

Article

DroneTFS-SemNet: A Semantic Learning Framework for Multi-Task Recognition of Drones

Yantian Shen ¹, Bohang Xu ¹, Na Wang ¹, Liang Ma ², Yang Yang ¹ , Yunxia Liu ³ , Hongjun Wang ¹ and Hua Li ^{1,*}

¹ School of Information Science and Engineering, Shandong University, Qingdao 266237, China; ytshen@mail.sdu.edu.cn (Y.S.); bohangu@mail.sdu.edu.cn (B.X.); wangna2023@sdu.edu.cn (N.W.); yyang@sdu.edu.cn (Y.Y.); hjw@sdu.edu.cn (H.W.)

² School of Computer Science and Technology, Shandong University, Qingdao 266237, China; liangma@sdu.edu.cn

³ Center for Optics Research and Engineering (CORE), Shandong University, Qingdao 266237, China; eyxliu@sdu.edu.cn

* Correspondence: sdlh@sdu.edu.cn

Abstract

The widespread deployment of unmanned aerial vehicles has raised growing privacy and security concerns, creating an urgent need for effective drone detection and identification. To address these challenges, this paper proposes a semantic learning framework termed Drone Time–Frequency–Semantic Network (DroneTFS-SemNet) for drone recognition. The main contributions are as follows. A Temporal Enhancement module is designed to improve model adaptability and robustness through radio frequency (RF) signal noise simulation, feature pre-extraction, learnable frequency-band partitioning, and an attention mechanism. A Cross-Channel Recalibration Module is introduced to promote cross-channel interaction of RF features via dilated convolutions and Channel Shuffle, thereby enhancing the model’s learning capability. Third, Feature Fusion Attention is employed to weight, align, and fuse RF features from different levels, thereby improving the model’s ability to represent differences among drone categories and operation modes. A semantic construction and decision module is further developed to extract high-level semantic features and accomplish drone detection, classification, and recognition. Experimental results on the DroneRF dataset show that DroneTFS-SemNet achieves a drone classification accuracy of 99.73% and an operation mode recognition accuracy of 99.65%, outperforming multiple comparison methods. These results demonstrate that the proposed model has strong recognition capability on this public dataset and under controlled experimental conditions, providing an effective method for RF-based drone recognition research.

Keywords: drone anomaly detection; drone recognition; RF signal recognition; semantic classification



Academic Editor: Jiann-Shiun Yuan

Received: 15 April 2026

Revised: 13 May 2026

Accepted: 15 May 2026

Published: 19 May 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the rapid advancement and widespread deployment of drone technology in applications such as precision spraying [1], path planning [2], and security patrol [3], drones have brought substantial convenience while also giving rise to a series of public safety and personal privacy concerns, including unauthorized airspace intrusion and covert surveillance, as illustrated in Figure 1. Therefore, achieving accurate and efficient drone identification and regulation has become an urgent and critical challenge.

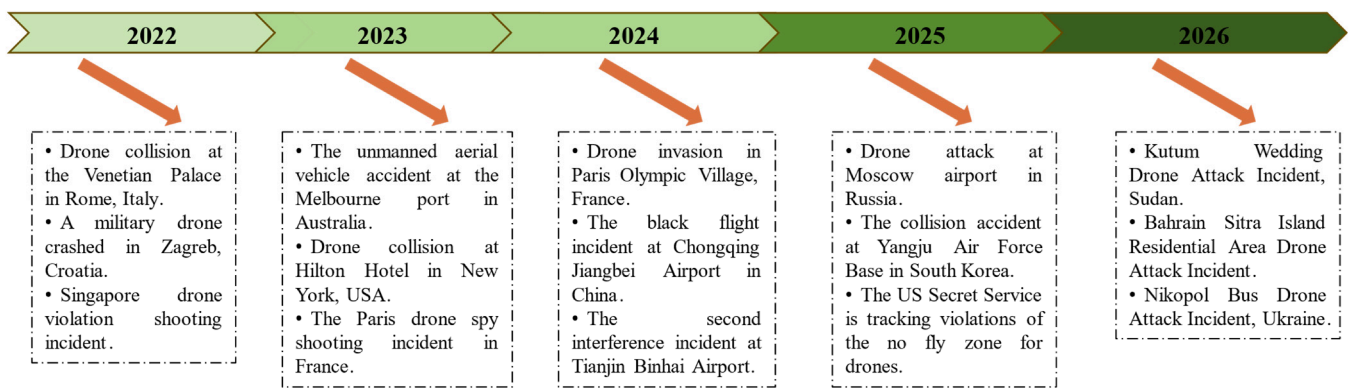


Figure 1. Drone Public Safety Risk Events (2022–2026).

At present, mainstream drone identification approaches can be broadly categorized into four types, namely radar-based, acoustic-based, vision-based, and RF-based methods. Table 1 summarizes their respective advantages, limitations, and applicable operating ranges.

Table 1. Comparison of mainstream methods for drone recognition.

Method	Distance	Advantage	Disadvantage
Radar	~8 km	Long—distance, high precision.	Weak ability to recognize target types.
Radio Frequency	~3 km	Relatively long distance, passive detection.	Prone to interference from electromagnetic signals.
Acoustics	~250 m	Short distance, low cost.	Prone to interference from noise.
Vision	~2 km	Intuitive information.	Greatly restricted by the environment

Radar-based drone identification mainly relies on the analysis of micro-Doppler effects. Zhong et al. [4] employed MobileNetV3, a lightweight deep neural network better suited to radar embedded hardware, to identify drone targets from three-dimensional data segments cropped from range-Doppler maps. Hong et al. [5] processed radar echoes to obtain micro-Doppler feature images and then classified drones through pattern recognition on these images. Zhang et al. [6] introduced a collaborative attention mechanism in convolutional neural networks. This mechanism fine-tunes the network's focus by dynamically adjusting the weights of different feature channels and spatial regions, allowing it to concentrate on the regions with the most information. Jiang et al. [7] learned directly from radar echo data and enabled the model to focus on the micro-Doppler signatures generated by the target, thereby enhancing drone identification performance.

Acoustic-based drone identification mainly depends on the acoustic signals generated during flight, including turbulent noise produced by the high-speed interaction between propellers and air, as well as electromagnetic and mechanical vibration noise from the motors. Ren et al. [8] used Mel-frequency cepstral coefficients and Gammatone cepstral coefficients as input features, and effectively addressed the challenge of medium- and long-range drone acoustic recognition through an innovative feature fusion and enhancement mechanism. Jafar et al. [9] proposed a novel frequency feature extraction method based on the Prony algorithm, which enables efficient drone detection and classification. Zhang et al. [10] developed a Bi-Long Short-Term Memory network with an OpenMax model for open-set classification. Liu et al. [11] addressed the difficulty caused by spectral overlap between drone sounds and background noise under low-SNR conditions by effectively combining lightweight design with an attention mechanism to improve feature extraction capability.

Vision-based drone identification is fundamentally achieved through the extraction and analysis of visual features from images or videos. Zou et al. [12] developed a multi-view feature fusion network, termed MVFF, to exploit complementary information across different viewpoints, thereby improving the accuracy of small-target detection in complex environments. Ulzhalgas et al. [13] identified moving objects via background subtraction and combined this process with a CNN for classification. Han et al. [14] integrated a spatial attention module into the network backbone to emphasize small-target information, enabling more accurate localization of small drones and further improving detection performance. Zhou et al. [15] proposed an improved YOLOv13 framework that further enhanced drone detection performance.

RF-based drone identification is primarily realized by analyzing the radio signal characteristics transmitted between drones and their controllers or ground stations. Yu et al. [16] adopted a frame-by-frame search strategy and combined it with a smoothed pseudo-Wigner-Ville distribution algorithm to implement a high-precision drone individual recognition method. Pankaj et al. [17] processed RF signals through wavelet-domain feature extraction and a three-stage wavelet decomposition scheme, achieving strong performance in drone model recognition. Rahman et al. [18] combined a CNN with a bidirectional long short-term memory network for accurate drone classification. Delleji et al. [19] designed a real-time micro-drone detection and identification system based on RF signal features and a customized multilayer perceptron model.

Existing RF-based drone recognition methods still have limitations in feature representation and multi-task collaborative processing. In particular, for drone detection, classification, and operation mode recognition, the modeling of fine-grained RF features still needs to be further improved. To address these issues, this paper proposes a semantic learning framework termed the Drone Time-Frequency-Semantic Network (DroneTFS-SemNet), which models drone RF signals from the temporal, frequency, and semantic perspectives, thereby completing drone detection, classification, and operation mode recognition tasks. The main contributions of this paper are as follows.

- First, a Temporal Enhancement (TE) module is designed to increase the diversity of training samples through noise perturbation, amplitude scaling, and random temporal discarding. A feature pre-extraction module is then used to obtain low-level representations of RF signals. On this basis, the Adaptive Attention Frequency-band Enhancement module (AAFE) partitions the spectrum using learnable frequency-band split points and enhances information from key frequency bands and important channels through an attention mechanism.
- Secondly, a Cross-Channel Recalibration Module (CCRM) is introduced to extract features through two layers of dilated convolution, while Channel Shuffle is incorporated to promote information flow across RF feature channels, strengthen cross-channel interaction, and refine channel importance, thereby enhancing the learning capability and expressive power of the model.
- Thirdly, Feature Fusion Attention (FFA) is employed to weight and align each feature map, effectively integrating RF feature information from different levels and enabling DroneTFS-SemNet to more accurately understand and represent complex patterns in the input data.
- Finally, the fused drone RF features are fed into the semantic construction and decision module to accomplish the final tasks of drone detection, classification, and recognition.

2. Materials and Methods

2.1. Definition of Core Tasks for Drone Identification and Supervision

The identification and supervision of unmanned aerial vehicles (UAVs) rely on a system composed of equipment, antennas, databases, etc. Detection, classification, and operation pattern recognition are the core tasks, jointly safeguarding airspace safety [20]. Figure 2 shows that in urban environments, various electromagnetic interferences can occur in factories, schools, and residential areas. When UAVs operate, radio frequency signals will overlap with signals from consumer electronics, communication facilities, and other sources in the airspace, forming a complex electromagnetic environment and increasing the difficulty of detection and identification. The antenna, database, and algorithm modules on the right side of Figure 2 have constructed a technical chain from signal acquisition to data support and function implementation, reflecting the need for drone supervision to rely on multi-device collaboration and algorithm model support to complete accurate identification tasks in complex urban scenes [21].

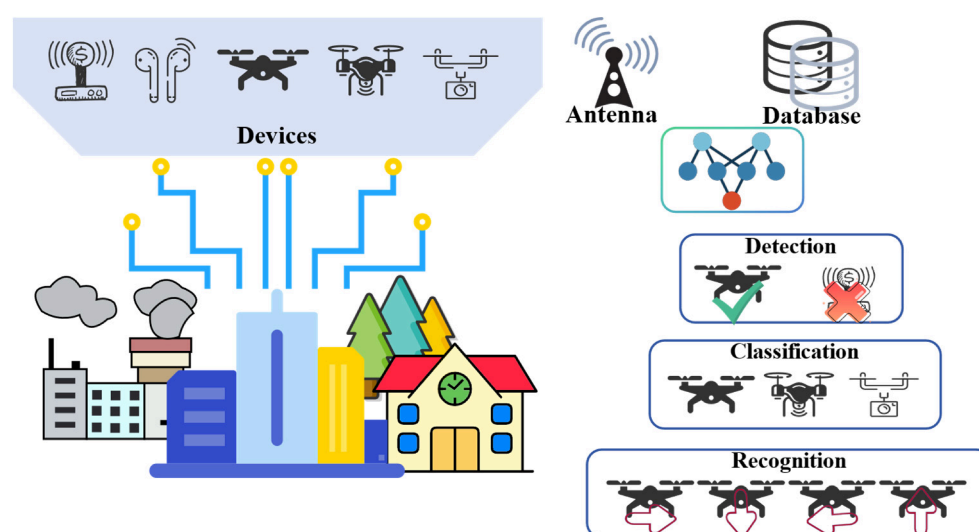


Figure 2. UAV Detection, Classification, and Recognition System in Urban Background.

As the first checkpoint of regulation [22], detection focuses on determining whether there are drones in the target area, laying a solid foundation for subsequent regulatory actions. If detection is missed, classification and recognition cannot be discussed.

The classification task aims to clarify “which type of drone it is”. Based on the signal characteristics of different drones stored in the database, it conducts in-depth analysis of the captured signals and accurately classifies them according to the differences in drone types. This enables regulators to grasp the basic attributes of drones and provides a basis for differentiated regulatory strategies.

Operation mode recognition focuses on the “operational status and intention” of UAVs [23]. By analyzing the signal spectrum and time-domain characteristics, combined with different flight stage feature templates in the database, the current operating mode of the drone can be identified to assist in determining whether it has violated regulations (such as cruising in a no-fly zone or taking off without authorization), which is a key link in precise supervision and intervention of drone behavior.

These three tasks progress layer by layer, detecting the existence of anchors, clarifying attributes through classification, recognizing operational patterns to gain insights into behavior, and jointly building a complete logical loop for drone recognition and supervision to help safeguard airspace safety and order.

2.2. The Inherent Characteristics and Challenges of RF Signals in UAVs

The signals of UAVs exhibit significant time-varying characteristics at different flight stages, which are closely related to factors such as flight stage, flight speed, and environmental interference. The communication between drones and remote controllers usually uses radio frequency signals:

$$s(t) = A(t) \cdot \cos(\omega(t)t + \varphi(t)), \quad (1)$$

where $A(t)$ represents the instantaneous amplitude of the signal, $\omega(t)$ is the instantaneous angular frequency, and $\varphi(t)$ is the instantaneous phase.

Some drones use frequency hopping communication technology, in which the angular frequency changes over time according to a specific hopping pattern to enhance the anti-interference ability of communication. If the frequency hopping pattern is ω_n , it will switch between these different frequency values at different time intervals T_n . The formula is as follows:

$$\omega(t) = \omega_n, (n-1)T_n \leq t < nT_n. \quad (2)$$

Noise is typically modeled as Gaussian white noise $n(t)$, with a probability density function of:

$$p(n) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{n^2}{2\sigma^2}}, \quad (3)$$

where σ^2 is the variance of noise. At this point, the received signal $r(t)$ is:

$$r(t) = s(t) + n(t). \quad (4)$$

The signal from the receiving end not only contains noise, but also a large amount of electromagnetic interference such as WiFi and Bluetooth. The interference signal $i(t)$ comes from various sources, so the received signal further becomes:

$$r(t) = s(t) + n(t) + i(t). \quad (5)$$

Multipath effects are also an important factor affecting the RF time-domain signals of UAVs. The signal propagates to the receiving end through different paths with varying lengths, resulting in different arrival times and phases of the signal at the receiving end. For the existence of N paths, the signal of the k_{th} path can be represented as:

$$s_k(t) = A_k(t) \cdot \cos(\omega(t)(t - \tau_k) + \varphi_k(t)), \quad (6)$$

where τ_k is the propagation delay of the k_{th} path and $\varphi_k(t)$ is the phase of the k_{th} path signal.

The total received signal $r(t)$ is the sum of the signals from each path plus noise and interference:

$$r(t) = \sum_{k=1}^N s_k(t) + n(t) + i(t), \quad (7)$$

The power spectrum $P(f, t)$ of the drone signal describes the energy distribution of the signal in the frequency domain. The dynamic evolution of power spectrum is closely related to the flight mode of UAVs. Under different flight modes, the power spectrum of the signal will exhibit different peaks and fluctuations. For example, in flight and recording modes, due to the presence of video data transmission, the power spectrum will generate more fluctuations in the high-frequency range, while in hover mode, the power spectrum is more stable.

To quantitatively describe the changes in signal power spectrum, Short Time Fourier Transform (STFT) or Discrete Fourier Transform (DFT) is commonly used to analyze time-

frequency characteristics [24]. Figure 3 shows the time–frequency plots of different drone signals on the DroneRF dataset [25] used in this article. For the signal $r(t)$ received by the drone, its STFT is defined as:

$$STFT_r(t, f) = \int_{-\infty}^{\infty} r(\tau)w(\tau - t)e^{-j2\pi f\tau}d\tau, \quad (8)$$

where $w(t)$ is the window function.

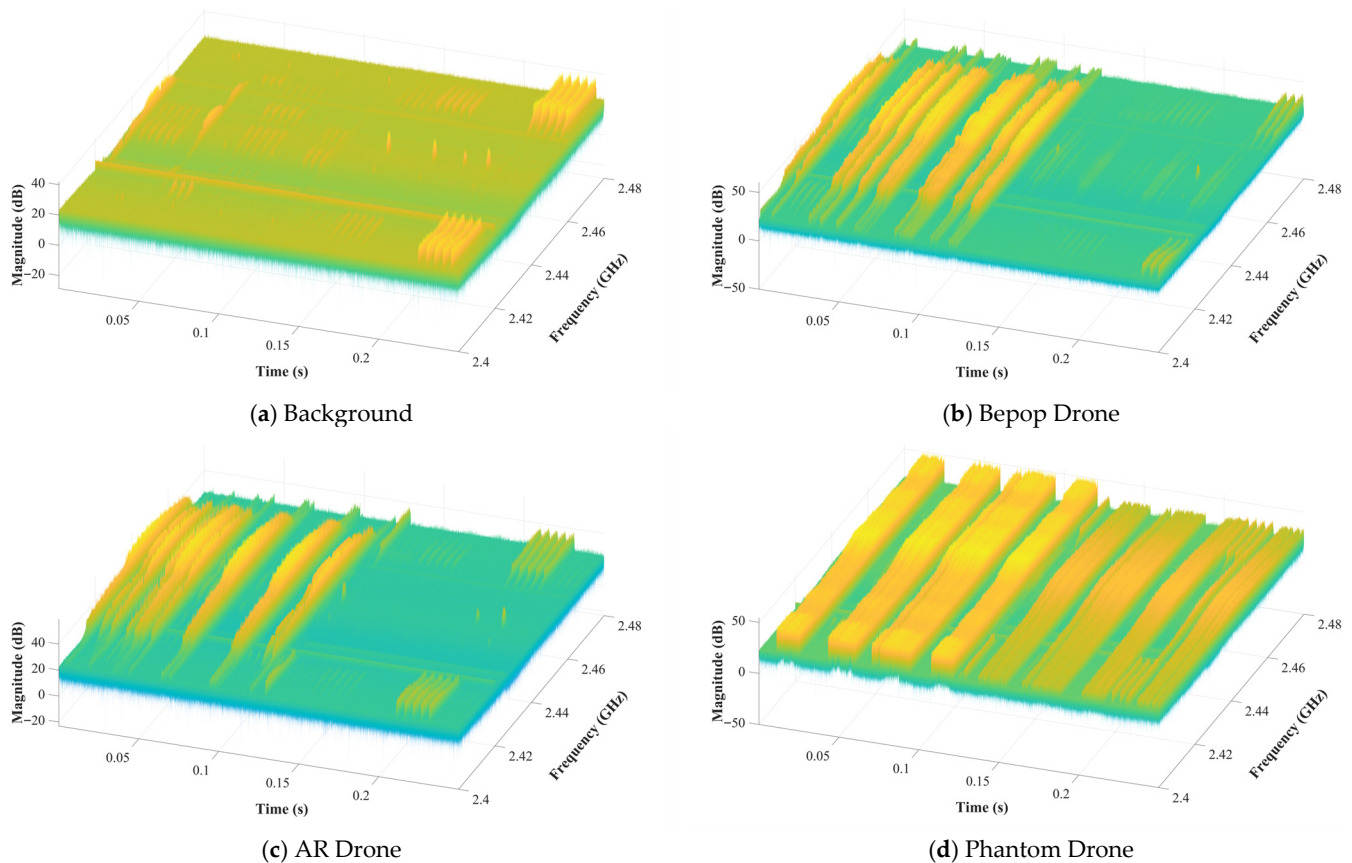


Figure 3. 3D time–frequency diagram of three types of drones and background RF activities in the DroneRF dataset.

Drone signals exhibit time-varying characteristics in different operating modes.

3. DroneTFS-SemNet

The proposed DroneTFS-SemNet is illustrated in Figure 4. First, the feature extraction module employs TE to perform temporal-domain enhancement on drone RF signals. The enhanced signals are then fed into the AAFE module, where preliminary frequency-domain enhancement is achieved through dynamic frequency-band partitioning and a dual-attention mechanism. Subsequently, a network denoted as CCRM is designed as the core of semantic modeling to generate deep semantic feature representations of drone samples and learn information related to drone categories and flight modes. Meanwhile, FFA dynamically integrates feature information from different levels through adaptive weighting. Finally, through the collaboration of the semantic construction and decision modules, the framework accomplishes drone signal detection, classification, and operation mode recognition.

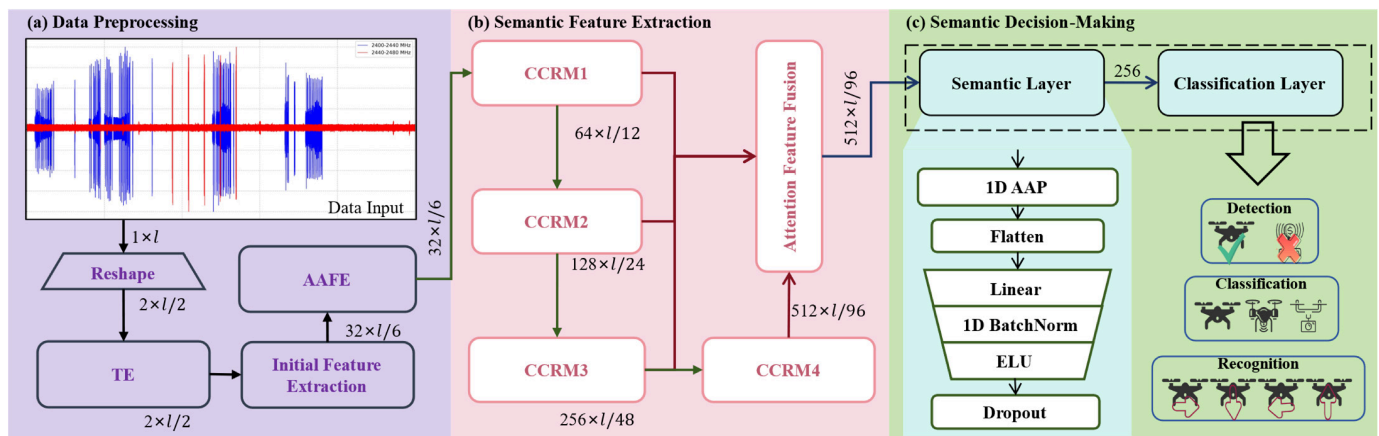


Figure 4. Architecture of the proposed DroneTFS-SemNet for drone recognition, consisting of three major modules: (a) data preprocessing, (b) semantic feature extraction, and (c) semantic decision.

3.1. DroneRF Dataset

DroneRF is an open-source drone RF signal dataset created by Allahham et al. [25] in 2019. The dataset records background RF activities without drones and remote-control communication signals from three consumer drones, including Parrot Bebop, Parrot AR Drone, and DJI Phantom 3. The data were collected by RF receiving equipment, which was mainly used to capture the communication signals between drones and their flight control modules. DroneRF consists of 227 RF activity segments, covering three drone types and background RF activities without drones. Each segment is divided into low-frequency and high-frequency bands, resulting in 454 original CSV files.

DroneRF supports three levels of recognition tasks. The first level is drone presence detection, which aims to determine whether a drone is present in the current RF signal. This task can be divided into two classes, namely background without drones and drone presence. The second level is drone type recognition, which further identifies the specific drone model after a drone is detected. The main classes include Parrot Bebop, Parrot AR Drone, and DJI Phantom 3. The third level is operation mode recognition, which aims to determine the current working state of the drone. Bebop and AR Drone contain four operation modes, namely switched on and connected to the controller, hovering, flying, and flying while recording. DJI Phantom 3 contains only the switched-on and connected-to-controller mode in this dataset.

The sample distribution of DroneRF is not fully balanced. The original data consist of 227 segments, including background segments without drones and segments from the three drone types. Existing information shows that the dataset usually contains 41 background segments, 84 Bebop segments, 81 AR Drone segments, and 21 Phantom 3 segments.

In this paper, each CSV file in the DroneRF dataset is first divided into frames. The time-amplitude plots of the DroneRF dataset are shown in Figure 5. The low-frequency and high-frequency signals are divided into multiple samples using a frame length of L , and each subsample is normalized by Min-Max normalization to eliminate the influence of scale. The normalized low-band sub-sample and high-band sub-sample are then concatenated row-wise to form a single sample with a dimension of $[2L, 1]$ for the same time interval.

Ultimately, the aggregated dataset contains normalized sub-samples of radio frequency signals from four classes. Each sub-sample consists of a concatenated sequence of low-band and high-band signals and can be directly used for drone radio frequency signal detection, classification, and recognition tasks.

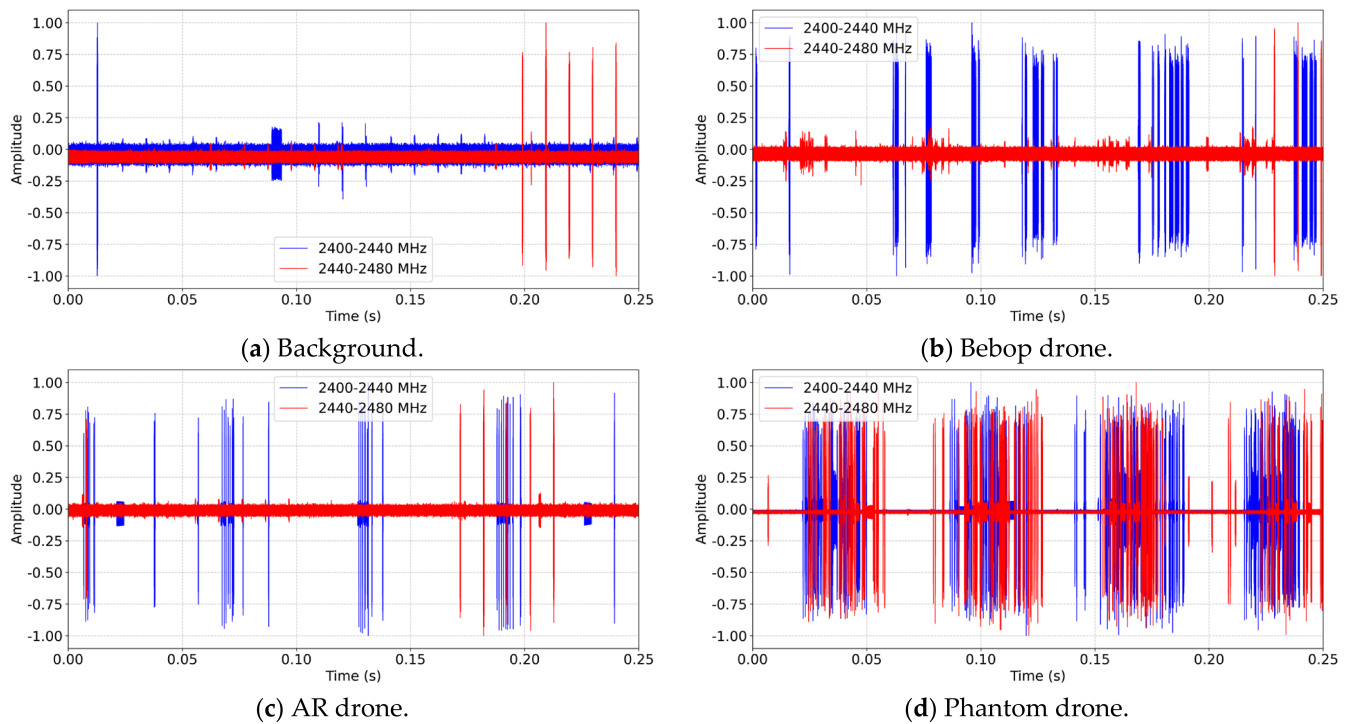


Figure 5. Time-amplitude plots of the three drones in the DroneRF dataset and the background radio frequency activity.

3.2. Temporal Enhancement

To improve the adaptability of DroneTFS-SemNet to input signal perturbations, the TE module enriches the training samples by constructing controlled disturbances. During training, the TE module adds random noise to the input signal and combines it with amplitude scaling and random temporal discarding to generate samples with different perturbation patterns. It should be noted that these perturbations are used to approximately simulate possible signal fluctuations during RF acquisition, but they cannot be fully equivalent to complex electromagnetic interference in real outdoor environments. The enhanced IQ signals are shown in Figure 6.

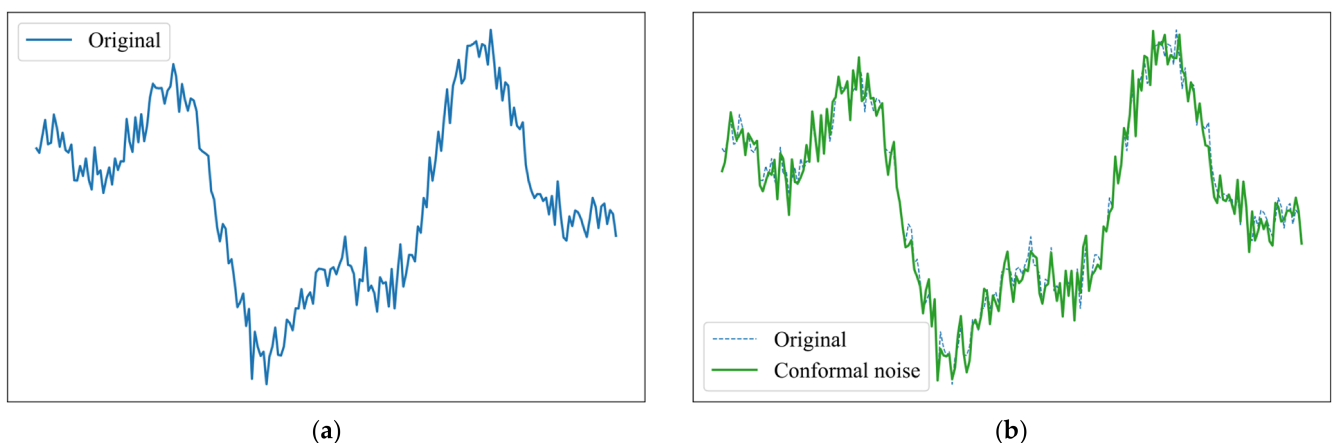


Figure 6. Cont.

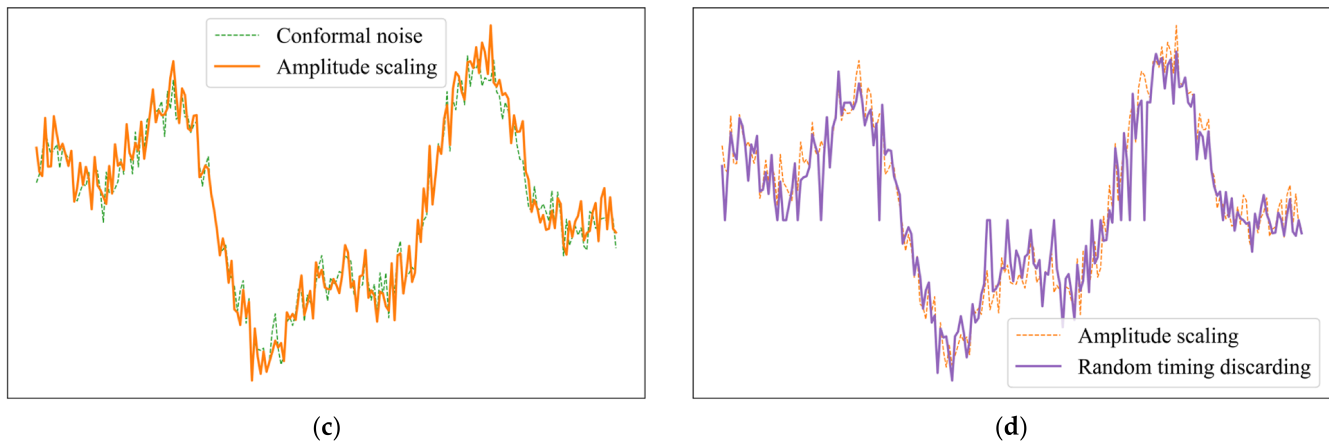


Figure 6. Illustration of the three-step enhancement process of the TE module on a random signal. (a) Original signal. (b) Comparison between the original signal and the signal enhanced with conformal noise. (c) Comparison between conformal-noise enhancement and amplitude-scaling enhancement. (d) Comparison between amplitude-scaling enhancement and random-timing-discarding enhancement.

3.2.1. Conformal Noise

The TE module adds standard Gaussian noise with the same shape as the input data $x \in \mathbb{R}^{c \times l}$. Specifically, a noise matrix $n \in \mathbb{R}^{c \times l}$ is defined, where each element n_{ij} is sampled from the standard normal distribution $\mathcal{N}(0, 1)$, and therefore has zero mean and unit variance. To control the noise intensity, a scaling factor $\alpha = 0.01$ is introduced to adjust the noise magnitude. After adding the noise to the input signal, the resulting noisy signal x_1 can be expressed as

$$x_1 = x + \alpha \cdot n, \quad (9)$$

where the n is sampled from the standard normal distribution, each element n_{ij} follows a distribution with mean 0 and variance 1.

3.2.2. Amplitude Scaling

Random amplitude scaling is then applied to the noisy signal x_1 to simulate random attenuation during signal transmission, thereby helping DroneTFS-SemNet adapt to different signal strengths:

$$x_2 = \beta \cdot x_1, \quad (10)$$

where x_2 denotes the amplitude-scaled data, and $\beta \in (0.99, 1.01)$ is a scaling factor sampled from a uniform distribution.

3.2.3. Random Timing Discarding

By randomly discarding a portion of the signal time steps, the TE module simulates information loss during transmission and further improves model robustness:

$$x_3 = \gamma \cdot x_2, \quad (11)$$

where x_3 denotes the data after temporal discarding, and $\gamma = 0.01$ is a randomly generated tensor with the same shape as the input signal.

The final output x_3 of the TE module is then fed into a series of convolution operations, batch normalization layers, and activation functions to extract the initial features of the drone radio frequency signals.

3.3. Adaptive Attention Frequency-Band Enhancement Module

The AAFE enhances the frequency components and spatiotemporal information in the input data through an adaptive attention mechanism, as shown in Figure 7. For the extracted initial features of drone RF signals, denoted by $I \in \mathbb{R}^{C \times l_t}$, the frequency-band partition points are first computed to determine the position of each frequency band in the spectrum. The learnable frequency-band partition points are initialized as $\varphi = [\varphi_1, \varphi_2, \varphi_3]$, which represent the relative positions used to divide different frequency bands. During training, these partition points are automatically adjusted according to the loss function, and the final learned partition points correspond to the most discriminative frequency-band boundaries in the signal.

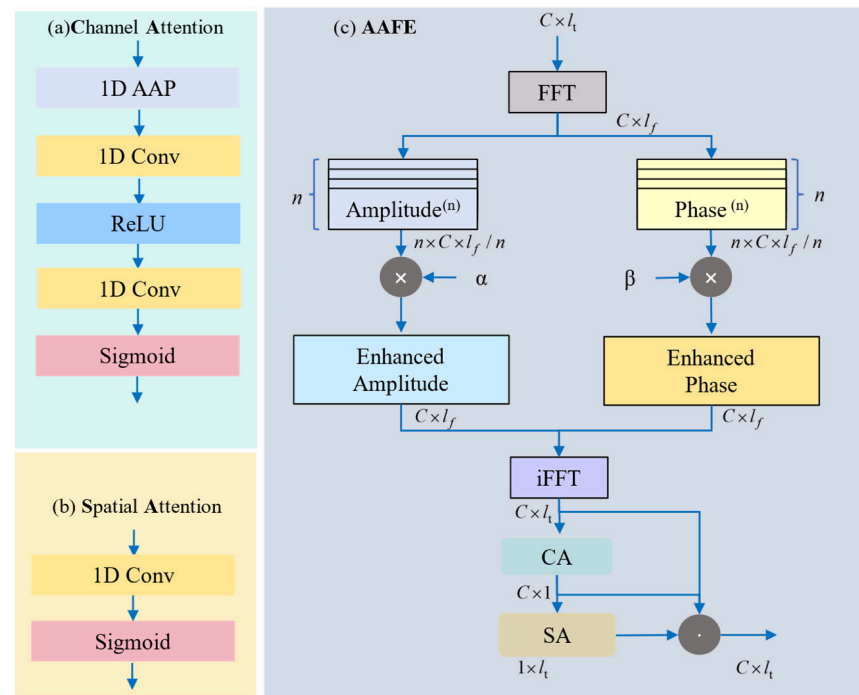


Figure 7. Structure diagram of AAFE. (a) Structure diagram of CA module. (b) Structure diagram of SA module. (c) AAFE.

Next, the input signal I is converted to the frequency domain through Fast Fourier Transform (FFT), and the input signal is converted to a complex frequency domain representation:

$$X_{\text{fft}} = \text{FFT}(I) \quad (12)$$

In the frequency domain, the signal is represented as amplitude ($|X_{\text{fft}}|$) and phase ($\angle(X_{\text{fft}})$). Through this information, the module can manipulate the frequency components of the signal. The length l_f of FFT is used to determine the absolute positions of frequency band division. By multiplying the division points φ of relative positions by the length of FFT, the actual frequency band division positions are obtained. The final division points cover the range from 0 to l_f :

$$\chi = [0, l_f \times \varphi_1, l_f \times \varphi_2, \dots, l_f \times \varphi_n, l_f] \quad (13)$$

Next, an enhancement operation is applied to each frequency band. For the current frequency band i , the amplitude A and phase P are enhanced in a learnable manner through the learnable enhancement parameters α_i and β_i :

$$\widehat{A} = A \times \alpha_i, \quad (14)$$

$$\widehat{P} = P \times \beta_i, \quad (15)$$

where $\alpha_i \in \mathbb{R}^{c \times 1}$ is the amplitude modulation coefficient of the i -th sub-band, During training, if $\alpha_i > 1$, it enhances the energy of this sub-band, and if $\alpha_i < 1$, it suppresses redundant energy. The phase spectrum of the current sub-band is modulated by $\beta_i \in \mathbb{R}^{c \times 1}$. The enhanced amplitude and phase information are combined into a complex form and converted back to a time—domain signal through the Inverse Fast Fourier Transform (iFFT):

$$\widehat{X} = \widehat{A} \cdot e^{j \cdot \widehat{P}}, \quad (16)$$

$$\widehat{I} = \text{IFFT}(\widehat{X}). \quad (17)$$

The obtained time-domain signal $\widehat{I}$ is the enhanced signal, The enhanced signal is further optimized through two attention mechanisms. First is the channel attention, which is used to screen important channels to highlight the channels that carry key UAV RF signals. Perform adaptive average pooling on $\widehat{I}$ along the length dimension to aggregate the overall energy of each channel, $c_2 = \text{AdaptiveAvgPool1d}(\widehat{I})$. Each element is the average energy of the corresponding channel: $c_2[b, c, 1] = \frac{1}{T} \sum_{t=1}^T \widehat{I}[b, c, t]$, where b and c represent the current channel index. Through two layers of 1D convolution and the Sigmoid function, convert the aggregated features into channel weights:

$$c_2 = \sigma(\text{Conv1d}(\text{ReLU}(\text{Conv1d}(c_1, c/r)), c)), \quad (18)$$

where σ is the Sigmoid function; a larger weight value indicates a more important channel. r represents the dimension reduction ratio. Multiply the enhanced signal and channel weights element-wise to amplify features of important channels:

$$I_{CA} = \widehat{I} \odot c_2. \quad (19)$$

This is the spatial attention, which is used to focus on key time-sequence positions. Apply 1D convolution to I_{CA} to extract local time-sequence features:

$$s_1 = \text{Conv1d}(I_{CA}). \quad (20)$$

Convert local features to spatial weights via the Sigmoid function, thus focusing on the importance of each time point:

$$s_2 = \sigma(s_1). \quad (21)$$

Element-wise multiply the channel-weighted signal and spatial weights to highlight features of key time points:

$$I_{SA} = I_{CA} \odot s_2. \quad (22)$$

After passing through the two attention modules, to avoid losing original signal information during the enhancement process, the module fuses the attention-weighted signal with the input signal via a residual connection:

$$f_{AAFE} = I_{SA} + \widehat{I}. \quad (23)$$

The final output not only contains the enhanced key features but also retains the basic structure of the original signal, improving the robustness of the module.

3.4. Cross-Channel Recalibration Module

The CCRM module plays a crucial role in the semantic feature extraction of drone RF signals. The design of CCRM aims to optimize features through cross-channel information exchange and attention mechanisms, thereby improving the representational power of signals. The CCRM model diagram is shown in Figure 8. After entering the CCRM module, the input feature map f_{AAFE} first passes through two convolutional layers. The convolution operation can be represented by the following formula:

$$\text{out}_1 = \text{ELU}\{\text{BN1}[\text{Conv1}(f_{AAFE})]\}, \quad (24)$$

where the convolution operation is implemented through a dilated convolutional layer. Next, the convolution output will undergo batch normalization and the ELU activation function will be applied.

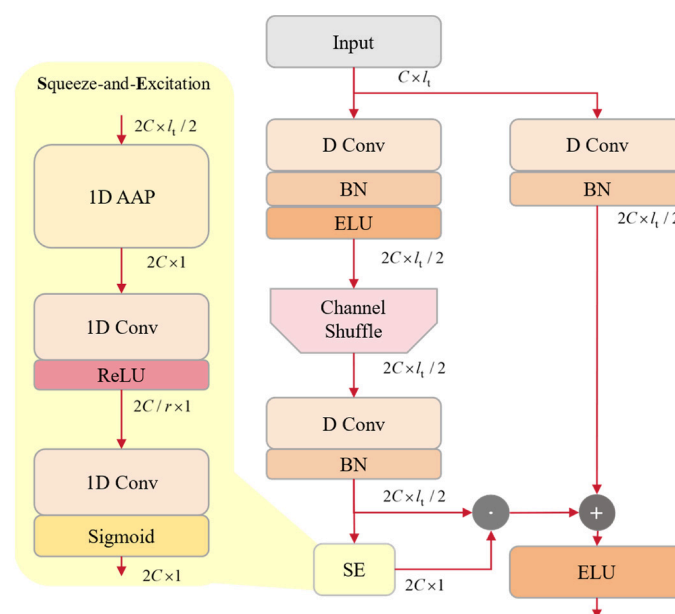


Figure 8. Diagram of the CCRM model.

At this time, the obtained output has a shape of $2C \times l_t/2$, where l_t is the sequence length of the input f_{AAFE} , and c is the number of channels of f_{AAFE} . For cross-channel information exchange, a channel shuffle operation is applied. First, the feature map is reshaped into a shape of $2C/\xi \times \xi \times l_t/2$, where ξ represents the number of groups for channel exchange. Then, the positions of channels and groups are exchanged through a transposition operation, and finally, it is reshaped back to a shape of $2C \times l_t/2$. The purpose of channel shuffle is to enhance the interaction between channels and prevent the localization of information. The shuffled feature map passes through the second convolutional layer:

$$\text{out}_2 = \text{BN2}[\text{Conv2}(\text{out}_{\text{shuffled}})]. \quad (25)$$

Finally, the output feature map out_2 undergoes the second convolution and normalization, with a shape of $2C \times l_t/2$. The feature map after convolution will be input to the SE module, which is used to recalibrate the weights of channels according to the global information of each channel.

First, perform global average pooling on the feature map out_2 to obtain the statistical information of each channel. Then, use two convolutional layers. The first layer is a one-dimensional convolution combined with the ReLU activation function, and the second layer is a one-dimensional convolution combined with the Sigmoid activation function to generate channel attention weights $\mathbf{w}_c$. Finally, use these attention weights $\mathbf{w}_c$ to weight the feature map out_2 :

$$\text{out}_3 = \text{out}_2 \cdot \mathbf{w}_c. \quad (26)$$

Each channel's feature values are scaled according to their corresponding weights, thereby enhancing the features of important channels and suppressing unimportant channels. To maintain information flow and avoid gradient vanishing, apply a residual connection to add the input feature map back to the feature map processed by the SE module:

$$\text{out}_4 = \text{out}_3 + \text{shortcut}(x), \quad (27)$$

where $\text{shortcut}(x)$ is a residual path generated through convolution and batch normalization operations.

The feature map after the residual connection will undergo a nonlinear transformation through the ReLU activation function:

$$f_{CCRM} = \text{ReLU}(\text{out}_4), \quad (28)$$

where f_{CCRM} is the feature map processed by the CCRM module.

The multi-layer CCRM gradually improves the feature expression ability of channel augmentation, combines increasing dilation coefficients to capture multi-scale features, and at the same time, uses residual connections and channel shuffle mechanisms to alleviate the gradient vanishing problem and improve the learning efficiency of the network. The specific parameters are as shown in the following Table 2.

Table 2. Core parameter configuration table of the fourth layer CCRM module.

Level	Input	Output	Kernel 1	Kernel 2	Dilation	Output
CCRM1	32	64	3	3	1	1667
CCRM2	64	128	3	3	2	834
CCRM3	128	256	3	3	4	417
CCRM4	256	512	3	3	8	209

3.5. Feature Fusion Attention

The FFA module is designed to perform weighted fusion of features from different layers. Its core idea is to dynamically adjust features based on the importance of each layer's feature map, that is, to determine the contribution of each layer's features in the final fusion by calculating the attention weight of each layer's features. Figure 9 shows the Feature Fusion Attention module, and the process can be described through the following steps.

For the output feature f_{CCRM} of each CCRM layer, the FFA calculates a weight coefficient through the SE module to reflect the importance of the features of that layer. For the feature map f_{CCRM} of each layer, the weight coefficient $\mathbf{a}_i$ can be calculated through the following steps: Global average pooling: First, perform global average pooling on each feature map to calculate the average value of each channel. Generate attention weights through convolution: Then, use a series of convolutional layers (here, 1×1 convolutions

are used) to calculate the attention weight of each channel. For the weight $\mathbf{a}_c$ of each channel, use the following formula:

$$\mathbf{a}_c = \sigma(\text{Conv2}(\text{ReLU}(\text{Conv1}(\mathbf{s}_c))))), \quad (29)$$

where σ is the Sigmoid activation function, and Conv1 and Conv2 are convolutional layers used to reduce and restore the number of channels, respectively. Weighted feature map: After obtaining the attention weights, multiply them with the original feature map to weight the features of each channel:

$$\psi_i = f_{CCRM} \odot \mathbf{a}_i, \quad (30)$$

where $\odot$ represents element-wise multiplication, and ψ_i is the weighted feature map of the i -th layer. This process generates a weighted feature map for each layer's feature map. The FFA module performs adaptive average pooling on each feature map to unify the length of the feature maps to the length $l_{\min}$ of the shortest feature map:

$$\psi_a = f_{AU}(\psi_i, l_{\min}), \quad (31)$$

where f_{AU} represents the operation of unifying the length of the feature map and adaptive average pooling, and ψ_a is the fused feature map after processing. Then, a 1×1 convolution is used to project the fused feature map to a unified number of channels. The following convolution is used:

$$f_{FFA} = \text{Conv1d}(\psi_a, C_{\max}), \quad (32)$$

where $C_{\max}$ is the maximum value of the number of output channels among the four layers of CCRM, and f_{FFA} corresponds to the final fused feature map.

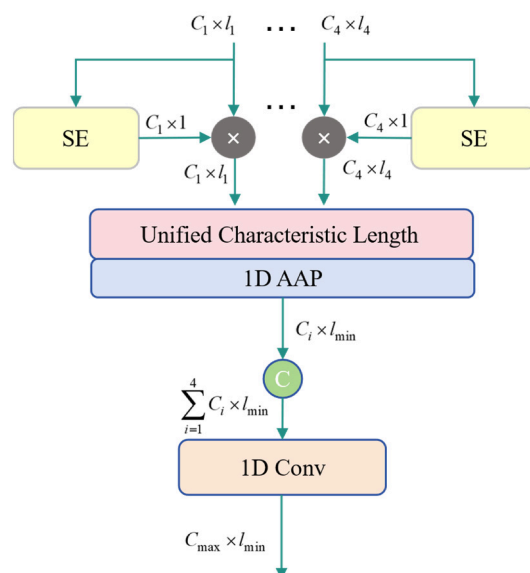


Figure 9. Diagram of the Feature Fusion Attention model.

3.6. Semantic Construction and Decision-Making

The core goal of the semantic construction and decision-making process is to extract effective features from the input signals and use them for accurate classification and recognition. The objective is to extract high-level semantic features from f_{FFA} .

The first step of semantic construction is performing global average pooling, compressing the feature dimension of f_{FFA} into a fixed-length vector. This way, the network no

longer depends on the length of f_{FFA} ; instead, it captures the global features of the entire feature through the average value. Formally, it can be expressed as:

$$\eta = \frac{1}{T} \sum_{s=1}^S f_{FFA}(s), \quad (33)$$

where $f_{FFA}(s)$ is the value of the input signal at feature point s , and S is the total number of features. After global pooling, the size of the output feature vector.

After the pooling layer, the output will be flattened into a one-dimensional vector as the input for the next layer of the network. Next, a fully connected layer is used to transform the features. In this semantic construction module, first, there is a linear layer combined with a normalization layer and a Dropout operation with a dropout rate of 0.1:

$$Semantic = BN[W_1 \cdot \eta + b_1]. \quad (34)$$

The classifier is responsible for making the final category prediction from the extracted semantic features. The classifier maps the output to the number of categories C through a fully connected layer.

3.7. Loss

In order to help the model converge quickly, this paper designs cross-entropy loss and center loss to assist the model in correctly classifying samples.

The design purpose of Center Loss is to bring the features of the same category closer to the center point of the category and push the features of different categories farther away, thereby enhancing the discrimination ability of the features learned by the network. The network maintains a center point for each category, which is randomly generated during initialization. Calculate the distance between the feature vector in each batch and the center $\mathbb{C}$ of its corresponding category. Compute the square distance matrix $\mathbb{R}$ between the feature vector and the category center. The formula used is:

$$\mathbb{R}[i, j] = \|x_i - \mathbb{C}_j\|^2. \quad (35)$$

Select the distances corresponding to the categories through masking. The final loss is obtained by summing these distances and restricting numerical stability through small and large values.

Cross-entropy loss is a standard loss function in classification tasks. It measures the difference between the predicted category probability and the true label, enabling the model to learn correct category predictions. The final loss is:

$$L = l_{cr} + \lambda \cdot l_{ce}, \quad (36)$$

where l_{cr} represents cross-entropy loss, l_{ce} represents Center Loss, and $\lambda = 0.03$ is the weight of Center Loss.

The combined loss function not only helps the model make correct predictions but also improves the model's feature learning ability, especially enhancing the separability and compactness between categories in the feature embedding space.

4. Experimental Settings

In this section, the effectiveness of the DroneTFS-SemNet is validated on the DroneRF dataset [25]. The dataset is first introduced in detail, followed by the data preprocessing procedure, experimental parameter settings, and evaluation metrics. In addition, ablation

studies and comparative experiments are conducted on the DroneRF dataset to verify the effectiveness and superiority of DroneTFS-SemNet.

4.1. Experimental Parameters and Evaluation Metrics

In this work, the Adam optimizer [26] is adopted to optimize the model parameters through gradient descent, with the learning rate set to 0.001. All experiments in this section are implemented using PyTorch 1.10.2+cu113 on a single NVIDIA A100 graphics processing unit with 40 GB of memory, together with a 20-core central processing unit and 150 GB of system memory.

Due to class imbalance in the dataset, K-fold cross-validation is employed during training. To more comprehensively evaluate the effectiveness of the proposed DroneTFS-SemNet, four statistical performance metrics are adopted, including Accuracy, Precision, Recall, and F1-score.

4.2. Experimental Results and Analysis

The experimental results in Tables 3–5 verify the effectiveness of DroneTFS-SemNet in drone detection and recognition tasks. Specifically, in the drone signal detection task, DroneTFS-SemNet mainly focuses on the differences in overall energy distribution, local amplitude variations, and frequency-band responses between background RF activities and drone communication signals. The TE module increases the diversity of input signals through controlled perturbations. The initial convolutional layer extracts local temporal fluctuation features, while the AAFE module further highlights discriminative frequency-band components and key temporal positions. These designs help the model distinguish between Background and Drone signals. As shown in Table 3, the Precision, Recall, and F1-score of the Background class remain stable under both $K = 5$ and $K = 10$.

Table 3. Performance of DroneTFS-SemNet on the drone signal detection task in terms of Precision (P), Recall (R), and F1-score.

Indicator \ Method	K = 5			K = 10		
	P	R	F1	P	R	F1
Background	0.99	0.99	0.99	1.00	1.00	1.00
Drone	1.00	1.00	1.00	1.00	1.00	1.00

Table 4. Performance of DroneTFS-SemNet on the drone classification task in terms of Precision (P), Recall (R), and F1-score.

Indicator \ Method	K = 5			K = 10		
	P	R	F1	P	R	F1
Background	1.00	1.00	1.00	1.00	1.00	1.00
Bebop	0.99	0.99	0.99	1.00	1.00	1.00
AR	0.99	0.99	0.99	1.00	1.00	1.00
Phantom	0.99	1.00	1.00	1.00	1.00	1.00

In the drone classification task, the model needs to further distinguish the RF fingerprints of different drone platforms. DroneTFS-SemNet mainly focuses on the energy distribution, spectral patterns, local pulse variations, and inter-channel response differences of different drone types in the low- and high-frequency bands. The FFA module performs weighted fusion of shallow local details and deep semantic features, enabling the model to jointly use short-time amplitude variations, frequency-band structural differences, and

cross-channel correlation information, thereby improving the separability among different drone classes.

Table 5. Performance of DroneTFS-SemNet on the operation mode recognition task in terms of Precision (P), Recall (R), and F1-score.

Indicator \ Method	K = 5			K = 10		
	P	R	F1	P	R	F1
Background	1.00	1.00	1.00	1.00	1.00	1.00
Bebop Mode1	1.00	0.99	1.00	1.00	1.00	1.00
Bebop Mode2	0.98	0.98	0.98	0.99	0.99	0.99
Bebop Mode3	0.99	1.00	1.00	1.00	1.00	1.00
Bebop Mode4	1.00	0.99	1.00	1.00	1.00	1.00
AR Mode1	0.98	0.98	0.98	0.98	0.99	0.99
AR Mode2	1.00	0.99	0.99	1.00	1.00	1.00
AR Mode3	1.00	1.00	1.00	1.00	1.00	1.00
AR Mode4	1.00	1.00	1.00	1.00	1.00	1.00
Phantom	1.00	1.00	1.00	1.00	1.00	1.00

For the operation mode recognition task, the model needs not only to identify the drone type but also to capture subtle RF variations of the same drone under different control states. DroneTFS-SemNet mainly focuses on short-time energy fluctuations, local temporal variations, frequency-band energy redistribution, and response differences between low- and high-frequency channels caused by different operation modes. AAFE helps enhance key frequency bands and important temporal positions, while CCRM captures correlated features between low- and high-frequency signals through cross-channel interaction. Therefore, the model can better distinguish the subtle differences among different flight modes of the same drone type.

Figure 10 presents the average Accuracy and Loss curves over the K-fold training process, where each fold is trained for 30 epochs. Task 1, Task 2, and Task 3 denote the drone signal detection task, drone classification task, and operation mode recognition task, respectively. When K = 5, the Loss values of all three tasks decrease rapidly and then gradually stabilize, while the Accuracy rises quickly and remains at a high level, indicating fast convergence and strong basic generalization ability of the model.

Figure 11 shows the average confusion matrices for drone signal detection. When K = 5 in Figure 11a, the recognition accuracy for both Background and Drone reaches 99%, with only a small number of misclassifications, including 0.75% for Background misclassified as Drone and 0.29% for Drone misclassified as Background. When K = 10 in Figure 11b, the accuracy is further improved and the misclassification is almost eliminated, with 0.47% for Background misclassified as Drone and 0.10% for Drone misclassified as Background.

Figure 12 shows the K-fold average confusion matrix for the drone classification task. The model achieves clear separation between Background and drone signals, indicating that it can effectively capture the energy and frequency-band response differences of drone communication signals relative to background RF activities. In contrast, slight confusion still exists between Bebop and AR. This is mainly because the two classes show high similarity in temporal amplitude variations, spectral distributions, and time–frequency structures. This phenomenon indicates that drone classification depends not only on overall energy differences but also on local frequency-band responses, short-time pulse structures, and fine-grained cross-channel features.

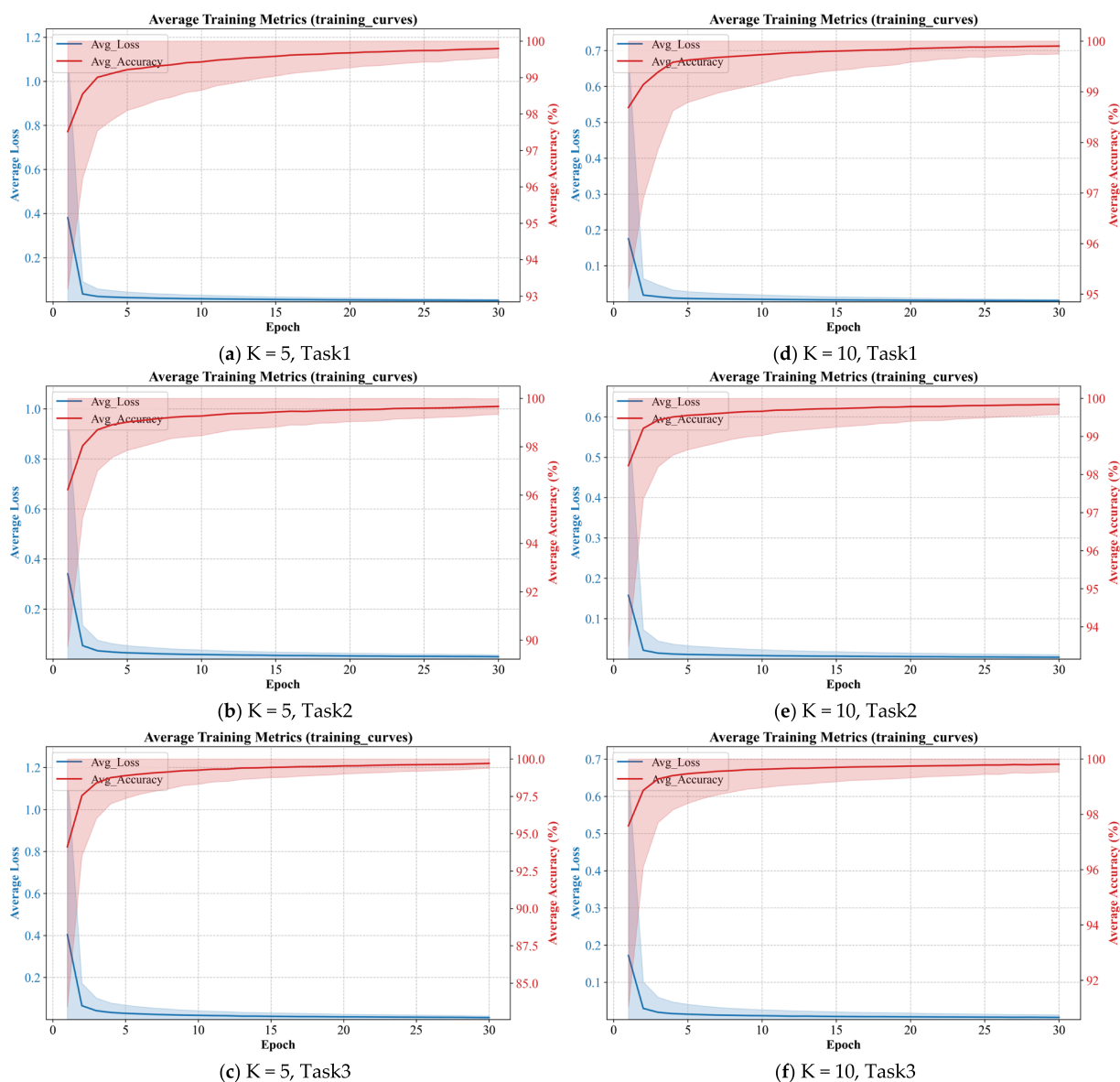


Figure 10. Average Accuracy and Loss curves over 30 epochs under K-fold cross-validation.

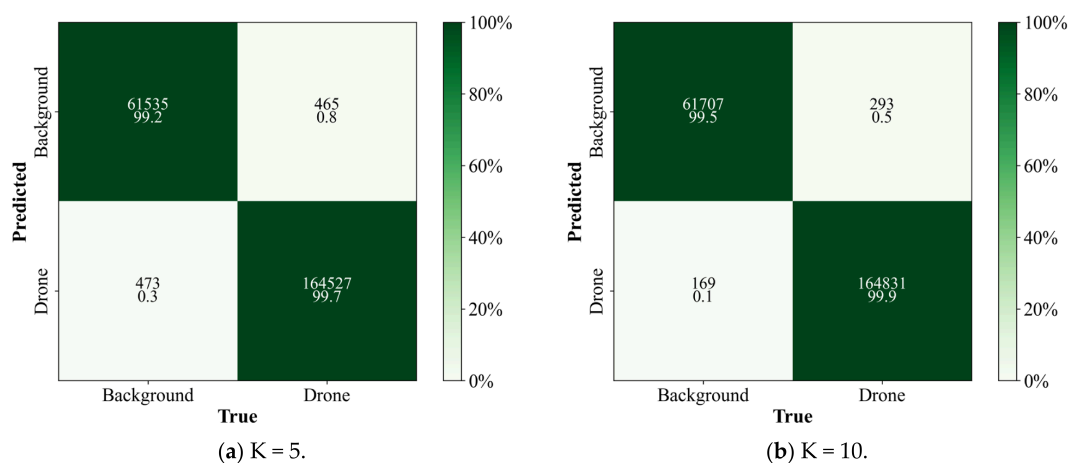


Figure 11. Average confusion matrices under K-fold cross-validation for Task 1.

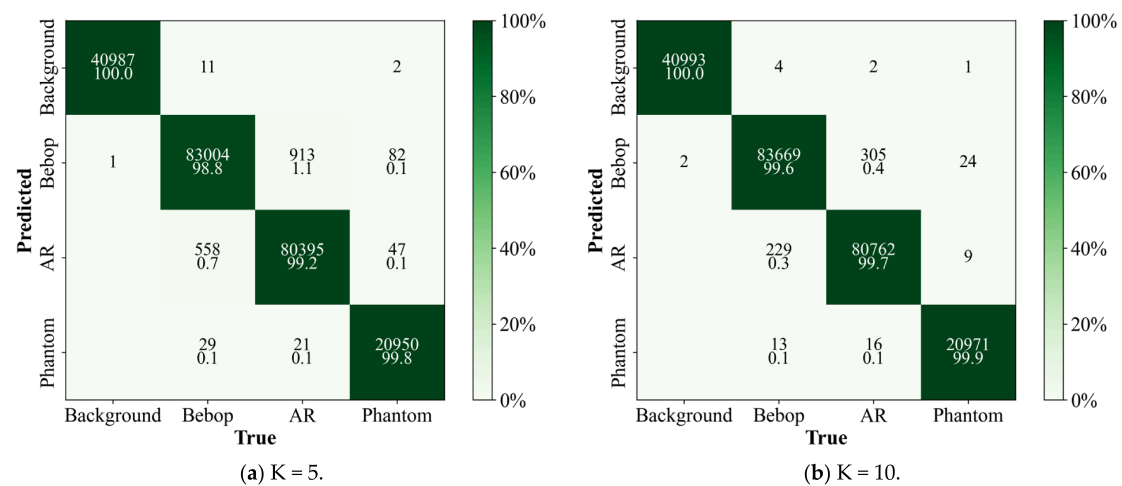


Figure 12. Average confusion matrices under K-fold cross-validation for Task 2.

Figure 13 shows the K-fold average confusion matrix for the drone operation mode recognition task. Compared with drone detection and drone type classification, operation mode recognition relies more heavily on fine-grained RF features. For the same drone type, different control commands may change the local temporal structure and frequency-band energy distribution of the communication signal. However, drones and remote controllers still continuously transmit basic information such as battery level, location, and status. Therefore, different operation modes still share some similar RF components. In this task, the model mainly focuses on short-time amplitude fluctuations, locally enhanced frequency-band regions, joint amplitude and phase variations, and the relationship between low- and high-frequency channels. These features help distinguish different operation modes within the same drone type.

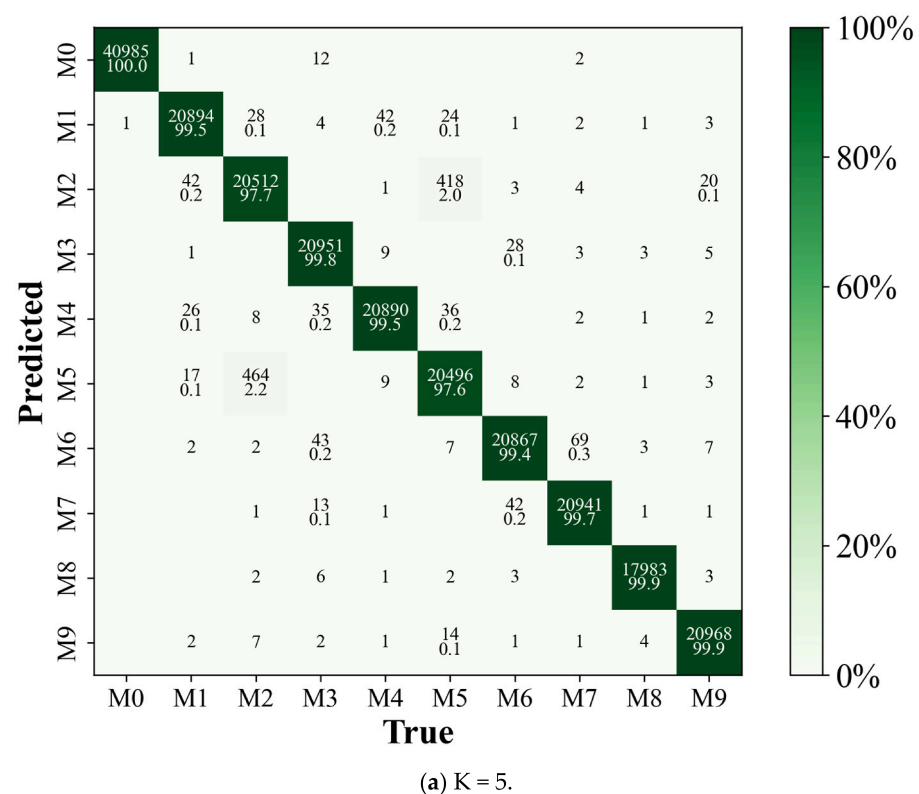


Figure 13. Cont.

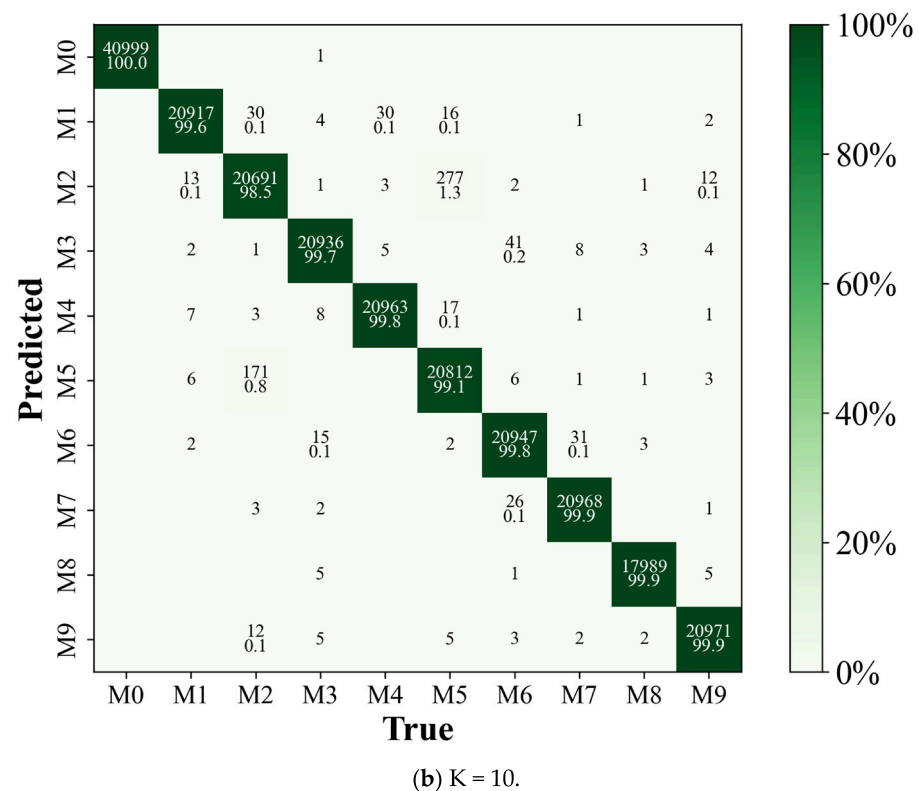
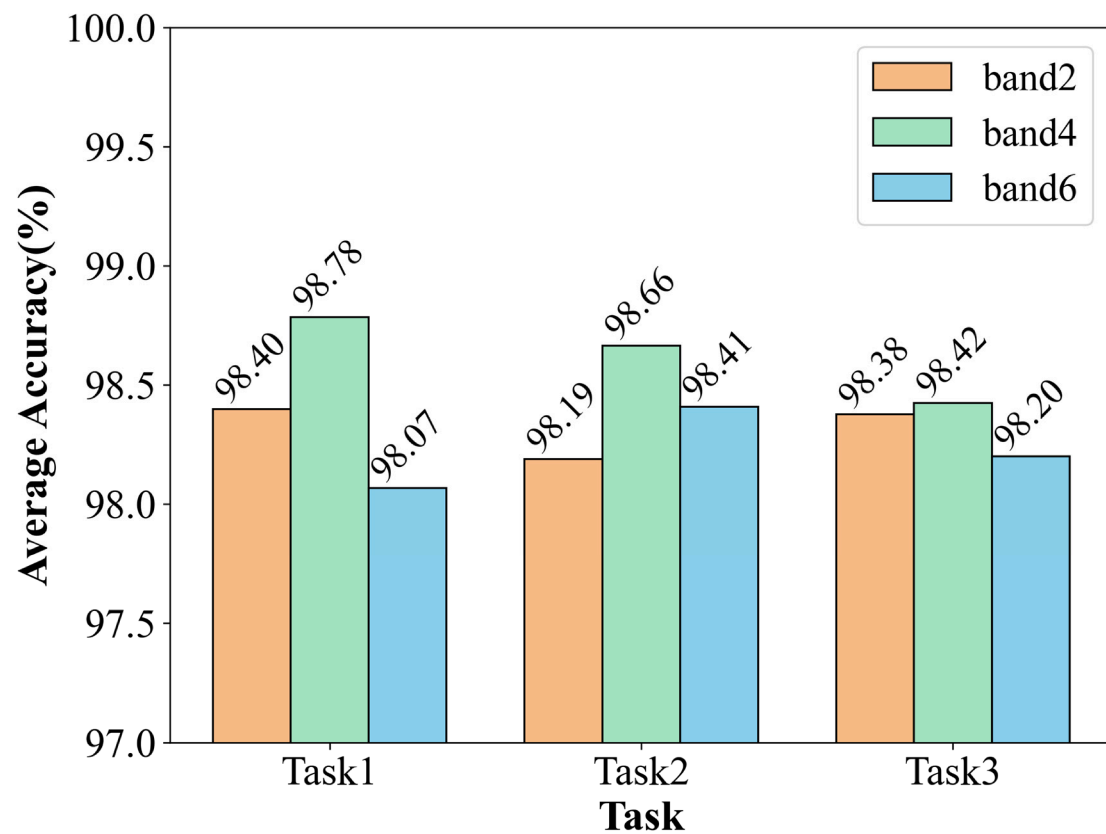


Figure 13. Average confusion matrices under K-fold cross-validation for Task 3, where M0, M1–M4, M5–M8, and M9 correspond to Background, Bebop Mode 1–4, AR Mode 1–4, and Phantom, respectively.

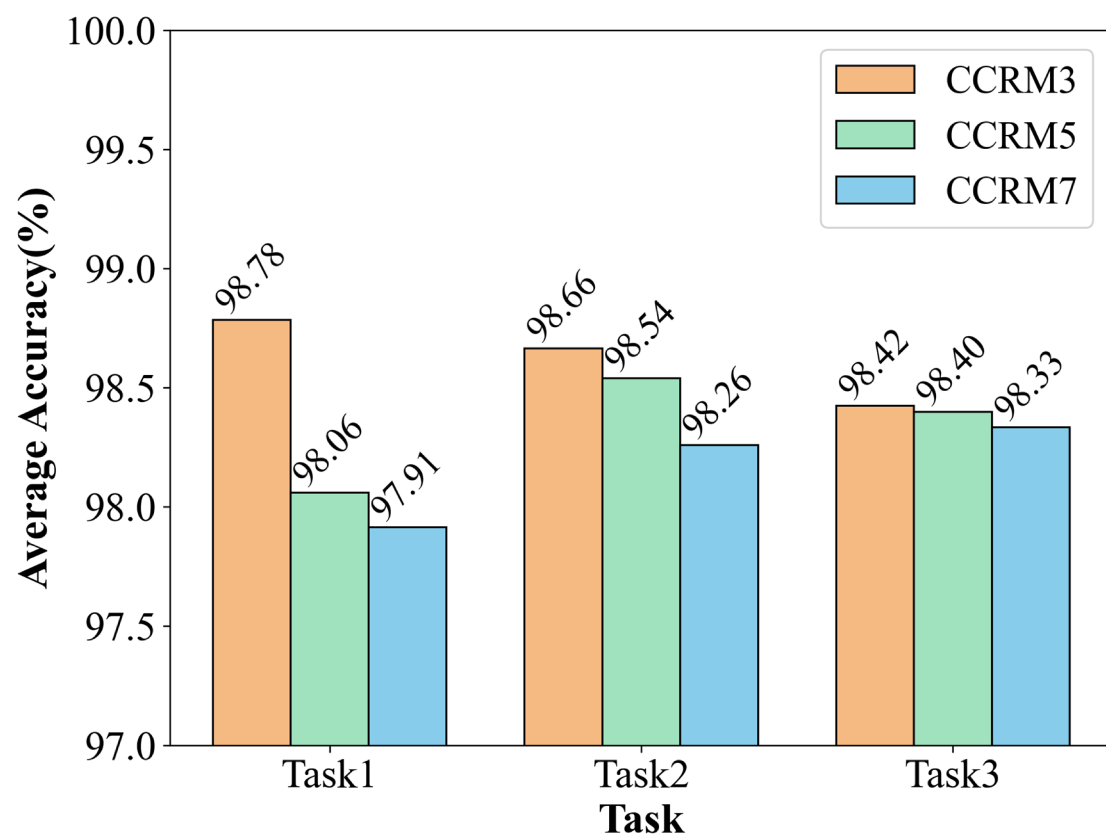
4.3. Analysis of the Impact of Model Parameters on Performance

To investigate the effects of the number of frequency-band partitions in the AAFE module, the convolution kernel size in the CCRM, and the number of channel groups used in channel shuffling on the classification performance of the DroneTFS-SemNet, experiments were conducted on the three tasks of the DroneRF dataset under different frequency-band numbers, kernel sizes, and channel group settings. The results are shown in Figure 14, where the reported values are the average accuracies obtained by five-fold cross-validation over 10 epochs.

As shown in Figure 14a, the influence of the number of frequency bands varies across different tasks. In essence, frequency-band partitioning performs spectrum segmentation and feature mining on RF signals. When the number of bands is set to 2, each band covers a broad spectral range, causing signal features to overlap and interfere within the same band, which makes it difficult for the model to precisely separate and extract discriminative features. In contrast, when the number of bands is increased to 6, although more local spectral details can be captured, the feature dimensionality and redundant information are also increased, making the model more prone to overfitting. By comparison, a four-band setting achieves a better balance between global and local feature extraction. It not only avoids feature aliasing caused by overly wide frequency bands, but also reduces the redundant interference introduced by overly fine partitioning, thereby better matching the spectral distribution characteristics of drone radio frequency signals in the DroneRF tasks.



(a) Number of frequency bands



(b) Convolution kernel size

Figure 14. Cont.

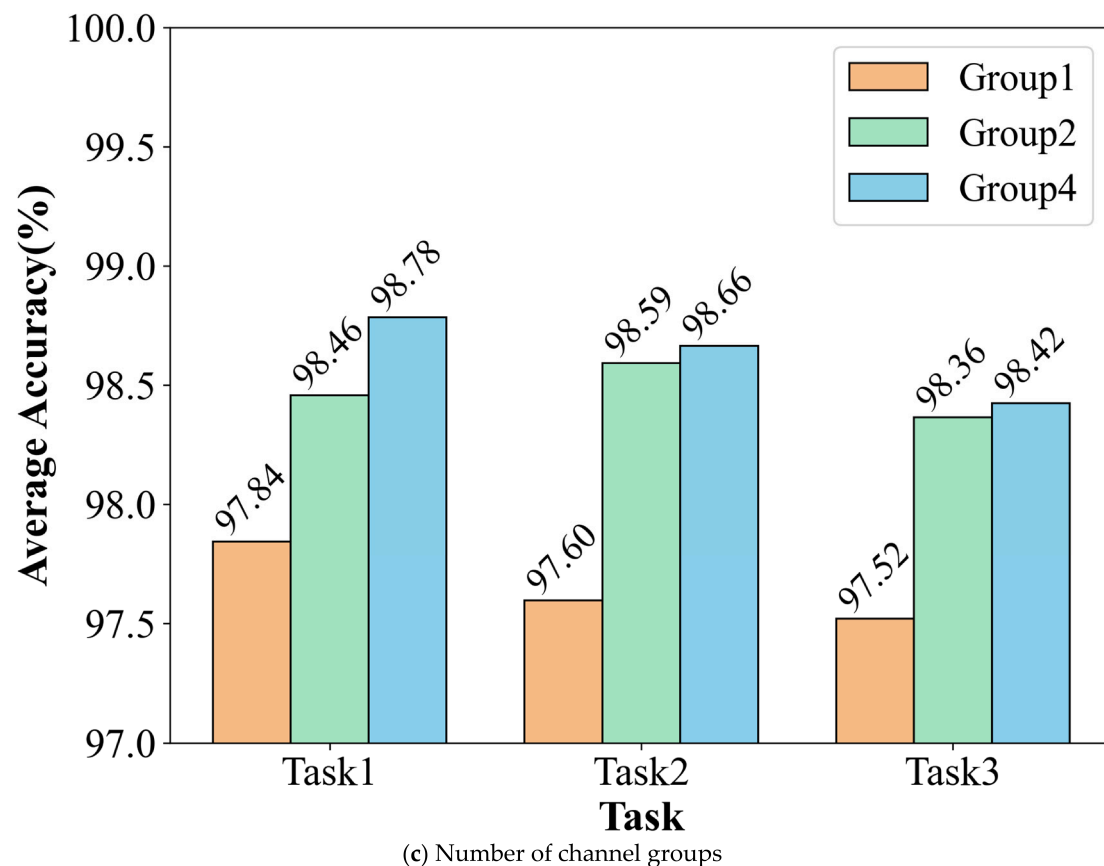


Figure 14. Analysis of the impact of different model parameters on model performance.

As shown in Figure 14b, a convolution kernel size of 3 yields the best classification performance across all three tasks. From the perspective of feature extraction, the kernel size determines the receptive field of the model and its ability to capture informative patterns. A smaller kernel provides a relatively limited receptive field and is therefore more suitable for focusing on subtle local features, which is beneficial for capturing and distinguishing the fine-grained signal differences present in the DroneRF dataset. By contrast, larger kernels with sizes of 5 and 7 provide a broader receptive field and can capture more global information, but they also tend to introduce irrelevant and redundant content, which may interfere with the identification of key features. In particular, in drone RF signal classification, local signal characteristics, such as pulses in specific frequency bands and modulation details, are crucial for classification decisions. Therefore, smaller kernels are better aligned with the required feature granularity.

The results in Figure 14c further show that the number of channel groups affects different tasks in different ways. When the group number is set to 1, each channel is processed independently, which severely limits feature interaction and prevents the model from fully exploiting complex correlations, resulting in relatively low accuracies across all tasks. When the group number is increased to 2, interaction between two-channel groups becomes possible and feature fusion is improved, but compared with the setting of 4 groups, the channel partition remains relatively coarse, leading to insufficient diversity in feature decomposition and recombination. In contrast, a group number of 4 enables finer-grained channel partitioning, allowing deeper interaction among different channel subsets and producing more diverse and task-relevant feature representations. This setting is therefore better suited to the feature extraction requirements of complex electromagnetic signals in the DroneRF dataset.

4.4. Ablation Study

To verify the contribution of each module to the performance of the DroneTFS-SemNet, ablation experiments were conducted in this section. Four ablation settings were designed and evaluated under five-fold cross-validation by comparing performance on the three tasks. Specifically, the four variants correspond to removing the TE (Net 1), removing the AAFE (Net 2), removing the CCRM (Net 3), and removing the FFA (Net 4), together with the complete DroneTFS-SemNet.

As shown by the ablation results in Table 6, the complete DroneTFS-SemNet achieves substantially higher average accuracy on all three tasks than all ablated variants, reaching 99.584% for Task 1, 99.266% for Task 2, and 99.334% for Task 3. In addition, its standard deviation is the best for Task 1 and the second best for Tasks 2 and 3. These results indicate that the collaboration of all modules effectively improves both classification accuracy and model stability. A comparison among the ablated variants further reveals clear differences in the impact of removing different modules. Net 2 shows the largest decrease in average accuracy on Task 1, with a drop of 2.216%, indicating that the Adaptive Attention Frequency-Band Enhancement module is critical to the strong performance on Task 1. Net 1 exhibits a certain decrease in average accuracy across all three tasks, demonstrating the importance of the Temporal Enhancement module to the overall performance.

Table 6. Accuracy and standard deviation of different methods in the ablation experiments, where AA denotes average accuracy and STD denotes standard deviation.

Indicator \ Task	Task1		Task2		Task3	
	AA (%)	STD	AA (%)	STD	AA (%)	STD
Net 1	97.902	0.926	97.996	0.911	97.904	0.932
Net 2	97.368	0.862	97.852	0.882	97.920	0.731
Net 3	98.536	1.441	99.000	0.535	98.800	0.626
Net 4	98.374	0.926	98.096	0.780	98.066	0.878
Ours	99.584	0.501	99.266	0.569	99.334	0.706

To further verify the contribution of each component in DroneTFS-SemNet, this paper conducts component-level ablation experiments under the 10-class task with 5-fold cross-validation. The complete DroneTFS-SemNet is used as the baseline. Specifically, noise injection, amplitude scaling, and random temporal discarding in TE are separately removed. Channel Shuffle, SE, dilated convolution, residual connection, and the second convolutional layer in CCRM are also separately removed. In addition, layer attention and projection in FFA are removed. All ablation models adopt the same data split and training settings, and the average accuracy is used for comparison. The results are shown in Figure 15.

As shown in Figure 15, the complete DroneTFS-SemNet achieves the highest accuracy. Removing the layer attention in FFA leads to the largest accuracy drop, with a decrease of 4.88 percentage points, indicating that adaptive weighting of multi-level features is the most critical for the 10-class task. Removing noise injection and random temporal discarding in TE decreases the accuracy by 3.83 and 3.10 percentage points, respectively, showing that TE helps improve the model's adaptability to input perturbations and local temporal variations. In CCRM, removing Channel Shuffle, dilated convolution, and SE decreases the accuracy by 2.99, 2.51, and 1.40 percentage points, respectively. This indicates that cross-channel information exchange, multi-scale receptive fields, and channel recalibration can all improve RF feature representation. In contrast, removing the residual connection and the second convolutional layer has a smaller effect, with accuracy drops of 0.97 and 0.67 percentage points, respectively, suggesting that they mainly contribute to training

stability and feature refinement. Overall, the results show that the performance improvement of DroneTFS-SemNet comes from the collaboration of multiple components, among which FFA attention, TE-based perturbation enhancement, and CCRM-based cross-channel modeling provide the most significant contributions.

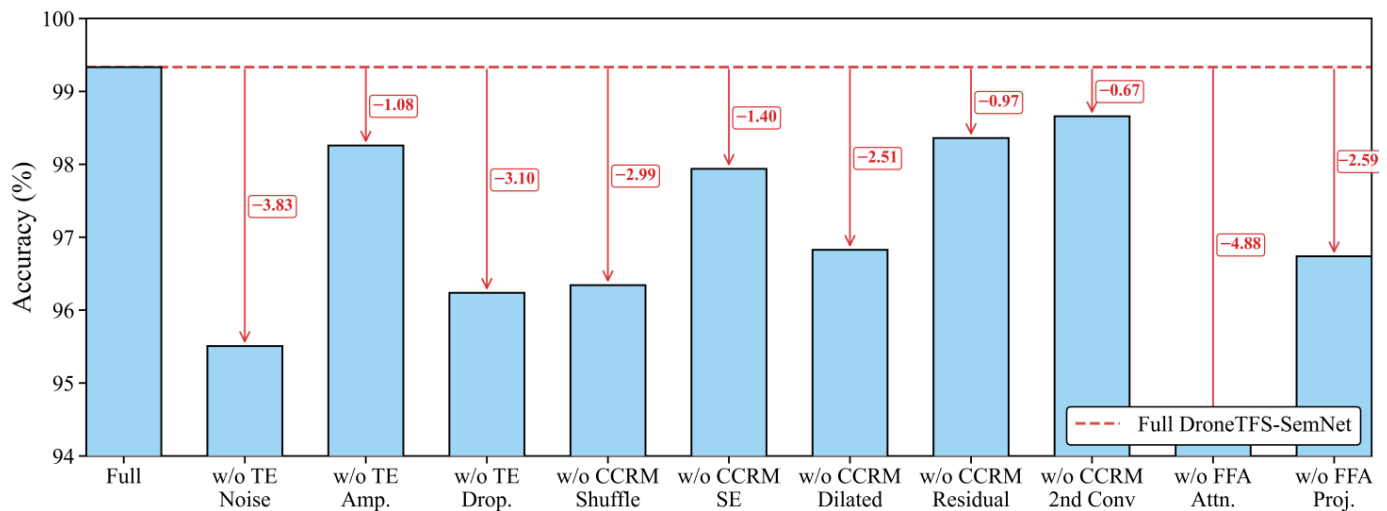


Figure 15. Accuracy comparison of the component-level ablation experiments.

4.5. Comparative Experiments and Result Analysis

To comprehensively evaluate the effectiveness of DroneTFS-SemNet in drone RF recognition tasks, this section compares it with several representative baseline methods on the DroneRF dataset. The evaluated tasks include UAV detection, UAV classification, and UAV operation mode recognition. The selected comparison methods include BISSIAM [27], DNN [28], 1D-CNN [29], MC-CNN [30], RF-UAVNet [31], DRNN [32], XGBoost [33], TFMS-CNN [34], Artificial Neural Network Model (ANNM) [35], Topology Compression Network for Drone RF Fingerprint Identification (RF-TCNet) [36], and M-DS-UAVNet [37]. To reduce the randomness caused by a single data split, DroneTFS-SemNet is evaluated using both 5-fold and 10-fold cross-validation strategies. Tables 7 and 8 report the average results over the test sets of all folds. In addition to accuracy, Precision, Recall, and F1-score are also reported to evaluate the recognition ability of the model from different perspectives, especially under class-imbalanced conditions.

Table 7 reports the Precision, Recall, and F1-score of different methods. For UAV detection, most advanced methods have already reached near-saturated performance. BISSIAM, XGBoost, and DroneTFS-SemNet all achieve near-perfect or perfect results. Thus, the advantage of DroneTFS-SemNet in this task is that it maintains high detection reliability while preserving the ability to support more complex classification and operation mode recognition. For UAV classification, the advantage of DroneTFS-SemNet is more evident. Compared with strong baselines such as BISSIAM, RF-UAVNet, and TFMS-CNN, DroneTFS-SemNet achieves a higher F1-score. This shows that the model can not only detect drones but also distinguish different drone types more accurately. This improvement may come from three aspects. First, TE improves the diversity of training samples through noise perturbation, amplitude scaling, and random temporal discarding, making the model more stable against signal strength changes and local missing information. Second, AAFE highlights key frequency components and important temporal positions through learnable frequency-band partitioning and channel-spatial attention. Third, CCRM and FFA further fuse semantic features at different levels, allowing the model to use both local details and global category information. For UAV operation mode recognition, DroneTFS-SemNet achieves the most significant improvement. This task is more difficult than detection and

drone type classification because different flight modes often come from the same drone type and their RF differences are more subtle. Traditional DNN, 1D-CNN, and DRNN have limited ability to model such fine-grained differences, so their performance drops clearly in this task. In contrast, DroneTFS-SemNet captures weak differences among operation modes through adaptive frequency-band enhancement, multi-scale semantic extraction, and multi-level feature fusion. Therefore, its F1-scores under both $K = 5$ and $K = 10$ are higher than most baselines, showing its effectiveness for fine-grained flight mode recognition.

Table 7. Accuracy comparison in terms of Precision (P), Recall (R), and F1-score.

Method	UAV Detection			UAV Classification			UAV Operation Mode Recognition		
	P	R	F1	P	R	F1	P	R	F1
BISSIAM	1	1	1	0.978	0.975	0.975	0.988	0.706	0.744
DNN	0.996	0.994	0.996	0.910	0.764	0.788	0.535	0.419	0.430
1D-CNN	0.998	0.995	0.996	0.913	0.817	0.845	0.589	0.555	0.551
MC-CNN	-	-	0.995	-	-	0.910	-	-	0.951
RF-UAVNet	-	-	0.998	-	-	0.985	-	-	0.951
DRNN	0.999	0.999	0.998	0.908	0.817	0.846	0.589	0.555	0.551
XGBoost	0.999	1	1	0.900	0.920	0.900	0.650	0.640	0.640
TFMS-CNN	-	-	0.999	-	-	0.989	-	-	0.868
RF-TCNet	-	-	-	0.950	0.949	0.949	-	-	-
M-DS-UAVNet	-	-	-	-	-	0.984	-	-	0.982
DroneTFS-SemNet ($K = 5$)	1	1	1	0.993	0.995	0.995	0.995	0.993	0.995
DroneTFS-SemNet ($K = 10$)	1	1	1	1	1	1	0.997	0.998	0.998

Table 8. Accuracy comparison in terms of Accuracy.

Method	Task1 (%)	Task12 (%)	Task3 (%)
BISSIAM	100	98.57	92.31
DNN	99.71	84.52	46.83
1D-CNN	99.81	85.45	59.19
MC-CNN	99.95	94.55	87.37
RF-UAVNet	99.85	98.55	95.33
DRNN	99.90	90.00	56.00
XGBoost	99.96	90.73	70.09
TFMS-CNN	99.89	98.56	87.67
ANNM	99.46	98.52	83.40
RF-TCNet	-	94.89	-
Ours ($K = 5$)	99.58	99.27	99.33
Ours ($K = 10$)	99.80	99.73	99.65

Table 8 further reports the Accuracy comparison. For UAV detection, several models achieve over 99% accuracy, indicating that this task is relatively easy on the DroneRF dataset. DroneTFS-SemNet is slightly lower than BISSIAM and XGBoost when $K = 5$, but improves to 99.80% when $K = 10$, reaching a level comparable to top baselines. Since this task is already close to saturation, small differences in detection accuracy should not be overinterpreted. More importantly, the model should maintain its advantage in more complex

classification and operation mode recognition tasks. For UAV classification, DroneTFS-SemNet outperforms existing high-performance baselines under both $K = 5$ and $K = 10$. Unlike detection, classification requires the model to distinguish RF fingerprints from different drone platforms. Therefore, it depends more on the modeling of frequency-band energy distribution, local temporal fluctuations, and cross-channel response differences. AAFE adaptively enhances key frequency bands, while CCRM recalibrates related information in low- and high-frequency channels. Together, they enlarge the feature gap among different drone classes. In addition, Center Loss further improves intra-class compactness in the semantic space, thereby enhancing classification stability. For UAV operation mode recognition, DroneTFS-SemNet achieves the largest improvement over most baselines. This result is practically meaningful because operation mode recognition is closely related to drone behavior understanding and low-altitude security monitoring. Compared with drone detection, this task needs to identify subtle RF variations of the same drone under different control states. These variations usually appear in short-time energy fluctuations, local temporal structures, specific frequency-band responses, and the correlation between low- and high-frequency channels. DroneTFS-SemNet uses multi-scale dilated convolution to enlarge the receptive field and uses FFA to fuse shallow local details with deep semantic information. Therefore, it is more suitable for capturing such fine-grained mode differences. The further improvement under $K = 10$ also shows the stability of the model under a more sufficient cross-validation setting.

Under the 10-class task with 5-fold cross-validation, this paper further compares the computational efficiency of DroneTFS-SemNet with baseline models such as DNN, 1D-CNN, and MC-CNN. The evaluated metrics include single-sample inference latency, parameter number, model size, and FLOPs. Single-sample inference latency measures the average time required to process one input sample. Parameter number and model size reflect the storage cost, while FLOPs indicate the computational complexity of one forward pass. The results are shown in Table 9. As shown in Table 9, DroneTFS-SemNet achieves a single-sample inference latency of 0.2396 ms, corresponding to a throughput of about 4173.62 samples/s. Although its inference speed is lower than that of lightweight baselines such as DNN, 1D-CNN, and MC-CNN, it still remains below the millisecond level, indicating high online processing efficiency on a high-performance GPU. Compared with DNN, DroneTFS-SemNet has far fewer parameters, and its model size is reduced from 39.7126 MB to 10.1508 MB. This shows that the proposed method does not rely on a large number of fully connected parameters. Instead, it improves recognition through convolutional feature extraction, frequency-band enhancement, cross-channel feature recalibration, and multi-level feature fusion.

Table 9. Comparison of inference efficiency and computational complexity of different models in the 10-class task.

Method	Latency (ms/Sample)	Parameters	Model Size (MB)	FLOPs (M)
DroneTFS-SemNet	0.2396	2,641,572	10.1508	1236.8160
DNN	0.0665	10,407,562	39.7126	20.8102
1D-CNN	0.0976	44,842	0.1800	234.2577
MC-CNN	0.0945	10,442	0.0500	55.6895
DRNN	0.1283	55,498	0.2140	50.8654
TFMS-CNN	0.1055	29,866	0.1283	487.2225

Meanwhile, DroneTFS-SemNet has 1236.8160 M FLOPs, which is higher than DNN, 1D-CNN, and MC-CNN. This is mainly because AAFE, CCRM, and FFA are introduced into the model, and CCRM uses multi-layer dilated convolution and channel recalibration. These designs increase the computational cost, but they also enhance the model's ability to represent fine-grained RF features. In particular, DroneTFS-SemNet achieves better recognition performance than lightweight baselines in UAV classification and operation mode recognition, indicating that the additional computational cost brings practical benefits in complex recognition tasks.

4.6. AAFE Feature Visualization

To verify the contribution of AAFE to UAV RF signal representation, we visualized its input and output features. First, the raw signal was fed into the initial convolutional layer to extract low-level features, producing the initial feature map before AAFE, as shown in Figure 16a. Next, the same feature was passed through the AAFE module to obtain the output feature map after frequency-band enhancement and attention-based reweighting, as shown in Figure 16b. Finally, the difference between the enhanced feature and the original feature was calculated to produce the difference response map, as shown in Figure 16c. The results show that the output feature of AAFE remains consistent with the input feature in overall texture distribution, indicating that the residual enhancement strategy preserves the basic structure of the original RF signal. At the same time, the enhanced feature exhibits stronger responses and clearer hierarchical patterns in several channels, and the difference map also shows obvious local enhancement and suppression regions. This indicates that AAFE mainly focuses on three types of RF features. The first is energy enhancement or attenuation in specific frequency bands. The second is amplitude variation at short-time local temporal positions. The third is response differences among different channels. These features are closely related to the spectral distribution, local pulse structures, and coordinated low–high-frequency variations of drone communication signals. Therefore, they provide more discriminative representations for drone detection, classification, and operation mode recognition.

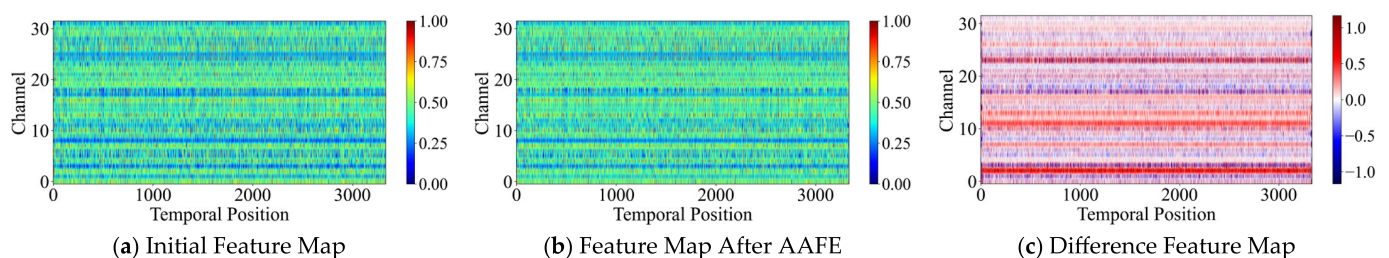


Figure 16. Visualization of feature maps before and after AAFE.

4.7. Robustness Analysis

To further evaluate the recognition stability of DroneTFS-SemNet under noise perturbation, Gaussian noise was added to the test sets of the 2-class, 4-class, and 10-class tasks under 5-fold cross-validation. The SNR was set from 5 dB to 20 dB with an interval of 5 dB. The results are shown in Figure 17, where Detection, Classification, and Operation Mode Recognition correspond to drone detection, drone classification, and operation mode recognition, respectively. As shown in Figure 17, different tasks are affected by noise to different degrees under low-SNR conditions. The Detection task still achieves an Accuracy of 72.69% at 5 dB, indicating that the model can still capture the overall difference between drone RF signals and background RF activity under relatively strong noise. In contrast, Classification and Operation Mode Recognition are more sensitive to noise. At 5 dB, their Accuracy values are 41.40% and 23.09%, and their F1-scores are 30.72% and 13.81%, re-

spectively. This suggests that drone type recognition and operation mode recognition rely more on local temporal structures, fine-grained frequency-band responses, and differences between low- and high-frequency channels, which are more easily masked under low-SNR conditions. When the SNR increases from 5 dB to 10 dB, the performance of Classification and Operation Mode Recognition recovers clearly. The F1-score of Classification increases by 28.07 percentage points, while the F1-score of Operation Mode Recognition increases by 50.74 percentage points, and its Accuracy also improves by 44.70 percentage points. This indicates that DroneTFS-SemNet is not only effective under high-SNR conditions, but can also quickly recover its ability to model key RF features during the early stage of SNR improvement. This result further shows that the combination of TE, AAFE, CCRM, and FFA helps improve feature recovery under noise perturbation.

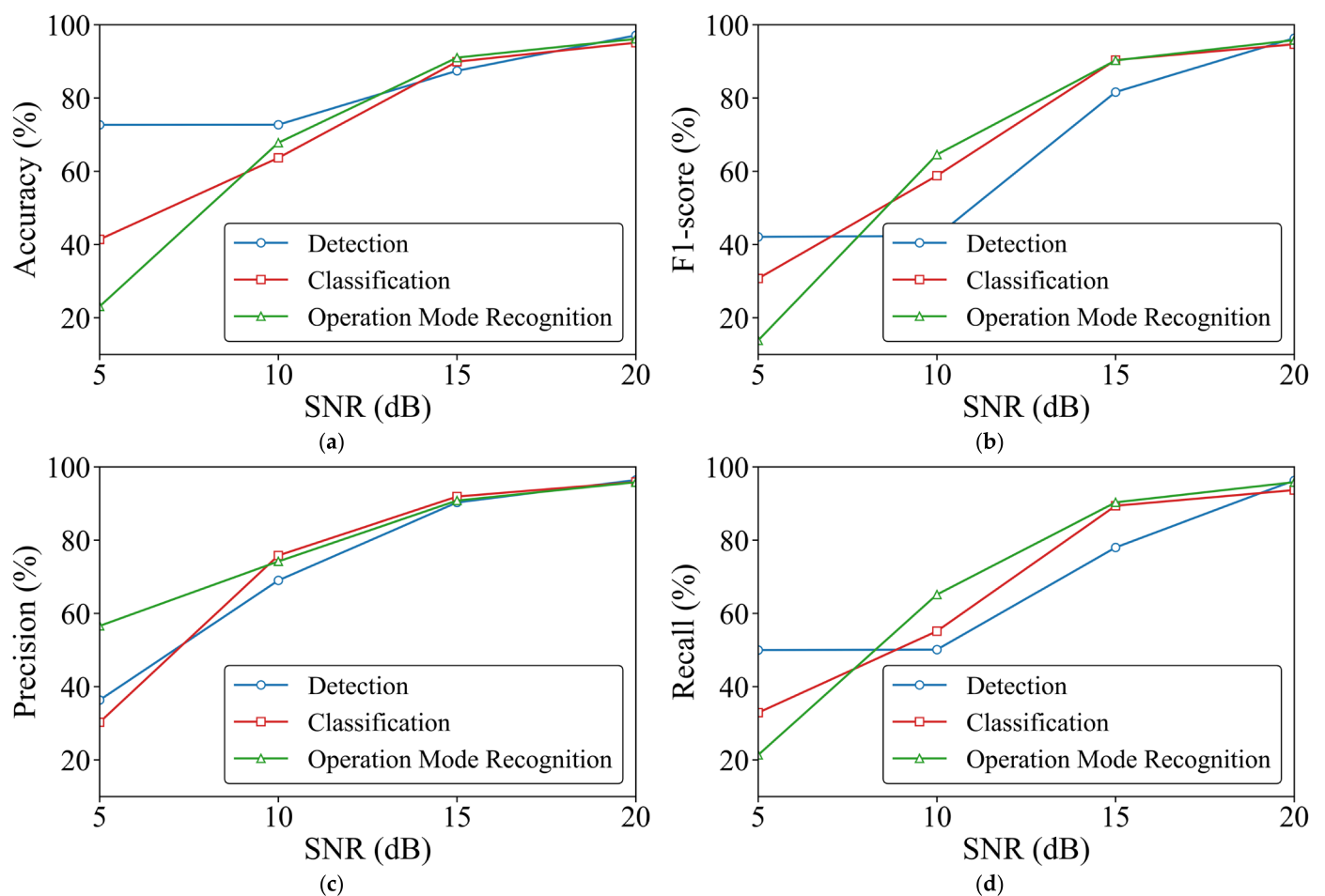


Figure 17. Robustness analysis of DroneTFS-SemNet under Gaussian-noise environments. (a) Accuracy, (b) F1-score, (c) Precision, and (d) Recall.

5. Limitations

Although DroneTFS-SemNet achieves strong performance on the three recognition tasks of the DroneRF dataset, its applicable scope should be discussed objectively. First, the experiments in this paper are mainly based on the DroneRF dataset. This public dataset is useful for fair comparison with existing methods. However, a single dataset cannot cover all variations in real RF-based drone monitoring scenarios, such as differences in acquisition devices, signal propagation distance, antenna direction, environmental noise, and background electromagnetic activity. Therefore, the current results mainly demonstrate the recognition ability of the model within the DroneRF dataset, and should not be directly generalized to all real-world scenarios.

Second, in practical applications, the training and test data may not follow the same distribution. When drone types, remote-control protocols, acquisition devices, or environmental conditions change, the spectral structure, amplitude distribution, and local temporal features of RF signals may also shift. Although the TE module improves sample diversity through noise perturbation, amplitude scaling, and random temporal discarding, and AAFE and CCRM enhance the modeling of key frequency bands and channel features, these designs still mainly improve robustness within the existing data distribution. For unseen drone types, unknown operation modes, and new RF environments, further cross-dataset testing and real-field validation are still required.

In addition, the anti-interference ability discussed in this paper is mainly evaluated under simulated perturbation conditions. Simulated noise can partly reflect signal fluctuations in practical environments, but real electromagnetic environments are usually more complex. They may include co-channel interference, multipath fading, non-stationary noise, signals from other wireless devices, and SNR variations caused by different monitoring distances. Therefore, the current experiments only show that the model has a certain adaptability to controlled perturbations, but they cannot fully represent its robustness in open and complex RF environments. Future work should include field-collected data to further evaluate the model under different SNRs, different acquisition distances, and multi-source interference conditions.

Finally, DroneTFS-SemNet combines TE, AAFE, CCRM, FFA, and the semantic decision module. This design improves multi-level feature representation, but it also increases computational cost. For offline analysis, the model can provide strong recognition performance. However, for edge devices, embedded platforms, or real-time RF monitoring systems, model size, inference latency, throughput, and memory usage still need further consideration. In the future, model pruning, knowledge distillation, lightweight convolution, and hardware acceleration can be introduced to optimize DroneTFS-SemNet and improve its deployment feasibility in practical drone monitoring systems.

6. Conclusions

To address the public safety and privacy risks caused by drone misuse, as well as the limitations of existing methods in RF feature representation, cross-channel information modeling, and multi-task collaborative processing, DroneTFS-SemNet provides an effective method for RF-based drone detection, classification, and operation mode recognition. Specifically, the TE increases the diversity of training samples through random noise perturbation, amplitude scaling, and random temporal discarding, providing richer input patterns for subsequent feature extraction. The CCRM further enhances the discriminability of target signals, drone categories, and operation modes by promoting cross-channel information interaction. The FFA enables effective multi-level feature fusion and fully integrates multi-dimensional RF information, thereby strengthening the model's ability to represent complex patterns. The semantic construction and decision module completes the unified transition from feature extraction to semantic inference, ultimately achieving accurate drone signal detection, classification, and recognition. Experimental results show that DroneTFS-SemNet achieves strong performance on the three tasks of the DroneRF dataset, with most F1-scores exceeding 0.99. This indicates that the model has strong recognition capability on this dataset and under controlled perturbation conditions. Compared with baseline methods such as BISSIAM, the proposed method achieves higher accuracy in drone classification and operation mode recognition, verifying the effectiveness of the proposed framework on the DroneRF dataset. It should be noted that the current experiments are mainly based on a public dataset and simulated perturbation conditions. Further validation on real field-collected data, cross-dataset scenarios, and more complex

interference conditions is still needed to evaluate the model's generalization ability and engineering applicability.

Author Contributions: Y.S.: Methodology, Original draft preparation; N.W.: Manuscript checking; Y.Y.: Funding acquisition; Y.L.: Visualization; B.X.: Methodology; H.W.: Manuscript checking; L.M.: Data acquisition; H.L.: Manuscript checking. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the National Key R&D Program of China [grant number 2023YFF0717402]. We thank all the anonymous reviewers who generously contributed their time and efforts. Their professional recommendations have greatly enhanced the quality of the manuscript. The authors would like to express their gratitude to all collaborators and institutions that provided valuable support throughout this study.

Data Availability Statement: No new data were created or analyzed in this study.

Conflicts of Interest: All authors of this article hereby declare that there are no conflicts of interest that may affect the objectivity and impartiality of the research results in the design, implementation, data analysis, and writing process of this study (or this article).

References

1. Graziano, B.; Sargolzaei, A. Cybersecurity of Unmanned Aerial Vehicles from a Control Systems Perspective: A Review. *Electronics* **2026**, *15*, 163. [\[CrossRef\]](#)
2. Cheng, L.; Shang, W. Multi-Head Attention DQN and Dynamic Priority for Path Planning of Unmanned Aerial Vehicles Oriented to Penetration. *Electronics* **2026**, *15*, 167. [\[CrossRef\]](#)
3. Belmonte-Hernández, A.; Grauby, B.; Fernández García, A.; Tardi, S.; Houge, T.; Bango, H.G.; Gutiérrez, Á. AI-Driven Tethered Drone Surveillance for Maritime Security in Ports and Coastal Areas. *Drones* **2026**, *10*, 268. [\[CrossRef\]](#)
4. Zhong, J.; Yang, J.; Yu, W.; Tan, K. Rotary-Wing UAV Target Recognition Method Based on Radar Measured Data. In Proceedings of the 2024 9th International Conference on Intelligent Computing and Signal Processing (ICSP), Xi'an, China, 9–21 April 2024. [\[CrossRef\]](#)
5. Hong, T.; Fang, C.; Hao, H.; Sun, W. Identification Technology of UAV Based on Micro-Doppler Effect. In Proceedings of the International Wireless Communications and Mobile Computing Conference, Harbin, China, 28 June–2 July 2021.
6. Zhang, L.; Tu, G.; Xu, Y.; Zhou, X. A Lightweight CNN-Based Method for Micro-Doppler Feature-Based UAV Detection and Classification. *Electronics* **2025**, *14*, 4831. [\[CrossRef\]](#)
7. Jiang, T.; Li, Q.; Huang, Z.; Zhuang, L. CT-RURnet: A novel network design for radar unmanned aerial vehicles recognition. *Meas. Sci. Technol.* **2025**, *36*, 026101. [\[CrossRef\]](#)
8. Ren, J.; Wang, Z.; Zhao, J.; Liu, X. A Drone Sound Recognition Approach Using Adaptive Feature Fusion and Cross-Attention Feature Enhancement. *Electronics* **2025**, *14*, 1491. [\[CrossRef\]](#)
9. Najafi, J.; Mirzakuchaki, S.; Shamaghdari, S. Autonomous Drone Detection and Classification Using Computer Vision and Prony Algorithm-Based Frequency Feature Extraction. *J. Intell. Robot. Syst.* **2025**, *111*, 8. [\[CrossRef\]](#)
10. Zhang, M.; Lu, L.; Wu, Y.; Yan, Z.; Sun, J.; Lin, F.; Ren, K. DroneAudioID: A Lightweight Acoustic Fingerprint-Based Drone Authentication System for Secure Drone Delivery. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 1447–1461. [\[CrossRef\]](#)
11. Liu, Z.; Fan, K.; Chen, Y.; Xiong, L.; Ye, J.; Fan, A.; Zhang, H. Deep Learning-Based Acoustic Recognition of UAVs in Complex Environments. *Drones* **2025**, *9*, 389. [\[CrossRef\]](#)
12. Zou, K.; Zhao, H.; Yan, X.; Wang, W.; Zhang, Y.; Zhang, Y. MVFF: Multi-View Feature Fusion Network for Small UAV Detection. *Drones* **2026**, *10*, 264. [\[CrossRef\]](#)
13. Seidaliyeva, U.; Akhmetov, D.; Ilipbayeva, L.; Matson, E.T. Real-Time and Accurate Drone Detection in a Video with a Static Background. *Sensors* **2020**, *20*, 3856. [\[CrossRef\]](#)
14. Sun, H.; Yang, J.; Shen, J.; Liang, D.; Ning-Zhong, L.; Zhou, H. TIB-Net: Drone Detection Network with Tiny Iterative Backbone. *IEEE Access* **2020**, *8*, 130697–130707. [\[CrossRef\]](#)
15. Zhou, M.; He, S.; Wang, C.; Wang, J. EFSL-YOLO: An Improved Model for Small Object Detection in UAV Vision. *Drones* **2026**, *10*, 243. [\[CrossRef\]](#)
16. Yu, J.; Qin, M.; Han, L.; Lu, S.; Zhou, Y.; Jiang, J. High-Precision Individual Identification Method for UAVs Based on FFS-SPWVD and DIR-YOLOv11. *Electronics* **2026**, *15*, 680. [\[CrossRef\]](#)
17. Choudhary, P.; Sihag, V.; Choudhary, G.; Dragoni, N. DrIfTeR: A Drone Identification Technique using R F signals. *Forensic Sci. Int. Digit. Investig.* **2025**, *54*, 301948. [\[CrossRef\]](#)

18. Rahman, M.H.; Baik, J.I.; Aziz, M.A.; Tabassum, R.; Sejan, M.A.S.; Song, H.-K. Cascaded learning empowered classification of UAVs using radio frequency under wireless interference. *Alex. Eng. J.* **2025**, *121*, 201–212. [\[CrossRef\]](#)
19. Delleji, T.; Slimeni, F.; Ayadi, A.; Salhi, R.; Siala, A. Real-Time Detection and Identification of mini-UAVs Using RF Signal Features and a Customized MLP Model. *J. Electr. Eng. Technol.* **2025**, *20*, 3585–3600. [\[CrossRef\]](#)
20. Khawaja, W.; Ezuma, M.; Semkin, V.; Erden, F.; Ozdemir, O.; Guvenc, I. A Survey on Detection, Classification, and Tracking of AAVs Using Radar and Communications Systems. *IEEE Commun. Surv. Tutor.* **2026**, *28*, 3272–3310. [\[CrossRef\]](#)
21. Shen, Y.; Yang, Y.; Liu, Y.; Wang, N.; Ma, L.; Wang, H.; Li, F. DFCT-net for automatic modulation recognition in UAV communication systems. *Neurocomputing* **2025**, *653*, 131200. [\[CrossRef\]](#)
22. Ahirrao, Y.V.; Yadav, R.P.; Kumar, S. RF-Based UAV Detection and Identification Enhanced by Machine Learning Approach. *IEEE Access* **2024**, *12*, 177735–177745. [\[CrossRef\]](#)
23. Li, M.; Hao, D.; Wang, J.; Wang, S.; Zhong, Z.; Zhao, Z. Intelligent Identification and Classification of Small UAV Remote Control Signals Based on Improved Yolov5-7.0. *IEEE Access* **2024**, *12*, 41688–41703. [\[CrossRef\]](#)
24. Zhao, R.; Li, T.; Li, Y.; Ruan, Y.; Zhang, R. Anchor-Free Multi-UAV Detection and Classification Using Spectrogram. *IEEE Internet Things J.* **2024**, *11*, 5259–5272. [\[CrossRef\]](#)
25. Allahham, M.H.D.S.; Al-Sa'd, M.F.; Al-Ali, A.; Mohamed, A.; Khattab, T.; Erbad, A. DroneRF dataset: A dataset of drones for RF-based detection, classification and identification. *Data Brief* **2019**, *26*, 104313. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015; p. 13.
27. Li, T.; Hong, Z.; Cai, Q.; Yu, L.; Wen, Z.; Yang, R. BisSiam: Bispectrum Siamese Network Based Contrastive Learning for UAV Anomaly Detection. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 12109–12124. [\[CrossRef\]](#)
28. Al-Sad, M.F.; Al-Ali, A.; Mohamed, A.; Khattab, T.; Erbad, A. RFased drone detection and identification using deep learning approaches: An initiative towards a large open source drone database. *Future Gener. Comput. Syst.* **2019**, *100*, 8697.
29. Al-Emadi, S.; Al-Senaïd, F. Drone detection approach based on radio-frequency using convolutional neural network. In Proceedings of the 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIOT), Doha, Qatar, 2–5 February 2020; p. 2934.
30. Allahham, M.S.; Khattab, T.; Mohamed, A. Deep learning for RF based drone detection and identification: A multi-channel 1-D convolutional neural networks approach. In Proceedings of the 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIOT), Doha, Qatar, 2–5 February 2020; p. 112117.
31. Huynh-The, T.; Pham, Q.-V.; Nguyen, T.-V.; Costa, D.B.D.; Kim, D.-S. RF-UAVNet: High-Performance Convolutional Network for RF-Based Drone Surveillance Systems. *IEEE Access* **2022**, *10*, 49696–49707. [\[CrossRef\]](#)
32. Gumaei, A.; Al-Rakhami, M.; Hassan, M.M.; Pace, P.; Alai, G.; Lin, K.; Fortino, G. Deep learning and blockchain with edge computing for 5G-enabled drone identification and flight mode detection. *IEEE Netw.* **2021**, *35*, 94–100. [\[CrossRef\]](#)
33. Ozturk, E.; Erden, F.; Guvenc, I. RF-based low-SNR classification of UAVs using convolutional neural networks. *arXiv* **2020**, arXiv:2009.05519.
34. Mandal, S.; Satija, U. Time–Frequency Multiscale Convolutional Neural Network for RF-Based Drone Detection and Identification. *IEEE Sens. Lett.* **2023**, *7*, 7003304. [\[CrossRef\]](#)
35. Mohaidat, T.; Niloy, S.I.; Khan, R.K.; Eda, J.; Niu, Z.; Sherif, A.; Khalil, K. RF-Based UAV Detection and Classification Using ANN Models. In Proceedings of the 2025 IEEE 4th International Conference on Computing and Machine Intelligence (ICMI), Mount Pleasant, MI, USA, 5–6 April 2025; pp. 1–5. [\[CrossRef\]](#)
36. Zuo, S.; Bai, X.; Zhang, C.; Liu, Y.; Wang, C.; Zhang, Y. RF-TCNet: A Lightweight Topology Compression Network for Drone RF Fingerprint Identification. *IEEE J. Radio Freq. Identif.* **2026**, *10*, 76–89. [\[CrossRef\]](#)
37. Fu, B.; Zhang, Y.; Wu, J.; Ye, F.; Zhang, W. An Efficient Dual-Branch Network and Multimodal Fusion Framework for Drone Identification. *IEEE Sens. J.* **2026**, *26*, 5241–5253. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.