SMNet: Multi-Drone Detection and Classification via Monitoring Flight Control Signals on Spectrograms

Ningning Yu, *Student Member, IEEE*, Jiajun Wu, Chengwei Zhou, *Senior Member, IEEE*, Zhiguo Shi, *Fellow, IEEE*, and Jiming Chen, *Fellow, IEEE*

***Abstract*—The widespread use of drones has provided numerous benefits, but it has also raised critical concerns regarding public safety and personal privacy due to many accidents stemming from their illegal use. Most radio frequency (RF) based methods for drone monitoring focus on the video-transmission signals (VTS), therefore bringing the potential risks that illicit drones can evade RF monitoring systems by disabling their VTS. To address this problem, we shift our attention to the flight control signals (FCS) that are indispensable for carrying the routine flight control instructions and logs. This paper proposes a new network designed specifically for FCS-oriented detection and classification, named the Spectrogram Monitoring Network (SMNet). Specifically, SMNet can identify the classes, frequency ranges, and time ranges of FCS by monitoring objects of interest on spectrograms. Different from traditional detectors, SMNet exploits the edge and corner features of FCS, which makes it perform well in complex electromagnetic environments. Then, we leverage our insights into texture features (TF) and position features (PF) found within FCS. By integrating TF-aware and PF-aware mechanisms, SMNet can effectively extract weak and scattered features of FCS to improve accuracy. Besides, we explore the application of SMNet to deal with multi-drone scenarios with a scale-limited dataset, which aims to reduce the demanded scale of data acquisition and labelling from the exponential level to the linear level. To this end, we utilize mixup-based data augmentation to generate pseudo-spectrograms that reflect multi-drone communication activities, thereby enhancing the generalizability of SMNet. Extensive experiments show that the proposed method outperforms the compared methods in terms of robustness against six typical types of noise while maintaining similar precision and recall.**

***Index Terms*—Anti-drone, flight control signal, multi-drone detection, signal detection and classification, spectrogram.**

I. Introduction

In recent years, drones (also known as Unmanned Aerial Vehicles, UAVs) have shown their great potential in agricultural irrigation, power inspection, geographical surveying and mapping, etc [1]–[5]. However, many accidents caused by misuse of drones, e.g., narcotics transportation, disruption to aviation, terrorist attacks, and privacy snooping, have also raised people's concerns about social security and personal privacy. Besides, drones are increasingly active in the military fields, playing key roles in intelligence gathering, position reconnaissance, and relay communication. Therefore, it is urgent to detect and classify drones that attempt to invade critical infrastructure, such as government offices, nuclear power plants, and hospitals.

There are mainly four sensing techniques for anti-drone, including vision, radar, acoustic, and radio frequency (RF) [6], [7]. Compared to the vision-based methods [8]–[10], acoustic-based methods [11]–[13], and radar-based methods [14]–[16], RF-based methods have the advantages of insensitivity to the non-line-of-sight scenarios and friendliness to the urban environment. Extensive works have adopted passive methods for drone detection and classification based on RF signals emitted by drones themselves [17]. In [18], Yan *et al.* performed data-free drone detection based on the periodicity of drone RF signals, which gets rid of a priori need for the data collection of drone signals. In [19], Zhang *et al.* proposed a drone identification method by using a convolutional neural network (CNN) and normalized cyclic prefix correlation spectrum (NCPCS), which is capable of effectively identifying drones at the low signal-to-noise ratio (SNR) regime. In [20], Xue *et al.* studied deep learning-based RF identification for drones with nonstandard waveforms, where a morphological-filtering-based carrier frequency offset estimation to address the problem of switching operating channels of drones is employed. To improve the classification efficiency, Cai *et al.* [21] proposed a low-complexity RF fingerprint identification (RFFI) method for drone identification, which uses lightweight multi-scale convolution (LMSC) blocks that can reduce the model size and enhance the feature extraction ability. The above-mentioned methods show that learning-based methods are popular and promising in RF-based drone detection and classification.

In recent works, RF-based drone detection and classification are regarded as a combined task consisting of object detection, localization, and classification on spectrograms. Basak *et al.* [22] was the first to propose a framework for combined multi-signal detection, localization, and classification. In their experiment, the You Only Look Once (YOLO)-lite based framework is proved to provide better performance compared

This work was supported in part by the National Natural Science Foundation of China under Grant U21A20456 and Grant 62401503, and in part by Zhejiang University Education Foundation Qizhen Scholar Foundation.

Ningning Yu, Jiajun Wu, Chengwei Zhou, and Zhiguo Shi are with the College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou 310027, China, and also with the Jinhua Institute of Zhejiang University, Jinhua 321037, China (e-mails: {nnyu, wujiajun, zhouchw, shizg}@zju.edu.cn).

Jiming Chen is with the State Key Laboratory of Industrial Control Technology, Zhejiang University, Hangzhou 310027, China, and also with the School of Artificial Intelligence, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: jmchen@ieee.org).

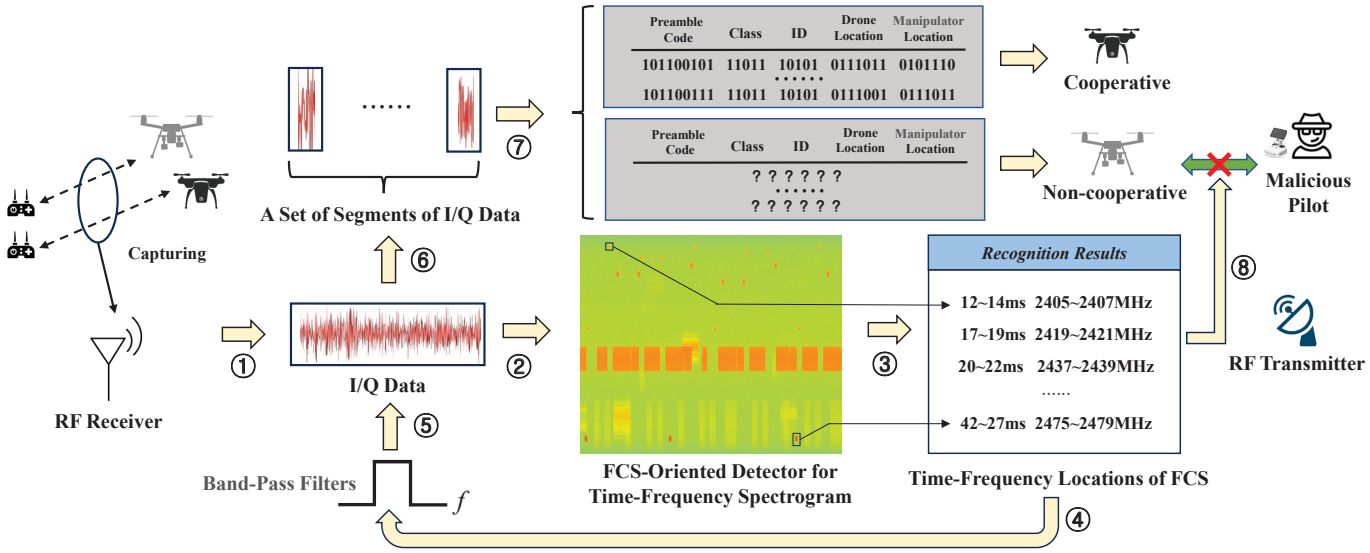


Fig. 1: Promising application of this work. The paradigm of RF-based drone surveillance consists of eight steps as follows. First, the RF receiver is deployed to collect signals from surrounding environments. Second, the captured I/Q data is transformed into time-frequency spectrograms. Third, an FCS-oriented detector is used to provide time and frequency ranges for each FCS. Fourth, a set of customized band-pass filters is set based on the frequency ranges of FCS. Fifth, the filters are applied to obtain the frequency bands of interest where the FCS are located. Sixth, the filtered signals are cut into several segments based on the time ranges of FCS. Seventh, the segments of signals are demodulated into binary codes based on the available protocols. Finally, targeted-frequency interference via an RF transmitter is implemented for the FCS that cannot be decoded.

to the Goodness-of-Fit (GoF) spectrum sensing and the Deep Residual Neural Network (DRNN). In [23], Zhao *et al.* proposed a keypoint-based anchor-free detector, which aims to address the problem of priori anchor mismatch. Besides, Zhao *et al.* also proposed an adversarial learning based data adaptation method, called spectrogram domain adaptation generative adversarial network (SDA-GAN), which effectively improves the cross-domain recognition performance of the detector towards a pair of practical scenarios.

However, the above-mentioned RF-based methods mainly focus on video-transmission signals (VTS) of drones, which brings potential risks that malicious drones can easily evade the RF surveillance system by consciously turning off VTS, called *Stealth* attacks [24]. To address this problem, we switch our attention to the flight control signals (FCS) of drones, which regularly carry flight logs and instructions between the drone and the ground controller, thus being more indispensable and reliable than the VTS.

It is worth pointing out that developing FCS-oriented detection and classification has the promising potential not only to efficiently decode identities for cooperative drones but also to guide intelligent countermeasures for non-cooperative drones (communication protocol unavailable), as shown in Fig. 1. Taking the urban low-altitude airway surveillance as a typical application, cooperative drones for airway transportation are strictly required to frequently report protocol-public broadcast signals (including information on identities and flight logs), in a similar form to FCS. The existing methods rely on searching the preamble code to determine the presence of the FCS in the captured In-phase/Quadrature (I/Q) data, which is unreliable under the condition of low SNR. As the FCS-oriented detection and classification can provide the frequency and time ranges of each FCS, it is efficient to focus on each FCS with customized bandpass filters to eliminate other interference. In this way, the decoding quality of cooperative drone signals can be improved. Besides, undecodeable FCS can be considered to be from illegal drones (non-cooperative drones). Based on the exact time-frequency positions of each FCS on spectrograms provided by FCS-oriented detection and classification, targeted-frequency electromagnetic countermeasures against drones are achievable, which will ensure the communication security of other wireless devices.

Due to the difference in RF characteristics between FCS and VTS, new challenges are introduced as follows. (a) **The sensitivity to noises.** The bandwidth of FCS is usually 1∼5MHz, and the time span is usually 0.5∼5ms. This results in the FCS appearing on the spectrogram accounting for 1∼3% of the whole, but the proportion of VTS is generally more than 10%. Therefore, noise interference can easily make the detector unable to focus on FCS. (b) **The variability in frequency.** It is relatively easy to detect and classify VTS because they always lie within a fixed frequency range unless the user manually switches the communication channel. However, the frequency of FCS keeps hopping over time, and its hopping range can reach up to 100MHz. (c) **The high costs of data labelling.** The FCS of same-brand drones are similar, leading to a huge risk in mislabelling. Especially when labelling a multi-drone RF dataset, professionals also need to carefully distinguish the identities of multiple FCS. To the best of our knowledge, there is currently no publicly published dataset for FCS-oriented multi-drone detection and classification. Therefore, insufficient training data will lead to a lack of generalization of the detector.

To address the above-mentioned challenges, we propose a

TABLE I: The main differences between this work and similar works.

Method	Signal Type	Number of Class	Multi-Drone Detection	Multi-Drone Dataset	Performance Evaluation
Proposed	FCS	15	Experiment	Unnecessary	Accuracy and Robustness
[19]	VTS	5	/	/	Accuracy and Robustness
[22]	FCS, VTS, and Wi-Fi	11	Simulation	/	Accuracy
[23]	VTS	5	Experiment	Necessary	Accuracy

FCS-oriented method for drone detection and classification, namely, Spectrogram Monitoring Network (SMNet). To improve the robustness, SMNet first preprocesses the original spectrograms combined with edge features and corner features that are noise-insensitive. Then, we analyze the texture features (TF) and the position features (PF) of FCS on spectrograms, therefore equipping SMNet with a TF-aware module and a PF-aware module to facilitate high-performance detection and classification. Specifically, the TF-aware module mainly provides the global perception of multi-scale feature maps, whereas the PF-aware module is designed with the self-attention mechanism to extract the global position characteristics of FCS, that is, the spatial correlation in the time-frequency domain, which is helpful to maintain noise-insensitivity while improving accuracy. Besides, combined with the mixup-based data augmentation, SMNet is further improved to deal with a data-constrained scenario where only a single-drone dataset is used for training while a multi-drone dataset is used for testing. Different from the conventional data augmentation techniques [20]–[22], this paper focuses on the matching of the proposed SMNet and the mixup, that is, the generalization performance gain brought by integrating the TF-aware and PF-aware mechanisms.

The main differences between this work and similar works are summarized in Table I. Different from the work [19] which evaluates the robustness only against the Gaussian noise, this work comprehensively evaluates the robustness against six typical noises, as well as more challenging mixed noises. Besides, although the work [22] studied FCS-based drone detection and identification, the proposed method has not been effective enough for the latest drones. It is because the newly-released drones improve communication anti-interference capabilities by using the FCS with smaller communication bandwidth, shorter communication duration, and more frequent transmission times, such as the DJI Mini series. In addition, this work evaluates a wider variety of drones than existing works to highlight the reliability of the proposed method. The main contributions of this paper are summarized as follows.

- As FCS become more vital in RF-based drone detection and classification, we propose a FCS-oriented method, namely SMNet, which integrates detection, localization, and classification of FCS on spectrograms.
- To improve the robustness and accuracy of SMNet, we combine the original spectrograms with edge and corner features that are insensitive to noises. Besides, we equip SMNet with a pair of TF-aware and PF-aware modules, which jointly facilitates feature extraction of FCS.
- By introducing the mixup-based data augmentation, SMNet is improved to detect and classify multi-drone test data while only single-drone data is available for learning. In this way, the scale of the training dataset for multi-drone detection and classification can be reduced from 2^D to D, where D is the number of drone classes.

The rest of this paper is organized as follows. In Section II, we analyze the key features of drone RF signals and present the problem formulation. In Section III, the proposed SMNet is elaborated. In Section IV, extensive experiments are provided in terms of accuracy, robustness, and generalizability comparisons. In Section V, we conclude this paper.

II. Signal Analysis and Problem Formulation

A. Drone RF Signal Analysis

In general, RF signals emitted by drones consist of FCS and VTS. Specifically, FCS carrying the instructions on flight control typically exist in the uplink channel, which are sent from the controller or mobile phones to the drones. Besides, FCS can also exist in the downlink channel to regularly report flight information such as position, speed, and altitude, which are sent from the drone to its controller. Compared to VTS, FCS are more reliable for anti-drone detection and classification, since the FCS of most commercial drones cannot be turned off by the user. In addition, VTS are easily terminated whereas FCS enable support communications in complex electromagnetic environments due to the unique modulation of FCS, as introduced as follows.

The FCS of most drones are modulated by frequency-hopping spread spectrum (FHSS), which can be denoted as

$$x_{\text{FCS}}(t) = c(t) \cdot A \sum_{k} W(t - kT_{\text{c}}) \cos\Big(2\pi f_k (t - kT_{\text{c}}) + \phi_k\Big), \tag{1}$$

where $t = 0, 1, \cdots, T-1$ with T being the length of sampling time, $k = 0, 1, \cdots, K-1$ with K being the number of frequency-hopping points (K usually ranges from 3 to 15 depending on the class of drones), $c(t)$ is the modulated data of control instructions or flight logs at the time slot t, A represents the modulated amplitude, $W(\cdot)$ is the rectangular window function, T_{c} is the frequency-hopping period, f_k is the center frequency, and ϕ_k is the initial phase. According to Eq. (1), the features of FCS can be divided into textural features (TF) and positional features (PF), which helps to clarify the subsequent design of the learning network for drone signals, i.e.,

$$\text{TF: } \{A, W(\cdot)\}, \tag{2}$$

$$\text{PF: } \{K, T_{\text{c}}, f_k, \Delta f_k\}, \tag{3}$$

where $\Delta f_k = |f_k - f_{k'}|$ with $k' \in \{0, 1, \cdots, K-1\}$ and $k \neq k'$. The TF refer to features with strong local correlations, whereas the PF refer to features with strong global correlations that own large shifts in both time and frequency domains.

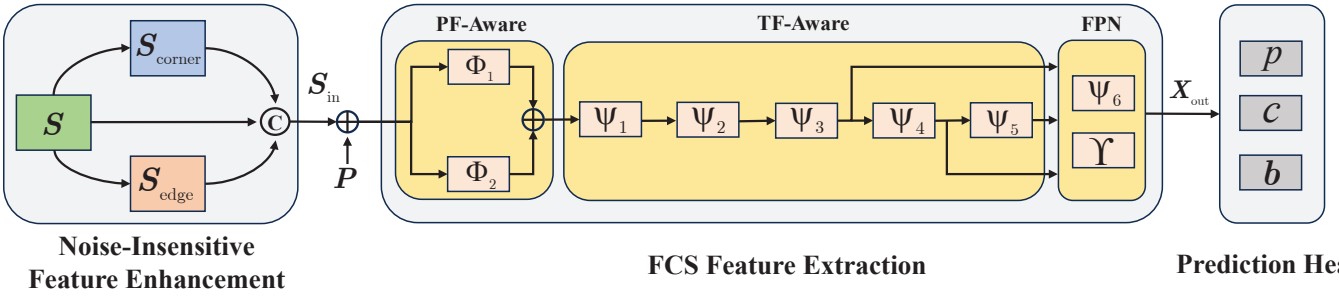


Fig. 2: Overview of the proposed SMNet, which consists of three key modules, namely, noise-insensitive feature enhancement, FCS feature extraction, and prediction head. Specifically, ⓒ denotes the operator of tensor concatenation, $\oplus$ denotes the operator of tensor summation, Φ denotes a self attention-based unit, Ψ denotes a standardized convolution-based unit, and Υ denotes the operator of enlarging size for tensors.

Considering a multi-drone scenario with the number of drones as D, the captured signals by the RF receiver can be denoted as

$$x(t) = \sum_{d=1}^{D} x_{\text{VTS}}^{d}(t) + \sum_{d=1}^{D} x_{\text{FCS}}^{d}(t) + x_{\text{IS}}(t), \tag{4}$$

where $x_{\text{VTS}}^{d}(t)$ and $x_{\text{FCS}}^{d}(t)$ denote the VTS and the FCS of the d-th drone, respectively. $x_{\text{IS}}(t)$ denotes the interference signals (IS) such as Bluetooth signals and Wi-Fi signals.

Since the components of the received signals are coupled in the time domain, the short-time Fourier transform (STFT) is employed to construct the time-frequency spectrogram to decouple the FCS with other signals, denoted as

$$\boldsymbol{S} = \begin{bmatrix} \boldsymbol{S}(0,0) & \boldsymbol{S}(0,1) & \cdots & \boldsymbol{S}(0,N-1)) \\ \boldsymbol{S}(1,0) & \boldsymbol{S}(1,1) & \cdots & \boldsymbol{S}(1,N-1) \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{S}(M-1,0) & \boldsymbol{S}(M-1,1) & \cdots & \boldsymbol{S}(M-1,N-1) \end{bmatrix}, \tag{5}$$

where N is the number of time frames, and M is the number of frequency points. Here,

$$\boldsymbol{S}(n,m) = \log_{10}\left| \sum_{t=0}^{M-1} x(t+nR)g(t) \cdot e^{\left(-\jmath 2\pi \frac{m}{M-1} t\right)} \right|, \tag{6}$$

where $\jmath$ represents the imaginary unit, R is the sliding length of the short-time analysis window, and $g(\cdot)$ is the Hamming window function. $|\cdot|$ indicates the modulus of the complex number.

B. Problem Formulation

In practical drone surveillance applications with complex electromagnetic environments, uncertain noise is inevitably introduced during the process of signal collection, spectrogram generation and conveyance. Therefore, a robust detector is required to detect all objects of interest, that is, a set of FCS, and provide their exact classes and time-frequency locations, expressed as

$$\left\{(\hat{c}_i, \hat{\boldsymbol{b}}_i)\right\}_{i=1,2,\ldots} = \mathcal{D}(\boldsymbol{S} + \boldsymbol{N}), \tag{7}$$

where $\mathcal{D}(\cdot)$ represents the detector, $\hat{c}_i$ is the predicted class of the i-th FCS, $\boldsymbol{N}$ is the noise matrix with the same size of $\boldsymbol{S}$, and $\hat{\boldsymbol{b}}_i = \left[\hat{b}_i^x, \hat{b}_i^y, \hat{b}_i^w, \hat{b}_i^h\right]$ is the bounding box of i-th FCS that contains the predicted coordinates (time frame and center frequency) and its height (frequency range) and width (time span), respectively.

III. Proposed Method: SMNet

The proposed SMNet consists of three key modules, i.e., noise-insensitive feature enhancement, FCS feature extraction, and prediction head, as shown in Fig. 2. In this section, we will elaborate on each module in detail.

A. Noise-Insensitive Feature Enhancement

The uncertain noise degrades the features of FCS, thereby making the detector hard to identify FCS. This work makes the detector free from the influence of noise by highlighting the features that are insensitive to noise, that is, edges and corners.

First, the gradients of the signal power with respect to the time domain and the frequency domain are respectively calculated by introducing two difference operators [25], which reflect the time and frequency frames of power changes in the spectrogram.

$$\boldsymbol{G}_{\text{T}} = \begin{bmatrix} 1 & 0 & 1 \\ 2 & 0 & -2 \\ 1 & 0 & -1 \end{bmatrix} * \boldsymbol{S}, \boldsymbol{G}_{\text{F}} = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * \boldsymbol{S}, \tag{8}$$

where $*$ denotes the operation of convolution. Then, the magnitude of the power gradient can be expressed as

$$\boldsymbol{G} = \sqrt{\boldsymbol{G}_{\text{T}}^2 + \boldsymbol{G}_{\text{F}}^2}. \tag{9}$$

To smooth the edges of FCS, the non-maximum suppression of $\boldsymbol{G}$ is expressed as

$$\hat{\boldsymbol{G}} = \begin{cases} \boldsymbol{G}(n,m), & \boldsymbol{G}(n,m) > \lambda_L \text{ and } \boldsymbol{G}(n,m) > \boldsymbol{G}(\delta_n, \delta_m), \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $\delta_n = n-\delta, n-\delta+1, \cdots, n+\delta$, $\delta_m = m-\delta, m-\delta+1, \cdots, m+\delta$, λ_L is a given lower-bounding threshold, and δ is a given value to determine the neighborhood. By introducing

an upper-bounding threshold λ_H, the edge spectrogram can be denoted as

$$\boldsymbol{S}_{\text{edge}}(n,m)=\begin{cases}1, & \hat{\boldsymbol{G}}(n,m)\geq\lambda_H,\\ 0.5, & \lambda_H>\hat{\boldsymbol{G}}(n,m)\geq\lambda_L,\\ 0, & \text{otherwise.}\end{cases}\tag{11}$$

The corner spectrogram can be denoted as [26]

$$\boldsymbol{S}_{\text{corner}}(n,m)=\begin{cases}1, & R(n,m)\geq\eta,\\ 0, & \text{otherwise,}\end{cases}\tag{12}$$

where η is a given threshold. $R(n,m)\in\boldsymbol{R}$ is the corner response derived as shown in Appendix A, denoted as

$$\boldsymbol{R}=ac-b^2-\gamma(a+c),\tag{13}$$

where $a=\boldsymbol{G}_{\text{T}}^2*\boldsymbol{W}$, $b=\boldsymbol{G}_{\text{T}}\boldsymbol{G}_{\text{F}}*\boldsymbol{W}$, $c=\boldsymbol{G}_{\text{F}}^2*\boldsymbol{W}$ with $\boldsymbol{W}$ being the Gaussian-weighted average window, and γ is a given threshold.

Finally, we combine the above-mentioned three spectrograms as the input of the detector, expressed as $\boldsymbol{S}_{\text{in}}=[\boldsymbol{S};\boldsymbol{S}_{\text{edge}};\boldsymbol{S}_{\text{corner}}]$, where $\boldsymbol{S}_{\text{in}}\in\mathbb{R}^{3\times N\times M}$.

B. FCS Feature Extraction

Existing detectors based on the TF-aware mechanism cannot perform well in FCS-oriented detection. On the one hand, the translation invariance of the TF-aware mechanism is useless for FCS-oriented detection, since FCS periodically appears at a specific set of frequencies according to a certain signal modulation. On the other hand, the TF mechanism is sensitive to noise, since the convolution operation is related to all inputs in the receptive field, even if one of the inputs has been perturbed by noise. As we summarized in Section II, PF of FCS are unique enough to distinguish drone classes, especially identifying same-brand drones with similar TF. Therefore, SMNet is specifically designed to combine the PF-aware mechanism and the TF-aware mechanism in tandem.

First, PF can be strengthened by adding a position encoding to the input, expressed as

$$\hat{\boldsymbol{S}}_{\text{in}}=\boldsymbol{S}_{\text{in}}+\boldsymbol{P},\tag{14}$$

where $\boldsymbol{P}$ is the position encoding as

$$\boldsymbol{P}(n,m)=\begin{cases}\sin(\omega_m n), & \text{if } m\%2=1,\\ \cos(\omega_m n), & \text{if } m\%2=0,\end{cases}\tag{15}$$

where $\%$ denotes the operation of calculating the remainder. ω_m is the hand-crafted frequency, and typically, $\omega_m=10000^{-\frac{2m}{M}}$.

Second, SMNet introduces the self-attention mechanism [27], [28] to construct the correlations of PF in the inputs. Since the positional correlations in both time and frequency domains are equally important, learning PF is performed simultaneously from these two domains, denoted as

$$\boldsymbol{X}_{512}=\Phi_1\big(\hat{\boldsymbol{S}}_{\text{in}}\big)+\Phi_2\big([\hat{\boldsymbol{S}}_{\text{in}}]^{\text{T}}\big),\tag{16}$$

where $\boldsymbol{X}_{512}$ indicates that the size of the output feature maps is 512×512, $[\,\cdot\,]^{\text{T}}$ denotes the matrix transpose. $\Phi_1(\cdot)$ and $\Phi_2(\cdot)$ are pair of self attention-based units, denoted as

$$\Phi(\boldsymbol{X}_{\text{in}})=\text{softmax}\left(\frac{\boldsymbol{X}_{\text{in}}\boldsymbol{Q}[\boldsymbol{K}]^{\text{T}}[\boldsymbol{X}_{\text{in}}]^{\text{T}}}{\sqrt{d_2}}\right)\boldsymbol{X}_{\text{in}}\boldsymbol{V},\tag{17}$$

where $\boldsymbol{X}_{\text{in}}$ is the input of $\Phi(\cdot)$, and $\text{softmax}(\cdot)$ denotes the SoftMax activation function. $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{d_1\times d_2}$ are learnable parameters, where d_1 and d_2 are the input dimension and the mapping dimension, respectively.

Third, we define a standardized convolution-based unit as $\Psi(\cdot)$, which sequentially contains a convolutional layer with the kernel size of 3×3, a batch normalization layer, and a max pooling layer with the stride of 2 to improve receptive fields. Since the features of FCS only account for a tiny proportion of the original feature map, it belongs to the concept of small target detection. Therefore, feature pyramid network (FPN) [29] is necessary to be introduced in SMNet, which can promote the recognition performance on small objects. Specifically, we first call different convolution-based modules in cascade to obtain three types of feature maps, namely, large 64×64, medium 32×32, and small 16×16 as

$$\begin{aligned}\boldsymbol{X}_{64}&=\Psi_3\Big(\Psi_2\big(\Psi_1(\boldsymbol{X}_{512})\big)\Big),\\ \boldsymbol{X}_{32}&=\Psi_4(\boldsymbol{X}_{64}),\\ \boldsymbol{X}_{16}&=\Psi_5(\boldsymbol{X}_{32}),\end{aligned}\tag{18}$$

where $\Psi_1(\cdot)$, $\Psi_2(\cdot)$, $\Psi_3(\cdot)$, $\Psi_4(\cdot)$, and $\Psi_5(\cdot)$ are convolution-based modules with different learnable parameters. Then, we define a size-enlargement operator as $\Upsilon(\cdot)$, which doubles the size of feature maps by padding the neighboring elements. In this way, the information of these three feature maps can be fused to obtain the fused feature maps as

$$\begin{aligned}\hat{\boldsymbol{X}}_{32}&=\boldsymbol{X}_{32}+\text{Conv}\big(\Upsilon(\boldsymbol{X}_{16})\big),\\ \hat{\boldsymbol{X}}_{64}&=\boldsymbol{X}_{64}+\text{Conv}\big(\Upsilon(\hat{\boldsymbol{X}}_{32})\big),\end{aligned}\tag{19}$$

where $\text{Conv}(\cdot)$ denotes as the learnable convolution layer with the kernel size of 1×1.

Finally, based on our investigation on the characteristics of most commercial drone signals, we found that the size of bounding boxes of FCS is mostly adapted to the medium feature map. Considering the delay-sensitivity in practical tasks, SMNet only retains the prediction output of the medium feature map which again integrates the information of the other two feature maps, denoted as

$$\boldsymbol{X}_{\text{out}}=\text{Conv}\big(\hat{\boldsymbol{X}}_{32}\big)+\Psi_6(\hat{\boldsymbol{X}}_{64})+\text{Conv}\big(\Upsilon(\boldsymbol{X}_{16})\big),\tag{20}$$

where $\Psi_6(\cdot)$ is a convolution-based module.

C. Prediction Head

In the prediction head, the output is obtain as $\hat{\boldsymbol{Y}}=\text{Conv}(\boldsymbol{X}_{\text{out}})$. Three kinds of information are provided for each grid in $\hat{\boldsymbol{Y}}$, namely, the confidence that FCS may exist, the bounding box of FCS, and the class of FCS. Accordingly, three loss functions are introduced to evaluate the prediction performance, i.e., confidence loss, offset loss, and classification loss.

Assuming that each map corresponds to B preset anchor boxes with a grid size of $G \times G$, the confidence loss is denoted as

$$\mathcal{L}_{\text{conf}} = -\sum_{i=1}^{G\times G}\sum_{j=1}^{B}\Big[\mathbb{I}_{i,j}^{\text{obj}} p_{i,j}\ln(score_i) + \mathbb{I}_{i,j}^{\text{noobj}}\ln(1-score_i)\Big], \tag{21}$$

where $score_i \in (0,1)$ is the predicted score about whether there is an object existing on the i-th grid. $\mathbb{I}$ is a mask vector. $\mathbb{I}_{i,j}^{\text{obj}} = 1$ when there exists an object on the j-th anchor box of the i-th grid, otherwise $\mathbb{I}_{i,j}^{\text{obj}} = 0$. On the contrary, $\mathbb{I}_{i,j}^{\text{noobj}} = 0$ when there exists an object on the j-th anchor box of the i-th grid, otherwise $\mathbb{I}_{i,j}^{\text{obj}} = 1$. $p_{i,j}$ denotes the similarity between the ground truth box and the anchor box, measured by

$$p_{i,j} = \frac{\min(b_i^w b_i^h, o_j^w o_j^h)}{\max(b_i^w b_i^h, o_j^w o_j^h)}, \tag{22}$$

where b_i^w and b_i^h are the width and the height of the object on the i-th grid, respectively. o_j^w and o_j^h are the width and the height of the j-th anchor box.

The offset loss measures the error between the predicted box and the ground truth box, denoted as

$$\mathcal{L}_{\text{offset}} = \sum_{i=1}^{G\times G}\sum_{j=1}^{B}\mathbb{I}_{ij}^{\text{obj}}||\boldsymbol{b}_i - \hat{\boldsymbol{b}}_i||_2^2, \tag{23}$$

where $\boldsymbol{b}_i = \left[b_i^x, b_i^y, b_i^w, b_i^h\right]$ denotes the time-frequency coordinates, width, and height of the ground truth box. $||\cdot||$ denotes the euclidean norm.

The classification loss employs the cross-entropy loss to compute the classification error of the predicted objects, denoted as

$$\mathcal{L}_{\text{class}} = -\sum_{i=1}^{G\times G}\sum_{j=1}^{B}\mathbb{I}_{ij}^{\text{obj}}\sum_{c=1}^{C}P(c_i)\ln\big(P(\hat{c}_i)\big) \tag{24}$$

where C is the number of classes, $P(\cdot)$ is the probability distribution function, and c_i is the ground truth class of the object on the i-th grid.

Finally, the total detection loss is

$$\mathcal{L}(\boldsymbol{Y}, \hat{\boldsymbol{Y}}) = \mathcal{L}_{\text{conf}} + \mathcal{L}_{\text{offset}} + \mathcal{L}_{\text{class}}, \tag{25}$$

where $\boldsymbol{Y} = \{(c_i, \boldsymbol{b}_i)\}_{i=1,2,\cdots}$ and $\hat{\boldsymbol{Y}} = \{(\hat{c}_i, \hat{\boldsymbol{b}}_i)\}_{i=1,2,\cdots}$ denote the set of the ground truth boxes and the predicted boxes, respectively.

D. Data-Constrained Multi-Drone Detection and Classification

In multi-drone scenarios, where the potential number of drone classes is denoted as D, an ideal dataset should encompass 2^D data classes. This principle is captured by the equation $2^D = \sum_{d=1}^{D}\frac{D!}{(D-d)!d!}$ where ! signifies the factorial, indicating that the invasions of different drone classes occurs independently. However, the explosive growth of drone classes renders it impractical to collect data for every conceivable combination of drones appearing simultaneously. This paper addresses a practical scenario where a multi-drone dataset is not available during training, aiming to reduce the need for extensive data acquisition and labelling. Despite this limitation, the proposed detector remains effective for detecting and classifying multiple drones using only a single-drone dataset. Consequently, the amount of required training data is significantly reduced from 2^D to D. However, the differences in sampling related to time, space, channel conditions, and the electromagnetic environment result in a distribution shift between the training and test sets, which leads to the problem of non-independent and identically distributed data (non-IID). To address the above-mentioned problem, the mixup-based [30] data augmentation is used in this work, as it aligns well with the FCS-oriented multi-drone detection and classification. First, compared to generative methods for data augmentation [31] [32], performing the mixup can preserve the features of FCS that make up a small proportion of spectrograms. Therefore, the non-IID problem can be addressed by enriching the training data with pseudo-spectrograms that are more similar to real-world multi-drones. Second, the increase in interference caused by the mixup can be effectively mitigated by the proposed SMNet, whose noise-insensitive mechanism offers a robust solution to further alleviate the non-IID issue.

In general, mixup can enrich the dataset by generating new training samples and their corresponding labels through linear interpolation in the classification task. For the combined task of object detection and classification, the new labels are obtained by integrating all ground truth boxes, which can be denoted as

$$\boldsymbol{S}' = \alpha\boldsymbol{S} + (1-\alpha)\tilde{\boldsymbol{S}}, \tag{26}$$

$$\boldsymbol{Y}' = \boldsymbol{Y} \cup \tilde{\boldsymbol{Y}}, \tag{27}$$

where $\alpha \in (0,1)$ is a weight parameter, $\boldsymbol{S}$ and $\tilde{\boldsymbol{S}}$ are the two real-world single-drone spectrograms, and their corresponding labels are $\boldsymbol{Y}$ and $\tilde{\boldsymbol{Y}}$.

E. Complexity Discussion

The time complexity of the proposed SMNet is analyzed as follows. In the noise-insensitive feature enhancement, the time complexity is $O\big((3|\boldsymbol{W}|^2+18)MN\big)$, where $|\boldsymbol{W}|$ is the window size of $\boldsymbol{W}$. In the PF-aware module, the time complexity is $O(M^2N + N^2M)$. In the TF-aware module, the time complexity is $O\big(\sum_{i=1}^{5}\frac{9}{4^{i-1}}MNC_{\text{in}}^{\Psi_i}C_{\text{out}}^{\Psi_i}\big)$, where $C_{\text{in}}^{\Psi_i}$ and $C_{\text{out}}^{\Psi_i}$ denote the number of input and output feature maps of $\Psi_i(\cdot)$, respectively. In the FPN, the time complexity is $O\big(\frac{9}{4^3}MNC_{\text{out}}^{\Psi_3}C_{\text{out}}^{\Psi_4} + \frac{1}{4^4}MNC_{\text{out}}^{\Psi_4}C_{\text{out}}^{\Psi_5} + \frac{1}{4^4}MNC_{\text{out}}^{\Psi_4}C_{\text{out}}^{\Psi_4}\big)$. Finally, in the prediction head, the time complexity is $O\big(\frac{1}{4^4}MNC_{\text{in}}^{\text{head}}B(D+5)\big)$.

IV. Experiment

A. Dataset

Based on the dataset *DroneRFb-Spectra* [33], this paper selects 15 typical classes of drones and annotates the boxes of FCS in their spectrograms. The dataset used in this paper is introduced in Table II. The experiment mainly includes two stages. The first stage performs single-drone detection and classification where both the training set and the test set include the drone classes numbered from A to L, and the

TABLE II: Introduction to the experimental dataset: classes, number of samples, and box sizes.

Index	Class	Average size of boxes		Number of samples
		Width (pixel)	Height (pixel)	
A	DJI Phantom 3	22	20	300
B	DJI Phantom 4 Pro	26	11	450
C	DJI MATRICE 200	26	10	501
D	DJI MATRICE 100	26	11	328
E	DJI Air 2S	8	15	237
F	DJI Mini 3 Pro	11	20	363
G	DJI Inspire 2	27	12	528
H	DJI Mavic Pro	9	11	245
I	DJI Mini 2	10	11	246
J	Vbar	22	10	306
K	Futaba T16IZ	19	21	338
L	DJI Phantom 4 Pro RTK	11	12	273
M	DJI Mavic 3 Pro	9	19	312
N	DJI Mini 2 SE	10	13	335
O	DJI Mini 4 Pro	9	32	349
P	DJI Mavic 3 Pro + DJI Mini 2 SE	/	/	148
Q	DJI Mavic 3 Pro + DJI Mini 4 Pro	/	/	246

† Due to different settings of sampling bandwidth, for class A to L, height: 1 pixel = 0.195MHz, while for class M to Q, height: 1 pixel = 0.156MHz. The sampling window is unified as 50ms, so for all classes, width: 1 pixel = 0.098ms.

TABLE III: Comparisons of recall and precision with respect to each class of drones.

Method / Class	Precision						Recall					
	Proposed	YOLOv3	YOLOv5	YOLOv7	CornerNet	CenterNet	Proposed	YOLOv3	YOLOv5	YOLOv7	CornerNet	CenterNet
DJI Phantom 3	0.8677	0.7409	0.8383	0.8701	0.8670	0.8644	0.9917	0.8582	0.9722	0.9976	0.9795	0.9582
DJI Phantom 4 Pro	0.7432	0.6606	0.6911	0.7491	0.7550	0.7590	0.9351	0.8295	0.8861	0.9804	0.9752	0.9488
DJI MATRICE 200	0.751	0.5911	0.6962	0.7632	0.7606	0.7370	0.9648	0.8083	0.9162	0.9908	0.9602	0.9416
DJI MATRICE 100	0.7612	0.6712	0.7003	0.7684	0.7673	0.7694	0.9414	0.8284	0.8938	0.9739	0.9504	0.9432
DJI Air 2S	0.9009	0.7330	0.7667	0.8595	0.8876	0.9021	0.9800	0.8192	0.8624	0.9594	0.9433	0.9680
DJI Mini 3 Pro	0.9014	0.6756	0.7409	0.8540	0.8794	0.8998	0.9937	0.7642	0.8252	0.9493	0.9322	0.9739
DJI Inspire 2	0.7495	0.6224	0.6911	0.7708	0.7566	0.7501	0.9287	0.8177	0.9143	0.9951	0.9379	0.8986
DJI Mavic Pro	0.8525	0.6364	0.6364	0.7441	0.8641	0.8895	0.9302	0.7071	0.7515	0.8482	0.8810	0.9440
DJI Mini 2	0.8579	0.6067	0.6551	0.7600	0.8498	0.9011	0.9434	0.7099	0.7462	0.8545	0.8968	0.9454
Vbar	0.8325	0.6456	0.7012	0.7984	0.8244	0.8268	0.9917	0.7838	0.8393	0.9561	0.9825	0.9601
Futaba T16IZ	0.7646	0.6728	0.7382	0.7647	0.7596	0.7548	1.0000	0.8679	0.9662	1.000	0.9966	0.9761
DJI Phantom 4 Pro RTK	0.8379	0.6643	0.6504	0.7628	0.8365	0.8732	0.9263	0.7171	0.7571	0.8702	0.9019	0.9224
Average	0.8184	0.6601	0.7088	0.7888	0.8173	**0.8273**	**0.9604**	0.7926	0.8609	0.9479	0.9448	0.9483

training set and the test set are randomly divided into the ratio of $8:2$. The second stage performs multi-drone detection and classification, where the training set uses single-drone classes numbered from M to O, and the test set uses multi-drone classes numbered from P to Q.

B. Compared Methods

We introduce five popular methods to compare with the proposed method, namely, YOLOv3 [34], YOLOv5 [35], YOLOv7 [36], CenterNet [37], and CornerNet [38]. Specifically, CenterNet and CornerNet belong to anchor-free approaches, while YOLO-series methods belong to anchor-based approaches. For the anchor-based approaches as well as the proposed SMNet, the preset anchors are set as the average size of boxes listed in Table II. The essential characteristics of compared methods are summarized as follows.

YOLOv3: An important feature of YOLOv3 is the multi-scale predictions, which makes it significantly better than previous versions of YOLO methods in both accuracy and efficiency. Specifically, the feature pyramid network (FPN) is used to fuse the feature maps output by the backbone network, and the head outputs predictions of large, medium, and small scales. In this way, the model can pay attention to objects of large, medium and small sizes at the same time. It is worth noting that the subsequent YOLO methods inherit the basic architecture of YOLOv3 including backbone (feature learning), neck (feature fusion), and head (prediction), and the proposed SMNet is also developed based on YOLOv3.

YOLOv5: The highlight of YOLOv5 is that it makes adaptive adjustments to the preset anchors based on the dataset, which is implemented by incorporating an Ultralytics algorithm called AutoAnchor. Besides, YOLOv5 employs the spatial pyramid pooling fast (SPPF) in the backbone, which helps to speed up the detection by combining multiple pooling layers of different sizes to learn multi-scale features.

YOLOv7: The main contributions of YOLOv7 are extended efficient layer aggregation network (E-ELAN), planned re-parameterized convolution, and coarse label assignment for auxiliary head.

CenterNet: The distinguishing feature of CenterNet is its improved head. Specifically, CenterNet constructs the heat map for the input through the backbone and further determines the center positions of objects based on the responses of the heat map. The center bias, lengths and widths of boxes are further optimized to develop an end-to-end and anchor-free detector.

CornerNet: Different from the CenterNet, CornerNet transforms object detection and classification into finding a pair

of key points, namely, the upper left corners and lower right corners of boxes.

C. Parameter Settings

We set the window size of $M = 2048$ for signal segments with the length of 50ms to implement the STFT. Furthermore, the size of spectrograms is compressed to 512×512 by downsampling. Like the conventional object detection and classification, we linearly normalize power values of the spectrograms into the range from 0 to 255. Corresponding to this normalized range, we set $\delta = 3$, $\lambda_H = 100$, $\lambda_L = 20$, $k = 0.05$, and $\eta = 0.01 \times \left(\boldsymbol{R}\right)_{\max}$, where $\left(\boldsymbol{R}\right)_{\max}$ denote the maximum of elements in $\boldsymbol{R}$. Note that the choice of δ, λ_H, λ_L, k, and η is not unique, which depends on the SNR of the received RF signals. In the dataset we used in this paper, which depicts a typical urban scenario including the interference from base stations and Wi-Fi devices, it is difficult to determine the best choice of the above-mentioned parameters.

To build the SMNet, we set the architecture parameters as follows: $M = N = 512$, $|\boldsymbol{W}| = 2$, $C_{\text{in}}^{\Psi_1} = 3$, $C_{\text{out}}^{\Psi_1} = C_{\text{in}}^{\Psi_2} = 32$, $C_{\text{out}}^{\Psi_2} = C_{\text{in}}^{\Psi_3} = 64$, $C_{\text{out}}^{\Psi_3} = C_{\text{in}}^{\Psi_4} = 128$, $C_{\text{out}}^{\Psi_4} = C_{\text{in}}^{\Psi_5} = 256$, $C_{\text{out}}^{\Psi_5} = 512$, $C_{\text{in}}^{\text{head}} = 768$, $D = 15$, and $B = 15$.

To perform the data augmentation based the mixup, we set $\alpha \in (0.3, 0.7)$ obeying the uniform distribution, which ensures the diversity of the synthesized spectrograms, especially simulating a low SNR situation when $\alpha \to 0.3$ or $\alpha \to 0.7$.

The proposed SMNet is deployed on a graphics processing unit (GPU) of GeForce RTX 3090 for training 250 epochs, equipped with the Stochastic Gradient Descent (SGD) optimizer with the learning rate of 1×10^{-5}.

Unless otherwise specified, the default threshold of intersection over union (IoU) between the ground truth box and the predicted box is 0.5 when performing comparison.

D. Accuracy Comparison

Two commonly-used indicators are introduced to evaluate the proposed SMNet, namely, precision and recall. Precision reflects the accuracy of positive predictions, and recall measures the proportion of actual positive instances that the model correctly identified [39].

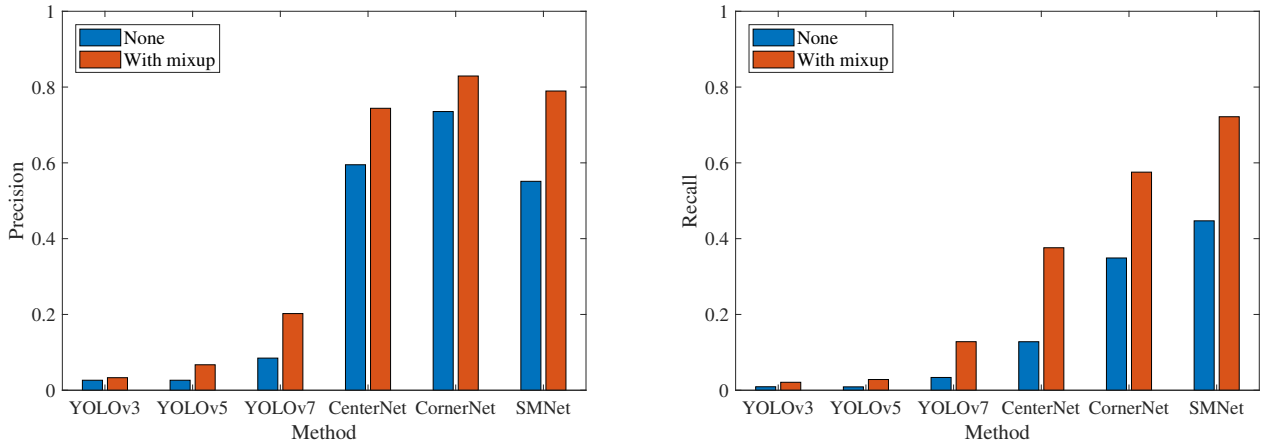
Single-drone: Table. III illustrates the accuracy of the proposed SMNet in comparison to the other methods, which shows that SMNet ranks first in terms of recall and second in terms of precision. This indicates that sequential integration of PF-aware and TF-aware mechanisms can achieve high-accuracy detection and classification. Besides, the precision gap between SMNet and CenterNet is mainly observed in the class DJI Mini 2, which has the smallest bounding boxes. Specifically, SMNet adopts the foundational architecture of YOLOv3 and extracts three feature maps with sizes of 16×16, 32×32, and 64×64, respectively. However, CenterNet focuses on the feature maps with the size of 128×128, which provides an advantage in detecting and classifying small-sized FCS.

Multi-drone: Fig. 3 shows the accuracy comparison for multi-drone detection and classification. In particular, we evaluate the applicability of the mixup-based data augmentation to the multi-drone scenario with a limited number of training data. First, it is obviously seen from Fig. 3 that YOLO-series methods fail in detecting FCS when there are two drones appearing on spectrograms, despite whether the FCS of each drone are trained. Even with the introduction of multi-drone pseudo-spectrograms generated by the mixup, the accuracy improvement of the YOLO-series methods is still poor. In contrast, anchor-free methods, i.e., CenterNet and CornerNet, own a relatively acceptable accuracy without employing the mixup. Besides, it is evident that the mixup-based data augmentation has significantly improved these three methods, namely, SMNet, CenterNet, and CornerNet. In particular, SMNet can perform best in term of recall with the introduction of the mixup. Finally, we focus on the required number of pseudo-spectrograms, which determines the number of training epochs for models, i.e., the overhead of training time. As shown in Fig. 4, when 250 pseudo-spectrograms (nearly one-quarter the scale of the single-drone dataset) are provided, there is an improvement in SMNet of approximately 20% in both precision and recall. In other words, SMNet employs the mixup-based data augmentation with an increase of the training time by 25% in exchange for an accuracy gain of nearly 20%. Besides, Fig. 4 indicates that simply increasing the number of pseudo-spectrograms does not yield a sustained improvement in accuracy. Notably, when the number surpasses 1750, precision and recall stabilize at 0.78 and 0.75, respectively. This finding emphasizes that the generated dataset, nearly twice the size of the original, is more than sufficient for optimizing the accuracy of SMNet.

E. Robustness Comparison

Six typical noises are introduced to verify the robustness of the proposed method, i.e., with the distribution of Gaussian, Impulse, Poisson, Rayleigh, Gamma, and Uniform, respectively. Each noise is rigorously tested at six distinct levels by fine-tuning its intensity or density. Specifically, Gaussian noise is set with different Gaussian distribution variances as $[0, 0.5, 1.0, 1.5, 2.0, 2.5]$. Impulse noise is set with different densities as $[0, 0.25, 0.5, 0.75, 1, 1.25] \times 10^{-4}$. Poisson noise is set with different densities as $[0, 0.25, 0.5, 0.75, 1, 1.25]$. Rayleigh noise is set with different intensities as $[0, 0.3, 0.6, 0.9, 1.2, 1.5]$. Gamma noise is set with different intensities as $[0, 0.3, 0.6, 0.9, 1.2, 1.5]$. Uniform noise is set with different intensities as $[0, 0.8, 1.6, 2.4, 3.2, 4]$.

We begin by assessing the robustness of the proposed method against a single type of noise. Fig. 5 and Fig. 6 demonstrate the robustness evaluation by respectively comparing precision and recall across varying noise levels, underscoring the advantages of our approach. It is obvious from Fig. 5 and Fig. 6 that YOLOv3 falls shortest in anti-noise. There are two reasons for this result. First, YOLOv3 lacks various multi-scale perception modules and nonlinear operations to suppress the noise interference on TF, which are supplemented in the E-ELAN of YOLOv7, the SPPF of YOLOv5, and the PF-aware module in SMNet. Second, YOLOv3 is essentially a detector that focuses on TF, which is more sensitive to noise compared to those detectors focusing on PF, e.g., CenterNet focusing on the PF of object centers, and CornerNet focusing on the PF



(a) Precision comparison

(b) Recall comparison

Fig. 3: Accuracy comparisons with the mixup.

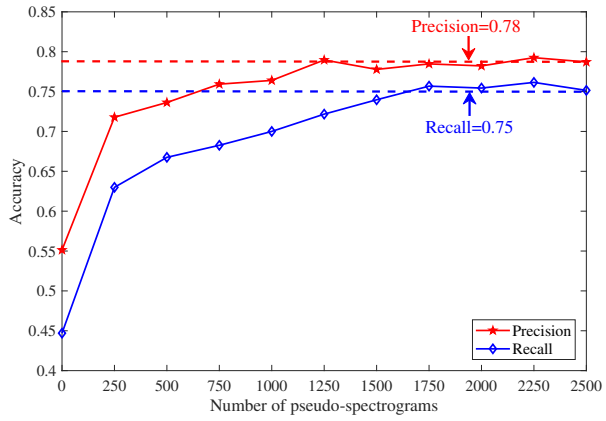


Fig. 4: Accuracy versus different number of pseudo-spectrograms.

of object corners. In Fig. 5, in addition to the capability of dealing with the impulse noise, the proposed SMNet achieves the highest precision, especially when the noise level is greater than 3. In Fig. 6, the proposed SMNet can achieve the best recall compared to other methods in all noise levels. These results indicate the effectiveness of the proposed PF-aware mechanism and the input spectrograms consisting of edge and corner features, both of which contribute to the noise-insensitivity performance of the proposed SMNet.

Then, we also consider the detection scenarios suffering from multiple mixed noises, therefore providing a robustness comparison under six mixed noises, as depicted in Fig. 7. To be fair, noise level of density and intensity is uniformly set as 5. In addition to the significantly higher accuracy than the comparative methods under the interference of mixed noises, we also discover a unique advantage of the proposed SMNet. By correlating Fig. 5, Fig 6, and Fig. 7, it becomes evident that the accuracy of SMNet is influenced primarily by the noise that has the most detrimental impact on it, remaining largely unaffected by other noise types. In other words, SMNet is insensitive to the number of noise types. Conversely, the accuracy of other comparative methods suffers considerably in the presence of mixed noises, indicating that their performance declines further as the number of noise types increases.

Last but not least, we summarize some common findings. First, the anchor-based methods are more sensitive to the impulse noise, while the anchor-free methods are more sensitive to the uniform noise. Second, as the noise level increases, the performance degradation of anchor-free methods is mainly reflected in the recall, and their performance degradation in precision is relatively slight.

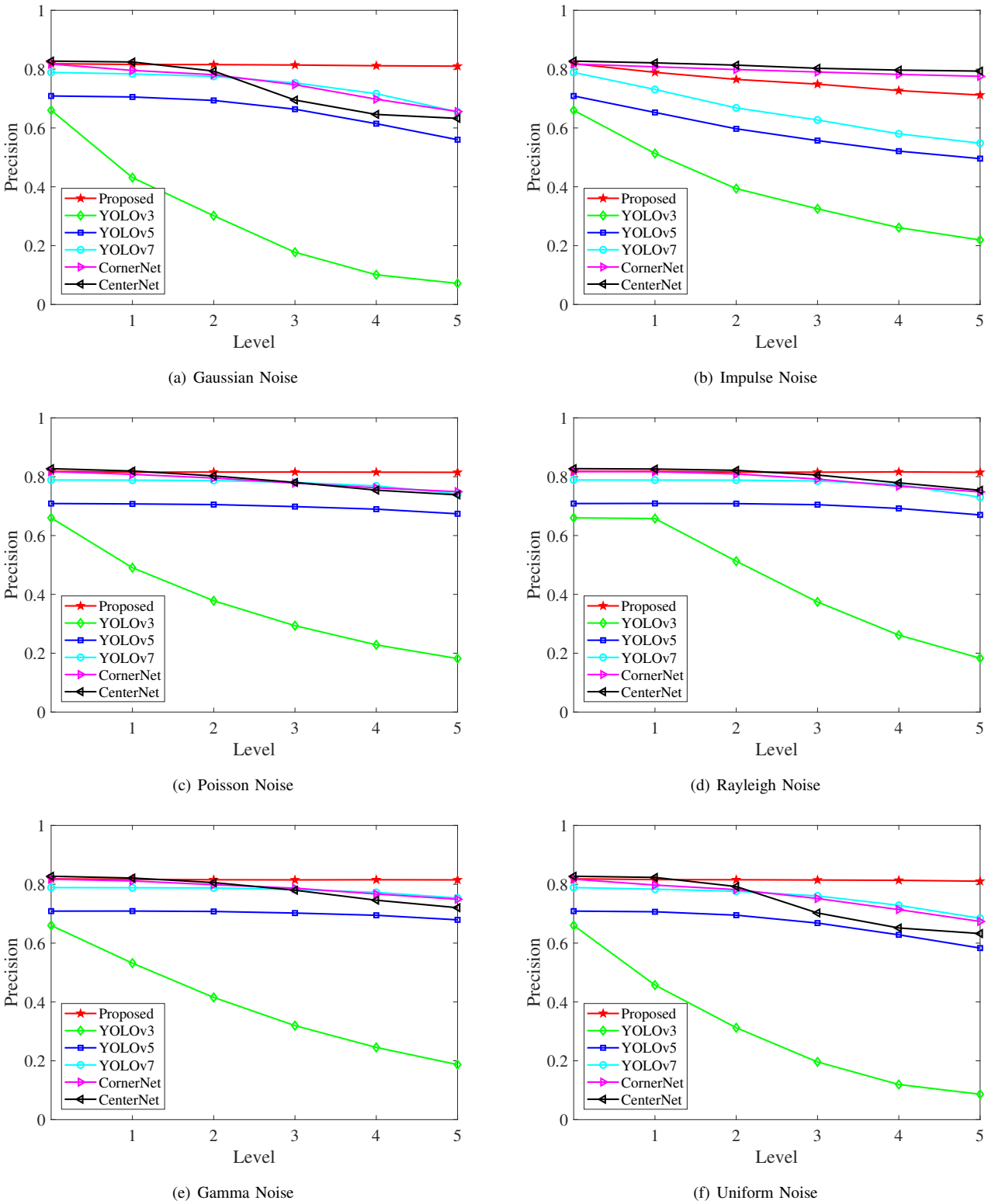
F. Other Comparison

While a default IoU threshold of 0.5 is commonly used, comparative experiments with various IoU thresholds are illustrated in Fig. 8. As the IoU threshold increases, SMNet shows a performance decline similar to that of the other two anchor-based methods. In contrast, CenterNet is less sensitive to the IoU threshold, which is suitable for tasks with high IoU requirements.

In addition, anchor-based methods are criticized for the determination of the number of preset anchors. Therefore, Fig. 9 indicates the precision comparison of the proposed SMNet versus different numbers of preset anchors. It is obvious from Fig. 9 that the precision of SMNet is not greatly affected by the reduction in the number of preset anchors, with the IoU threshold of 0.5. Therefore, it is feasible to appropriately reduce the number of anchors in exchange for the reduced computational complexity of SMNet in tasks with a lower requirement of IoU threshold. However, when the IoU threshold is set to 0.75, there is a substantial degradation in the precision of SMNet as the number of anchors is reduced.

G. Visualization

Fig. 10 illustrates the detection and classification results of single-drone scenarios. In Fig. 10(b), Fig. 10(d), and Fig. 10(i), even though the FCS have an SNR below 5dB, SMNet can still provide accurate prediction boxes. Besides, FCS detection by SMNet is reliable in the inevitable presence of interference

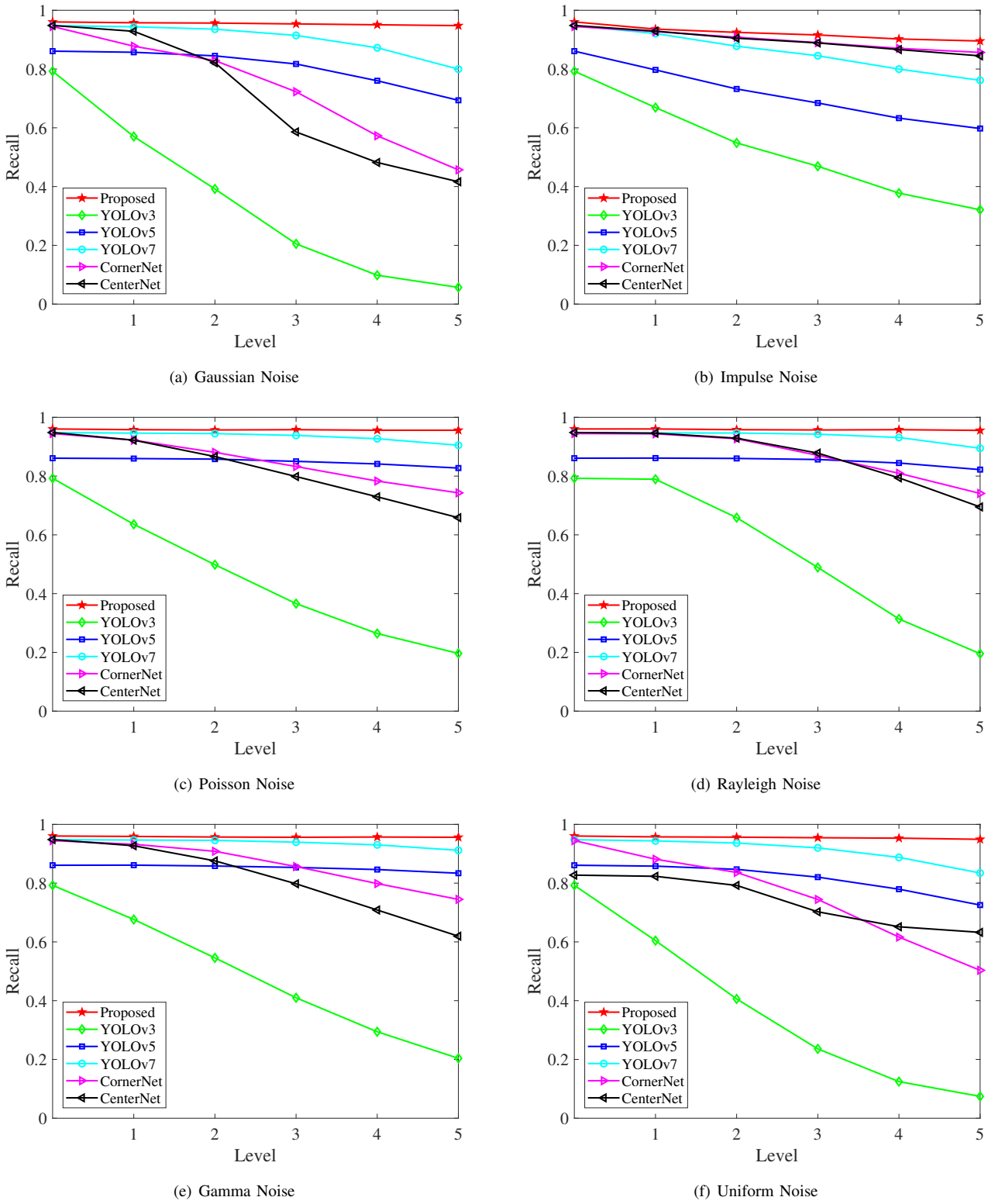


(a) Gaussian Noise (b) Impulse Noise

(c) Poisson Noise (d) Rayleigh Noise

(e) Gamma Noise (f) Uniform Noise

Fig. 5: Robustness evaluation by comparing precision versus different levels of noise intensity or density.

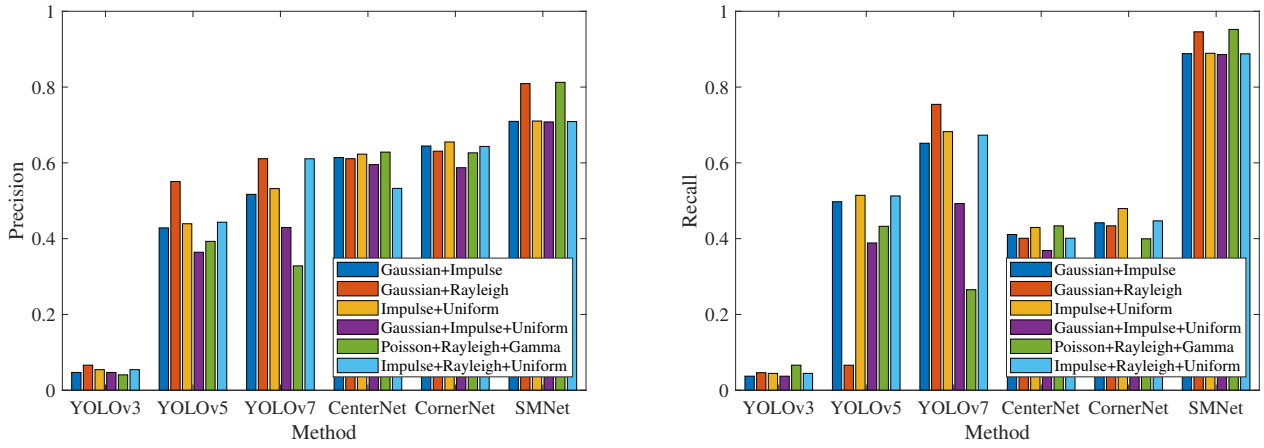


(a) Gaussian Noise (b) Impulse Noise

(c) Poisson Noise (d) Rayleigh Noise

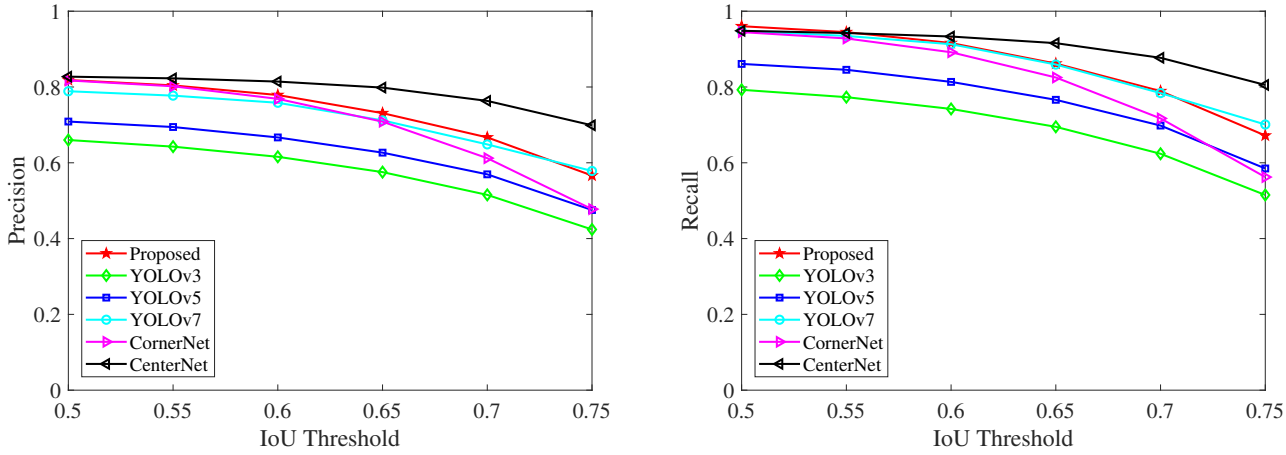
(e) Gamma Noise (f) Uniform Noise

Fig. 6: Robustness evaluation by comparing recall versus different levels of noise intensity or density.



(a) Precision comparison (b) Recall comparison

Fig. 7: Robustness comparison against six mixed noises.



(a) Precision comparison (b) Recall comparison

Fig. 8: Precision and recall comparisons versus different IoU thresholds.

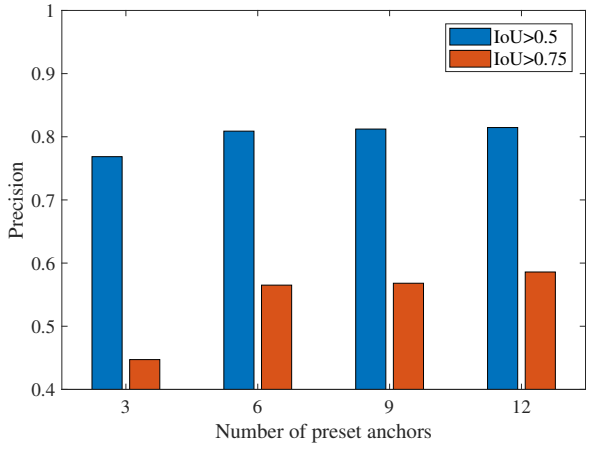
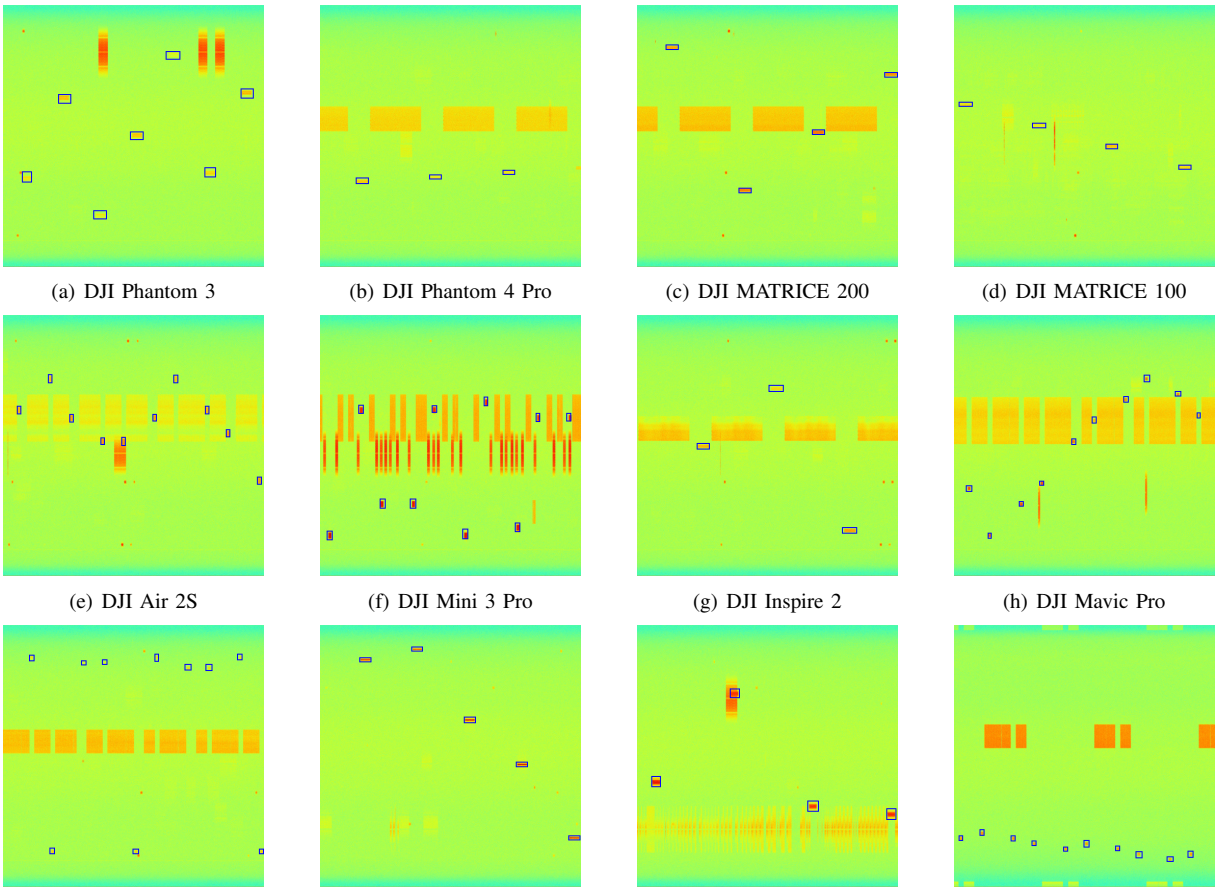


Fig. 9: Precision versus different number of preset anchors.

signals from Wi-Fi devices and base stations, as shown in Fig. 10(a), Fig. 10(f), Fig. 10(j), and Fig. 10(k). Furthermore, as indicated in Fig. 10(b), Fig. 10(c), Fig. 10(d), and Fig. 10(g), the TF of FCS are the same for DJI Phantom 4 Pro, DJI MATRICE 200, DJI MATRICE 100, and DJI Inspire 2, but their PF are significantly different. For example, the frequency shift (i.e., Δf_m) of DJI MATRICE 200 and DJI Inspire 2 is significantly larger than that of DJI Phantom 4 Pro and DJI MATRICE 100, which highlights the necessity of introducing the PF-aware mechanism.

Fig. 11 illustrates the detection and classification results in multi-drone scenarios. It should be noted that the presented spectrograms are sampled from real-world scenarios of multi-drone communication activities, which cannot be learned by SMNet during the training stage. By generating pseudo-spectrograms to simulate the data distribution of multi-drone scenarios, SMNet is also able to identify real-world test spectrograms rather than just pseudo-spectrograms. From Fig. 11(a), although DJI Mavic 3 Pro and DJI Mini 2 SE have similar FCS, SMNet can successfully recognize all FCS. In Fig. 11(b), most FCS can be detected by SMNet, except for a small amount of FCS from DJI Mavic 3 Pro. As the

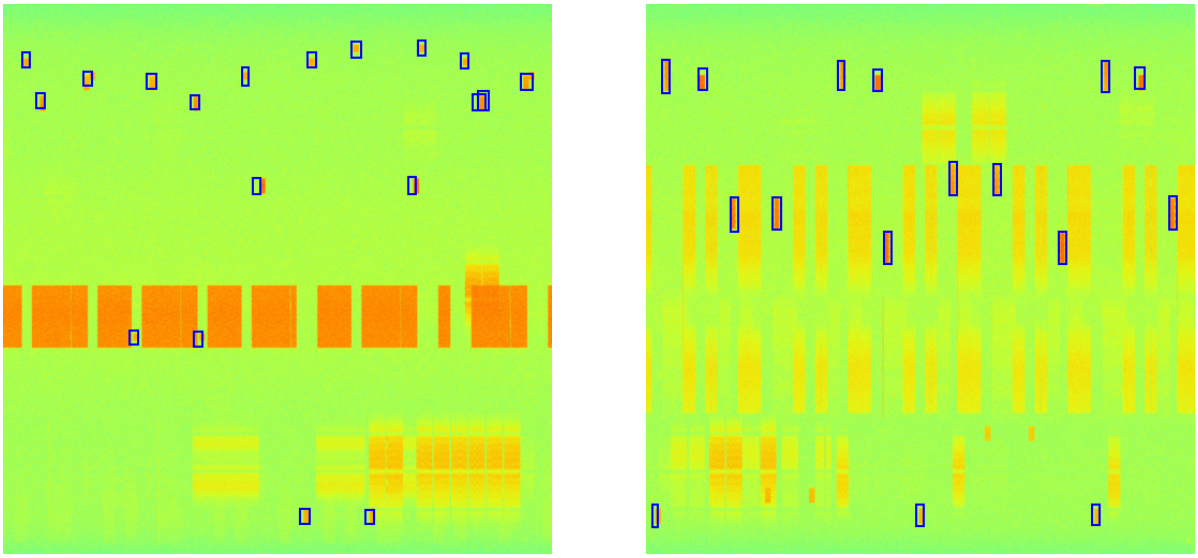


(a) DJI Phantom 3 (b) DJI Phantom 4 Pro (c) DJI MATRICE 200 (d) DJI MATRICE 100

(e) DJI Air 2S (f) DJI Mini 3 Pro (g) DJI Inspire 2 (h) DJI Mavic Pro

(i) DJI Mini 2 (j) Vbar (k) Futaba T16IZ (l) DJI Phantom 4 Pro RTK

Fig. 10: Visualization of single-drone detection and classification.



(a) DJI Mavic 3 Pro + DJI Mini 2 SE (b) DJI Mavic 3 Pro + DJI Mini 4 Pro

Fig. 11: Visualization of multi-drone detection and classification.

bandwidth of VTS of DJI Mini 4 Pro is 40MHz in Fig. 11(b), the spectrum resources are limited, so the FCS of DJI Mavic 3 Pro are easily affected by interference signals, resulting in missed detection.

V. Conclusion

This paper proposed a robust method for FCS-oriented detection and classification, named SMNet. Specifically, we enabled the SMNet to be noise-insensitive by combining the features of edges and corners with the input spectrograms. We equipped the SMNet with a pair of PF-aware and TF-aware modules to improve its detection and classification accuracy, based on the feature analysis of FCS on spectrograms. Particularly, SMNet is applied to a data-constrained scenario, where the generalizability of multi-drone detection and classification is improved by employing the mixup-based data augmentation. The accuracy, robustness, and generalizability of the SMNet are verified through extensive experiments compared with existing methods.

In future development, the proposed method will be applied to the urban low-altitude airway surveillance system, ensuring the safety of airway transportation by detecting, locating in time and frequency domains, and decoding drone signals. As the required technologies are shown in Fig. 1, designing RF transmitters with the FCS-oriented targeted-frequency suppression will be the focus of our subsequent research.

Appendix A
Deriving Corner Responses

Given an observation window $\boldsymbol{W}$ of an element $S(n,m) \in \boldsymbol{S}$, the power change can be quantified as

$$E(n,m) = \sum W(\Delta n, \Delta m)\big[S(n+\Delta n, m+\Delta m) - S(n,m)\big]^2, \tag{28}$$

where Δn and Δm are offsets with respect to the time domain and the frequency domain, respectively. $W(\Delta n, \Delta m) \in \boldsymbol{W}$ generally follows the Gaussian distribution. Based on the Taylor Formula, E can be approximately expressed as

$$E(n,m) \approx [\Delta n\ \Delta m] \underbrace{\begin{bmatrix} \sum \boldsymbol{W} * \left(\frac{\partial \boldsymbol{S}}{\partial n}\right)^2 & \sum \boldsymbol{W} * \left(\frac{\partial \boldsymbol{S}}{\partial n}\frac{\partial \boldsymbol{S}}{\partial m}\right) \\ \sum \boldsymbol{W} * \left(\frac{\partial \boldsymbol{S}}{\partial n}\frac{\partial \boldsymbol{S}}{\partial m}\right) & \sum \boldsymbol{W} * \left(\frac{\partial \boldsymbol{S}}{\partial m}\right)^2 \end{bmatrix}}_{\boldsymbol{\Lambda}} [\Delta n\ \Delta m]^{\mathrm{T}}. \tag{29}$$

Corners are different from edges, reflecting large power gradients in both time and frequency domains, which is equivalent to requiring the eigenvalue of $\boldsymbol{\Lambda}$ to be approximate and non-negative. Therefore, following the work [26], corner responses can be expressed as

$$R(n,m) = \det(\boldsymbol{\Lambda}) - \gamma[\mathrm{tr}(\boldsymbol{\Lambda})]^2, \tag{30}$$

where $\det(\boldsymbol{\Lambda})$ and $\mathrm{tr}(\boldsymbol{\Lambda})$ are the determinant and the trace of $\boldsymbol{\Lambda}$, respectively.

References

[1] W. Feng, Y. Lin, Y. Wang, J. Wang, Y. Chen, N. Ge, S. Jin, and H. Zhu, "Radio map-based cognitive satellite-UAV networks towards 6G on-demand coverage," *IEEE Trans. Cognit. Commun. Networking*, vol. 10, no. 3, pp. 1075–1089, 2024.

[2] Y. Liu, B. Zhang, D. Guo, H. Wang, and G. Ding, "Joint precoding design and location optimization in joint communication, sensing and computing of UAV systems," *IEEE Trans. Cognit. Commun. Networking*, vol. 10, no. 2, pp. 541–552, 2024.

[3] B. He, X. Ji, G. Li, and B. Cheng, "Key technologies and applications of UAVs in underground space: A review," *IEEE Trans. Cognit. Commun. Networking*, vol. 10, no. 3, pp. 1026–1049, 2024.

[4] A. Caruso, S. Chessa, S. Escolar, J. Barba, and J. C. López, "Collection of data with drones in precision agriculture: Analytical model and LoRa case study," *IEEE Internet Things J.*, vol. 8, no. 22, pp. 16 692–16 704, 2021.

[5] I. Bisio, C. Garibotto, H. Haleem, F. Lavagetto, and A. Sciarrone, "Traffic analysis through deep-learning-based image segmentation from UAV streaming," *IEEE Internet Things J.*, vol. 10, no. 7, pp. 6059–6073, 2023.

[6] X. Shi, C. Yang, W. Xie, C. Liang, Z. Shi, and J. Chen, "Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 68–74, 2018.

[7] Z. Zhang, Z. Shi, and Y. Gu, "Ziv-Zakai bound for DOAs estimation," *IEEE Trans. Signal Process.*, vol. 71, pp. 136–149, 2023.

[8] H. Wang, X. Wang, C. Zhou, W. Meng, and Z. Shi, "Low in resolution, high in precision: UAV detection with super-resolution and motion information extraction," in *Proc. IEEE ICASSP*, Rhodes Island, Greece, June 2023.

[9] X. Min, W. Zhou, R. Hu, Y. Wu, Y. Pang, and J. Yi, "Lwuavdet: A lightweight UAV object detection network on edge devices," *IEEE Internet Things J.*, vol. 11, no. 13, pp. 24 013–24 023, 2024.

[10] S. K. Mistry, S. Chatterjee, A. K. Verma, V. Jakhetiya, B. N. Subudhi, and S. Jaiswal, "Drone-vs-bird: Drone detection using YOLOv7 with CSRT tracker," in *Proc. IEEE ICASSP*, Rhodes Island, Greece, June 2023.

[11] Z. Shi, X. Chang, C. Yang, Z. Wu, and J. Wu, "An acoustic-based surveillance system for amateur drones detection and localization," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 2731–2739, 2020.

[12] Y. Sun, W. Wang, L. Mottola, J. Zhang, R. Wang, and Y. He, "Indoor drone localization and tracking based on acoustic inertial measurement," *IEEE Trans. Mobile Comput.*, vol. 23, no. 6, pp. 7537–7551, 2024.

[13] J. Kim, M. Y. Wang, and E. T. Matson, "Self-supervised drone detection using acoustic data," in *Proc. IEEE IRC*, Laguna Hills, CA, Dec. 2023, pp. 67–70.

[14] Y. Yang, F. Yang, L. Sun, T. Xiang, and P. Lv, "Echoformer: Transformer architecture based on radar echo characteristics for UAV detection," *IEEE Sensors J.*, vol. 23, no. 8, pp. 8639–8653, 2023.

[15] A. N. Sayed, O. M. Ramahi, and G. Shaker, "RDIwS: An efficient beamforming-based method for UAV detection and classification," *IEEE Sensors J.*, vol. 24, no. 9, pp. 15 230–15 240, 2024.

[16] X. Zhu, L. Tu, S. Zhou, and Z. Zhang, "Robust variability index CFAR detector based on Bayesian interference control," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, Art no. 5105809, 2022.

[17] H. Fu, S. Abeywickrama, L. Zhang, and C. Yuen, "Low-complexity portable passive drone surveillance via SDR-based signal processing," *IEEE Commun. Mag.*, vol. 56, no. 4, pp. 112–118, 2018.

[18] Y. Yan, N. Yu, C. Zhou, and Z. Shi, "A UAV detection method based on signal periodicity and its implementation on FPGA," in *Proc. IEEE/CIC ICCC*, Hangzhou, China, Aug. 2024, pp. 1098–1103.

[19] H. Zhang, T. Li, N. Su, D. Wei, Y. Li, and Z. Wen, "Drone identification based on normalized cyclic prefix correlation spectrum," *IEEE Trans. Cognit. Commun. Networking*, vol. 10, no. 4, pp. 1241–1252, 2024.

[20] C. Xue, T. Li, Y. Li, Y. Ruan, R. Zhang, and O. A. Dobre, "Radio-frequency identification for drones with nonstandard waveforms using deep learning," *IEEE Trans. Instrum. Meas.*, vol. 72, Art no. 5503713, 2023.

[21] Z. Cai, Y. Wang, Q. Jiang, G. Gui, and J. Sha, "Toward intelligent lightweight and efficient UAV identification with RF fingerprinting," *IEEE Internet Things J.*, vol. 11, no. 15, pp. 26 329–26 339, 2024.

[22] S. Basak, S. Rajendran, S. Pollin, and B. Scheers, "Combined RF-based drone detection and classification," *IEEE Trans. Cognit. Commun. Networking*, vol. 8, no. 1, pp. 111–120, 2022.

[23] R. Zhao, T. Li, Y. Li, Y. Ruan, and R. Zhang, "Anchor-free multi-UAV detection and classification using spectrogram," *IEEE Internet Things J.*, vol. 11, no. 3, pp. 5259–5272, 2024.
[24] I. Bisio, C. Garibotto, F. Lavagetto, A. Sciarrone, and S. Zappatore, "Blind detection: Advanced techniques for WiFi-based drone surveillance," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 938–946, 2019.
[25] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-8, no. 6, pp. 679–698, 1986.
[26] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vision Conference*, Manchester, UK, 1988.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in NIPS*, Long Beach, CA, Dec. 2017.
[28] P. Deng, S. Hong, J. Qi, L. Wang, and H. Sun, "A lightweight transformer-based approach of specific emitter identification for the automatic identification system," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2303–2317, 2023.
[29] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE CVPR*, Honolulu, HI, Jul. 2017, pp. 2117–2125.
[30] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," 2018. [Online]. Available: https://arxiv.org/abs/1710.09412.
[31] L. Liu, M. Muelly, J. Deng, T. Pfister, and L.-J. Li, "Generative modeling for small-data object detection," in *Proc. IEEE/CVF ICCV*, Seoul, Korea (South), Oct. 2019, pp. 6073–6081.
[32] I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. M. Botvinick, S. Mohamed, and A. Lerchner, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. ICLR*, Toulon, France, Apr. 2017.
[33] N. Yu, J. Wu, C. Zhou, Z. Shi, and J. Chen, "Open set learning for RF-based drone recognition via signal semantics," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 9894–9909, 2024.
[34] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018. [Online]. Available: https://arxiv.org/abs/1804.02767.
[35] G. Jocher, "YOLOv5 by Ultralytics," May 2020. [Online]. Available: https://github.com/ultralytics/yolov5.
[36] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," 2022. [Online]. Available: https://arxiv.org/abs/2207.02696.
[37] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," 2019. [Online]. Available: https://arxiv.org/abs/1904.07850.
[38] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," 2019. [Online]. Available: https://arxiv.org/abs/1808.01244.
[39] J. R. Terven, D. M. Cordova-Esparza, and J. A. Romero-González, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS." *Mach. Learn. Knowl. Extr.*, vol. 5, no. 4, pp. 1680–1716, 2023.

Ningning Yu (Student Member, IEEE) received the B.S. and M.S. degrees in Information Engineering from Zhejiang University of Technology, Hangzhou, China, in 2018 and 2021, respectively. He is now pursuing the Ph.D. degree in Electronic Science and Technology from Zhejiang University, Hangzhou, China. His research interests include anti-drone technology, spectrum recognition, and machine learning.

Jiajun Wu received the B.S. degree in Electronic Science and Technology from Zhejiang University, Hangzhou, China, in June 2024. He is now pursuing the Ph.D. degree in Electronic Science and Technology from Zhejiang University, Hangzhou, China. His research interests include anti-drone technology, signal processing, and pattern recognition.

Chengwei Zhou (Senior Member, IEEE) received the Ph.D. degree in Electronic Science and Technology from Zhejiang University, Hangzhou, China, in June 2018. He was a Visiting Researcher at University of Technology Sydney, Sydney, NSW, Australia, from April 2017 to October 2017. He served as a Postdoctoral Research Fellow at the State Key Laboratory of Industrial Control Technology, the College of Control Science and Engineering, Zhejiang University, Hangzhou, China, from 2018 to 2020. Since June 2020, he has been with the College of Information Science and Electronic Engineering, Zhejiang University, where he is currently a Tenure-Track Professor. His current research interests are in the areas of array signal processing, direction-of-arrival estimation, and adaptive beamforming.

Zhiguo Shi (Fellow, IEEE) received the B.S. and Ph.D. degrees in Electronic Engineering from Zhejiang University, Hangzhou, China, in 2001 and 2006, respectively. Since 2006, he has been a Faculty Member with the College of Information Science and Electronic Engineering, Zhejiang University, where he is currently a Full Professor. From 2011 to 2013, he visited the Broadband Communications Research Group, University of Waterloo, Waterloo, ON, Canada. His research interests include array signal processing, localization, and internet-of-things.

Prof. Shi was the recipient of the 2019 IET Communications Premium Award, and coauthored a paper that received the 2021 IEEE Signal Processing Society Young Author Best Paper Award. He was also the recipient of the Best Paper Award from ISAP 2020, IEEE GLOBECOM 2019, IEEE WCNC 2017, IEEE/CIC ICCC 2013, and IEEE WCNC 2013. He was the General Co-Chair of IEEE SAM 2020 and served as an Editor for the IEEE Network. He is currently serving as an Associate Editor for the IEEE Signal Processing Letters, IEEE Transactions on Vehicular Technology, and Journal of the Franklin Institute. He is an elected member of the Sensor Array and Multichannel (SAM) Technical Committee of the IEEE Signal Processing Society.

Jiming Chen (Fellow, IEEE) received the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2005. He is currently a Professor with the Department of Control Science and Engineering, and the Vice Dean with the Faculty of Information Technology, Zhejiang University. His research interests include IoT, networked control, and wireless networks. He serves on the editorial boards of multiple IEEE Transactions, and the General Co-Chairs for IEEE RTCSA'19, IEEE Datacom'19, and IEEE PST'20. He was a recipient of the 7th IEEE ComSoc Asia/Pacific Outstanding Paper Award, the JSPS Invitation Fellowship, and the IEEE ComSoc AP Outstanding Young Researcher Award. He is an IEEE VTS Distinguished Lecturer. He is a fellow of the CAA.