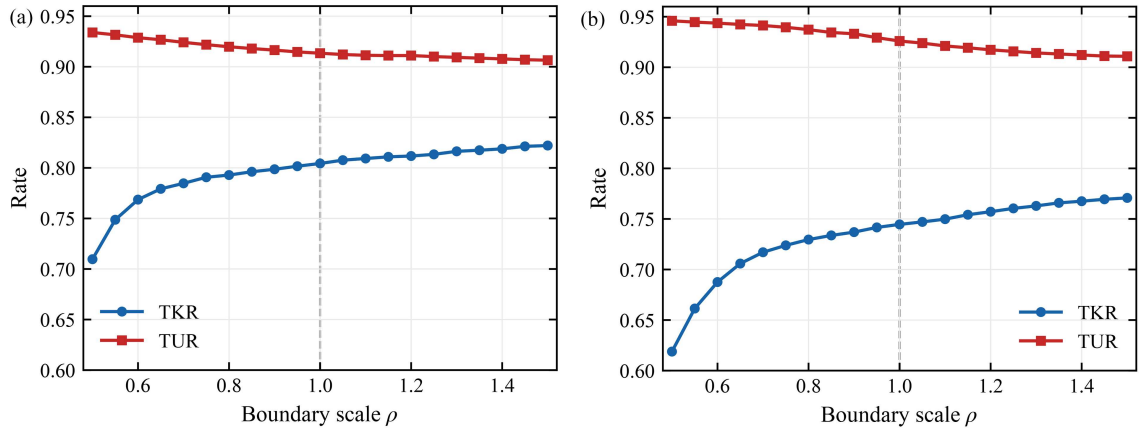

Figure 5 Threshold sensitivity in scenario II: (a) original input; (b) one fixed Level-3 input realization ($\sigma = 1.5$). Each class-specific boundary is scaled by ρ, while all upstream outputs and distance statistics remain fixed. The dashed vertical line denotes the default operating point $\rho = 1$.

Table 7 Unknown-class clustering performance of K-means and GMM. For each scenario, the better result is shown in bold.

Scenario	K-means + silhouette	GMM + silhouette
II	**0.4986**	0.3565
III	0.3923	**0.4868**
IV	**0.5024**	0.2455
V	**0.5188**	0.4714
VI	**0.4245**	0.3779

are recomputed as ρ changes; no retraining, fine-tuning, or repeated network inference is involved. The default value is not reselected using unknown test samples.

As shown in Figure 5, increasing ρ monotonically raises TKR and lowers TUR, reflecting the expected trade-off between accepting known samples and rejecting unknown samples. Around the default boundary, changing ρ from 0.9 to 1.1 changes TKR/TUR by only $+1.06/-0.51$ percentage points for the original input and $+1.28/-1.20$ percentage points for the Level-3 input. The response is therefore smooth and locally stable rather than tied to an exact threshold. A more pronounced degradation occurs only when the boundary is excessively tightened. Relative to $\rho = 1$, setting $\rho = 0.5$ reduces TKR by 9.47 and 12.58 percentage points for the original and Level-3 inputs, respectively, while improving TUR by only 2.05 and 2.00 percentage points. Hence, an overly strict boundary mainly increases false rejection of known samples without a comparable gain in unknown rejection. Conversely, the moderate TUR decrease up to $\rho = 1.5$ shows that unknown proposals are not readily absorbed into the known domains under boundary expansion. The Level-3 input follows the same monotonic pattern as the original input, although its TKR remains lower, suggesting that the basic threshold-response behavior is preserved under the tested noise perturbation.

5.2.5 *K-means versus GMM for unknown clustering*

To examine whether the additional covariance flexibility of a Gaussian mixture model (GMM) benefits unknown-class clustering, we replace K-means with a GMM with diagonal covariance matrices. All upstream detections, rejected candidates, augmented semantics, and candidate numbers of clusters are kept identical. For both methods, the number of unknown clusters is selected by maximizing the Silhouette Coefficient over the same candidate range. The resulting UP scores for the two clustering methods are reported in Table 7.

Although GMM improves UP in scenario III, K-means performs better in the other four evaluated scenarios. This result indicates that the additional covariance flexibility of GMM does not consistently translate into better unknown-class clustering performance. The center, cluster, and intra-frame consistency losses explicitly regularize the learned semantics toward compact and well-separated groups, making centroid-based Euclidean clustering well aligned with the learned semantic structure. Moreover, the rejected candidate buffer contains false proposals and mis-rejected known samples, and estimating component-specific covariance parameters can increase GMM's sensitivity to such noisy samples. We therefore retain K-means for its stronger overall performance and lower model complexity.

Table 8 Detailed comparisons of performance under different model configurations. Bold values indicate the best results across model configurations, including ties.

Unknown class (index)	Original		$\boldsymbol{s}_0 = \boldsymbol{s}_t$		$\boldsymbol{s}_0 = \boldsymbol{s}_p$		w/o $\mathcal{L}_{\text{intra}}$		$\boldsymbol{s}_{\text{aug}} = \boldsymbol{s}_{\text{out}}$	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
DJI Mavic 3 Pro (N)	**0.7316**	**0.7807**	0.2066	0.3066	0.3208	0.3857	0.4767	0.5193	**0.7316**	**0.7807**
DJI Mini 4 Pro (O)	**0.4341**	**0.5286**	0.2661	0.5424	0.3182	0.4642	0.1964	0.2413	0.4020	0.4949
TKR	**0.8044**		0.7637		0.7924		0.7981		**0.8044**	
TUR	0.9134		0.9154		0.8978		**0.9229**		0.9134	
KP	0.9338		0.9248		0.9168		**0.9448**		0.9338	
UP	**0.4986**		0.3187		0.3504		0.3009		0.4857	

5.2.6 *Ablation study*

To validate the effectiveness of the key components in the proposed method, a comprehensive ablation study is conducted under the setting of scenario II. The quantitative results are summarized in Table 8.

Across all configurations, KP remains within a relatively narrow range from 0.9168 to 0.9448, whereas UP varies from 0.3009 to 0.4986. This contrast indicates that the ablated components primarily affect the fine-grained clustering of unknown signals rather than the precision of known-set acceptance.

First, the importance of the proposed semantic initialization is verified. By comparing the original method with the single-stream variants, namely $\boldsymbol{s}_0 = \boldsymbol{s}_t$ and $\boldsymbol{s}_0 = \boldsymbol{s}_p$, it is evident that the fusion of both features yields the optimal performance. Notably, the $\boldsymbol{s}_p$ demonstrate a more significant impact on performance than the texture features $\boldsymbol{s}_t$. Relying solely on texture features results in a sharp decline in UP from 0.4986 to 0.3187, and the precision for the unknown DJI Mavic 3 Pro drops drastically. This suggests that, while internal signal texture provides basic discrimination, the geometric and location priors provided by the position stream are critical for distinguishing adjacent signal bursts.

Then, the necessity of the proposed intra-frame consistency loss $\mathcal{L}_{\text{intra}}$ is investigated. As shown in Table 8, removing this physical constraint leads to a substantial degradation in unknown classification, with UP dropping significantly to 0.3009. While TKR and TUR remain relatively stable, the ability to correctly cluster unknown signals is severely impaired. This result validates that $\mathcal{L}_{\text{intra}}$ plays a pivotal role in aggregating disjoint signal bursts from the same drone instance.

Finally, the contribution of the semantic augmentation strategy is evaluated. By comparing the proposed method with the non-augmented setting where $\boldsymbol{s}_{\text{aug}} = \boldsymbol{s}_{\text{out}}$, we observe that fusing the raw texture features $\boldsymbol{s}_t$ enhances the clustering performance for novel patterns. Specifically, for the DJI Mini 4 Pro, the augmentation strategy improves the recall from 0.4949 to 0.5286 and precision from 0.4020 to 0.4341. This indicates that $\boldsymbol{s}_{\text{out}}$, being deeply optimized for known classes, tends to filter out features that are irrelevant to the closed set but actually informative for unknown drones. Semantic augmentation effectively compensates for this information loss, providing a more generalized representation for open-set clustering.

6 Conclusion

In this paper, we proposed a coarse-to-fine framework for open-set multi-drone FCS detection and classification. The framework integrates class-agnostic energy-based proposal generation, texture-position semantic initialization, Transformer-based feature refinement, and multi-objective semantic optimization with intra-frame consistency constraints. Experiments on real-world drone RF data demonstrate that the proposed framework can effectively localize FCS bursts, recognize known drone classes, and reject and cluster unknown classes across different open-set scenarios. The results further show that the proposed design provides a practical balance among detection performance, open-set recognition capability, and computational efficiency.

Several limitations remain to be addressed. First, the current framework is static and cannot yet promote newly identified unknown classes into the known set without model updating; future work will therefore investigate incremental learning strategies that incorporate new classes while mitigating catastrophic forgetting. Second, the present system relies solely on RF observations, and its reliability may degrade under strong electromagnetic interference or severe signal collisions that are not extensively represented in the current scenarios. Future studies will explore collision-aware processing in more densely occupied time-frequency environments and multi-modal fusion with complementary sensing modalities, such as FMCW radar micro-Doppler signatures, to improve the validation of unknown targets and robustness to interference. Finally, although the current implementation demonstrates

favorable inference efficiency, a comprehensive device-level evaluation of latency, memory usage, and energy consumption remains necessary for edge deployment. Lightweight and more advanced feature extractors will therefore be systematically investigated to further improve the accuracy–efficiency trade-off for practical low-altitude anti-drone applications.

Acknowledgements This work was supported by National Key Research and Development Program of China (Grant No. 2025YFF0514600), National Natural Science Foundation of China (Grant Nos. 62531019, 62401502), Pioneer and Leading Goose R&D Program of Zhejiang (Grant No. 2026LDC01013(JT)), and Fundamental Research Funds for the Central Universities (Grant No. 226-2026-00060).

References

1 Zhang T, Chen C, Xu Y, et al. Joint task scheduling and multi-UAV deployment for aerial computing in emergency communication networks. Sci China Inf Sci, 2023, 66: 192303
2 Wang C, Li X, Gu Y, et al. An adaptive 3D reconstruction method for asymmetric dual-angle multispectral stereo imaging system on UAV platform. Sci China Inf Sci, 2024, 67: 182305
3 Zhou X, Xiong J, Zhao H, et al. Joint UAV trajectory and communication design with heterogeneous multi-agent reinforcement learning. Sci China Inf Sci, 2024, 67: 132302
4 Zhang Y, Shan H, Hou F, et al. Uncrewed aerial vehicle communications for low-altitude economy: leveraging integrated satellite-terrestrial networks. IEEE Wireless Commun, 2026, 33: 145–154
5 Wang Y, Ran D. China's low-altitude economy needs to grow with more caution. Nature, 2025, 638: 891
6 Huang X. The small-drone revolution is coming—scientists need to ensure it will be safe. Nature, 2025, 637: 29–30
7 Shi X, Yang C, Xie W, et al. Anti-drone system with multiple surveillance technologies: architecture, implementation, and challenges. IEEE Commun Mag, 2018, 56: 68–74
8 Bisio I, Garibotto C, Lavagetto F, et al. Blind detection: advanced techniques for Wi-Fi-based drone surveillance. IEEE Trans Veh Technol, 2019, 68: 938–946
9 Dong Y, Jiang X, Zhou H, et al. SR2CNN: zero-shot learning for signal recognition. IEEE Trans Signal Proces, 2021, 69: 2316–2329
10 Bendale A, Boult T E. Towards open set deep networks. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, 2016. 1563–1572
11 Yu N, Wu J, Zhou C, et al. Open set learning for RF-based drone recognition via signal semantics. IEEE Trans Inf Foren Sec, 2024, 19: 9894–9909
12 Wang M, Lin T, Wang L, et al. Uncertainty-inspired open set learning for retinal anomaly identification. Nat Commun, 2023, 14: 6757
13 Basak S, Rajendran S, Pollin S, et al. Combined RF-based drone detection and classification. IEEE Trans Cogn Commun Netw, 2022, 8: 111–120
14 Zhao R, Li T, Li Y, et al. Anchor-free multi-UAV detection and classification using spectrogram. IEEE Internet Things J, 2024, 11: 5259–5272
15 Yu N, Wu J, Zhou C, et al. SMNet: multi-drone detection and classification via monitoring flight control signals on spectrograms. IEEE Trans Cogn Commun Netw, 2025, 11: 3245–3259
16 Xie Y, Jiang P, Gu Y, et al. Dual-source detection and identification system based on UAV radio frequency signal. IEEE Trans Instrum Meas, 2021, 70: 1–15
17 Al-Sa'd M F, Al-Ali A, Mohamed A, et al. RF-based drone detection and identification using deep learning approaches: An initiative towards a large open source drone database. Future Gener Comp Sy, 2019, 100: 86–97
18 Xu C, Chen B, Liu Y, et al. RF fingerprint measurement for detecting multiple amateur drones based on STFT and feature reduction. In: Proceedings of IEEE Integrated Communications, Navigation and Surveillance Conference (ICNS), Herndon, 2020
19 Kılıç R, Kumbasar N, Oral E A, et al. Drone classification using RF signal based spectral features. Eng Sci Tech an Int J, 2022, 28: 101028
20 Medaiyese O O, Ezuma M, Lauf A P, et al. Wavelet transform analytics for RF-based UAV detection and identification system using machine learning. Pervasive Mob Comput, 2022, 82: 101569
21 Xu C, He F, Chen B, et al. Adaptive RF fingerprint decomposition in micro UAV detection based on machine learning. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Toronto, 2021. 7968–7972
22 Shoufan A, Al-Angari H M, Sheikh M F A, et al. Drone pilot identification by classifying radio-control signals. IEEE Trans Inf Foren Sec, 2018, 13: 2439–2447
23 Wang Q, Wang L, Yu L, et al. An ID-based robust identification approach toward multitype noncooperative drones. IEEE Sens J, 2023, 23: 10179–10192
24 Rajendran S, Meert W, Lenders V, et al. Unsupervised wireless spectrum anomaly detection with interpretable features. IEEE Trans Cogn Commun Netw, 2019, 5: 637–647
25 Xue C, Li T, Li Y, et al. Radio-frequency identification for drones with nonstandard waveforms using deep learning. IEEE Trans Instrum Meas, 2023, 72: 1–13
26 Medaiyese O O, Ezuma M, Lauf A P, et al. Hierarchical learning framework for UAV detection and identification. IEEE J Radio Freq Id, 2022, 6: 176–188
27 Zhang H, Li T, Li Y, et al. RF-based drone classification under complex electromagnetic environments using deep learning. IEEE Sens J, 2023, 23: 6099–6108
28 Wang Z, Cao Z, Xie J, et al. RF-based drone detection enhancement via a generalized denoising and interference-removal framework. IEEE Signal Proc Let, 2024, 31: 929–933
29 Li J, Li J, Lv L, et al. Drone remote identification based on Zadoff-Chu sequences and time-frequency images. IEEE Trans Cogn Commun Netw, 2026, 12: 1634–1648
30 Chen X, Yin Z, Liu F, et al. Multiple UAV flight control signals extraction based on optimized local maximum with FPGA implementation. IEEE Trans Instrum Meas, 2026, 75: 1–15
31 Soltani N, Reus-Muns G, Salehi B, et al. RF fingerprinting unmanned aerial vehicles with non-standard transmitter waveforms. IEEE Trans Veh Technol, 2020, 69: 15518–15531
32 Li T, Wen Z, Long Y, et al. The importance of expert knowledge for automatic modulation open set recognition. IEEE Trans Pattern Anal, 2023, 45: 13730–13748
33 Han H, Li W, Feng Z, et al. Proceed from known to unknown: jamming pattern recognition under open-set setting. IEEE Wirel Commun Le, 2022, 11: 693–697
34 Mao Y, Liang H, Zhang Z, et al. A few-shot open-set UAV recognition method based on multidomain prototype learning. IEEE Internet Things J, 2025, 12: 45913–45916
35 Guo R, Yan S, Wang J. LOSDR: lightweight open-set drone recognition via RF semantics alignment. IEEE Internet Things J, 2026, 13: 6928–6945
36 Gao N, Zeng T, Chen B, et al. Multi-domain supervised contrastive learning for UAV radio-frequency open-set recognition. IEEE J Sel Area Comm, 2026, 44: 4083–4098
37 Joseph K J, Khan S, Khan F S, et al. Towards open world object detection. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, 2021. 5826–5836

38 Gupta A, Narayan S, Joseph K J, et al. OW-DETR: open-world detection transformer. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, 2022. 9225–9234
39 Zhang S, Ni Y, Du J, et al. Open-world objectness modeling unifies novel object detection. In: Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, 2025. 30332–30342
40 He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016. 770–778
41 Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017. 2117–2125
42 Zhu X, Su W, Lu L, et al. Deformable DETR: deformable transformers for end-to-end object detection. In: Proceedings of International Conference on Learning Representations (ICLR), 2021
43 Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers. In: Proceedings of the European Conference on Computer Vision (ECCV), 2020. 213–229
44 Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection. In: Proceedings of IEEE/CVF International Conference on Computer Vision (ICCV), Honolulu, 2017. 2980–2988
45 Nie F, Li Z, Wang R, et al. An effective and efficient algorithm for k-means clustering with new formulation. IEEE Trans Knowl Data En, 2023, 35: 3433–3443
46 Yu N, Mao S, Zhou C, et al. DroneRFa: a large-scale dataset of drone radio frequency signals for detecting low-altitude drones. J Electron Inf Technol, 2024, 46: 1147–1156
47 Ren J, Yu N, Mao S, et al. DroneRFb-DIR: an RF signal dataset for non-cooperative drone individual identification. J Electron Inf Technol, 2025, 47: 573–581
48 Finn H M, Johnson R S. Adaptive detection mode with threshold control as a function of spatially sampled clutter-level estimates. RCA Rev, 1968, 29: 414–464
49 Rohling H. Radar CFAR thresholding in clutter and multiple target situations. IEEE Trans Aero Elec Sys, 1983, AES-19: 608–621