

Article

Acoustic Recognition of Unmanned Aerial Vehicles in a Gas Discharge Noise Environment Based on Adaptive Time–Frequency Denoising and Multichannel Feature Fusion

Chenkai Wang ¹, Zhiheng Zhang ², Weihua Kong ², Minxuan Zhong ² and Guoqing Yang ^{2,*}

¹ International Engineering College, Xi'an University of Technology, Xi'an 710048, China; 3232241006@stu.xaut.edu.cn

² School of Electrical Engineering, Xi'an University of Technology, Xi'an 710054, China; 2251921110@stu.xaut.edu.cn (Z.Z.); 3201712092@stu.xaut.edu.cn (W.K.); 3221712087@stu.xaut.edu.cn (M.Z.)

* Correspondence: yanggq@xaut.edu.cn

Featured Application

This work provides a cost-effective acoustic recognition method for protecting critical power infrastructure, including substations and overhead transmission lines, against unauthorized UAV intrusions.

Abstract

Background: The acoustic recognition accuracy of Unmanned Aerial Vehicles (UAVs) is severely degraded by strong gas discharge noise in power substations and overhead transmission lines. **Methods:** This paper proposes a UAV acoustic recognition method based on adaptive time–frequency denoising and multichannel feature fusion. The method integrates adaptive Wiener filtering with time–frequency masking techniques, employs a six-channel microphone array for collaborative noise reduction, and constructs a CNN-LSTM hybrid deep learning model for UAV type recognition. **Results:** Experimental results demonstrate that the two-stage denoising achieves a signal-to-noise ratio (SNR) improvement of 19.08 dB and a spectral fidelity of 0.89. Under extreme noise conditions at –10 dB SNR, the proposed method achieves 87.4% recognition accuracy, representing a 23.7% improvement over the traditional MFCC-SVM method and a 12.3% improvement over single-stage denoising strategies. In unknown environments, the recognition accuracy remains at 87–91%, exhibiting strong robustness and generalization ability. **Conclusions:** The proposed method effectively mitigates strong discharge noise interference and provides a reliable acoustic recognition solution for UAV intrusion detection in power facilities.

Keywords: UAV acoustic recognition; gas discharge noise; hierarchical denoising; adaptive Wiener filtering; time–frequency masking; CNN-LSTM



Academic Editor: John S. Allen

Received: 22 May 2026

Revised: 24 June 2026

Accepted: 25 June 2026

Published: 29 June 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 development of UAV technology, UAVs have been widely applied in power inspection, agricultural plant protection, logistics delivery, and other fields [1,2]. Unauthorized UAV intrusions into critical power facilities such as substations and overhead transmission lines occur frequently, posing serious threats to the safe operation of power grids [3]. Therefore, developing effective UAV monitoring and early-warning technologies is of significant practical importance.

Acoustic monitoring has become an important means of UAV detection due to its advantages of low cost, strong concealment, and immunity to lighting conditions [4,5]. In power-facility environments, gas discharge noise such as corona-discharge noise and spark discharge noise severely restrict the performance of acoustic monitoring systems [6]. This noise exhibits both wideband stationary noise and random impulsive characteristics, covering the main frequency bands of UAV rotor noise. Under extreme conditions, the SNR can be as low as -10 dB, leading to a dramatic decline in recognition accuracy of traditional methods.

Existing UAV acoustic recognition methods mainly include traditional Mel-frequency cepstral coefficient (MFCC)-based methods [7] and deep learning-based methods [8]. MFCC-SVM/HMM methods have been widely applied, but MFCC features show poor robustness in high-noise environments [9]. In terms of deep learning, CNNs can extract spatial features from mel-spectrograms [10,11], LSTMs can model temporal dynamics of acoustic signals [12], and CNN-LSTM hybrid architectures perform well in audio classification [13]. Attention mechanisms can automatically learn feature importance and improve discriminative ability [14]. However, existing research primarily targets clean or moderate noise conditions, and the performance under extremely low-SNR conditions still needs improvement [15].

In acoustic denoising, multichannel microphone arrays achieve noise reduction through spatial filtering [16,17], and multichannel averaging can provide SNR gain [18]. Wiener filters are optimal linear filters under the minimum mean-square error criterion [19,20], and time–frequency masking can effectively suppress high-energy noise components [21,22]. Single-stage denoising methods struggle to simultaneously handle wideband characteristics and random impulsive features [23]. Multimodal fusion can improve detection performance [24,25], but deployment complexity is relatively high [26,27].

However, robust UAV acoustic recognition in gas discharge noise environments remains insufficiently investigated. Unlike ordinary ambient noise, gas discharge noise around power facilities contains both broadband quasi-stationary components caused mainly by corona-discharge and intermittent impulsive components caused mainly by spark discharge. These noise components overlap with UAV rotor harmonics in the time–frequency domain and may severely mask the acoustic signatures used for recognition. Existing UAV acoustic recognition methods are usually evaluated under clean, ordinary environmental, or moderately noisy conditions, and their robustness under extremely low-SNR gas discharge interference is still limited. Therefore, the research gap addressed in this paper is the lack of a robust UAV acoustic recognition framework specifically designed for strong gas discharge noise environments around power facilities.

From an engineering perspective, this problem is relevant because unauthorized UAV intrusions near substations and overhead transmission lines may threaten the safe operation of power infrastructure. Acoustic monitoring provides a low-cost and passive detection solution, but its effectiveness depends strongly on whether the recognition model can resist gas discharge noise. This motivates the proposed adaptive time–frequency denoising and multichannel feature fusion framework.

To address these issues, this paper proposes a UAV acoustic recognition method that integrates adaptive time–frequency denoising with multichannel feature collaboration, suppressing gas discharge noise through hierarchical denoising to achieve high-precision recognition under extremely low-SNR conditions.

2. Materials and Methods

2.1. Gas Discharge Noise Generation Mechanism

The generation mechanism of gas discharge noise can be summarized into electromagnetic vibration, thermal expansion, and plasma oscillation mechanisms. Gas discharge in power facilities mainly includes two forms: Corona discharge and spark discharge. Corona-discharge generates ion wind propagating at subsonic speeds, producing continuous noise. Spark discharge generates shock waves propagating at supersonic speeds, producing impulsive noise.

To characterize the gas discharge noise considered in this study, corona-discharge noise was measured in a laboratory setup simulating the terminal of the same high-voltage conductor line under different applied voltages. Specifically, the noise spectra shown in Figure 1 were obtained under 8 kV, 14 kV, and 17 kV applied-voltage conditions. Therefore, Figure 1 is based on our own laboratory measurements rather than an external reference. The figure is used to illustrate the spectral characteristics of corona-discharge acoustic interference under different voltage levels and its relevance to UAV acoustic recognition near power facilities.

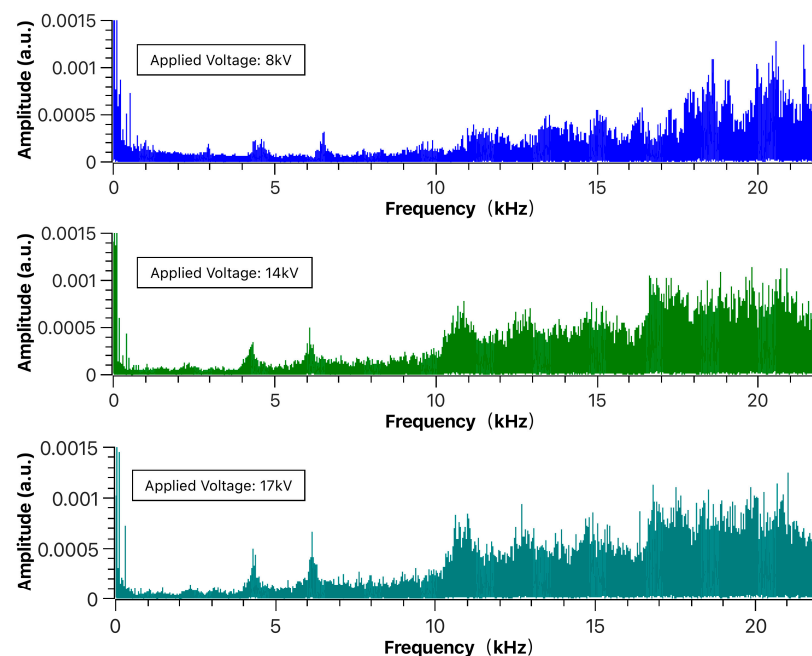


Figure 1. Measured spectra of corona-discharge noise.

As shown in Figure 1, the spectral distributions of corona-discharge noise differ significantly under different applied voltages. With the increase in applied voltage, the discharge state and acoustic radiation characteristics change, resulting in noticeable differences in spectral energy distribution. This indicates that gas discharge noise in power-facility environments is not a fixed background interference, but varies with operating voltage and discharge intensity. Such voltage-dependent spectral variation may overlap with UAV rotor acoustic components and mask useful recognition features. Therefore, single-stage time-domain or frequency-domain denoising methods may be insufficient, and a joint time–frequency analysis with hierarchical denoising is required.

Furthermore, the voltage-dependent spectral differences indicate that the discharge noise characteristics are closely related to the discharge state, which increases the variability of acoustic interference in practical power-facility environments.

2.2. Adaptive Wiener Filtering

Adaptive Wiener filtering is adopted as the first denoising stage because it provides an optimal linear estimation framework under the minimum mean-square error (MMSE) criterion when the signal and noise power spectra can be estimated [19,20]. In the gas discharge noise environment considered in this study, corona-discharge noise contains broadband quasi-stationary components that can be suppressed by estimating the noise power spectrum and applying a frequency-dependent Wiener gain. Compared with overly aggressive filtering, Wiener filtering can reduce background interference while preserving the harmonic structure of UAV rotor sounds. However, Wiener filtering alone is insufficient for intermittent impulsive components caused by spark discharge. Therefore, time–frequency masking is introduced as the second denoising stage to suppress localized high-energy impulses, forming a complementary two-stage denoising framework.

Step 1: Time–Frequency Transform

Short-time Fourier transform is performed on the i -th channel signal $x_i(t)$ to convert it to the time–frequency domain:

$$X_i(f, t) = \int_{-\infty}^{+\infty} x_i(\tau) w(\tau - t) e^{-j2\pi f\tau} d\tau \quad (1)$$

where τ is the integration dummy variable and $w(\cdot)$ is the Hanning window with a window length of 512 samples and hop length of 256 samples.

Step 2: Noise Power Spectrum Estimation

Assuming the first T_{frame} frames do not contain the target signal, the noise power spectral density estimation is

$$\hat{P}_n(f) = \frac{1}{T_{frame}} \sum_{k=0}^{T_{frame}-1} |X_i(f, k)|^2 \quad (2)$$

where T_{frame} denotes the number of initial frames used for noise power spectrum estimation and is set to 10 in this study.

Step 3: Signal Power Spectrum Estimation

The Minimum Controlled Recursive Averaging (MCRA) method is adopted to update the signal power spectral density:

$$\hat{P}_s(f, t) = \alpha_s \hat{P}_s(f, t-1) + (1 - \alpha_s) |X_i(f, t)|^2 \quad (3)$$

where α_s is the smoothing factor, set to 0.98.

Step 4: Wiener Gain Function Calculation

Based on the estimated power spectral density, calculate the time–frequency Wiener gain function:

$$H_i(f, t) = \frac{\hat{P}_s(f, t)}{\hat{P}_s(f, t) + \hat{P}_n(f)} \quad (4)$$

To avoid excessive attenuation, a gain lower bound constraint is introduced:

$$H_i(f, t) = \max\left(\frac{\hat{P}_s(f, t)}{\hat{P}_s(f, t) + \hat{P}_n(f)}, H_{min}\right) \quad (5)$$

where H_{min} is the gain lower bound threshold, and is set to 0.1 in this paper.

Step 5: Time–Frequency Filtering

Apply the Wiener gain function to the time–frequency representation:

$$Y_i(f, t) = X_i(f, t) \cdot H_i(f, t) \quad (6)$$

Step 6: Signal Reconstruction

Reconstruct the enhanced time-domain signal through inverse short-time Fourier transform (ISTFT):

$$x'_i(t) = \text{ISTFT}[Y_i(f, t)] \quad (7)$$

2.3. Time–Frequency Masking

Time–frequency masking is a time–frequency domain denoising method based on energy thresholds, leveraging energy distribution differences to achieve selective suppression [21]. Under gas discharge noise environments, reasonable energy thresholds can effectively identify and suppress high-energy discharge noise regions while preserving UAV target signal time–frequency components. This paper combines time–frequency masking with adaptive Wiener filtering to form a two-layer denoising architecture, achieving collaborative suppression of wideband background noise and impulse noise. The specific implementation steps are as follows:

First, perform STFT on the enhanced signal after Wiener filtering to convert it to the time–frequency domain:

$$X'_i(f, t) = \text{STFT}[x'_i(t)] \quad (8)$$

where $x'_i(t)$ denotes the time-domain signal of the i -th channel after adaptive Wiener filtering enhancement, and $X'_i(f, t)$ represents the time–frequency domain complex coefficients obtained by applying the short-time Fourier transform (STFT) to $x'_i(t)$. The STFT parameters remain consistent with those defined previously, utilizing a Hanning window with a window length of 512 points and a frame shift of 256 points, ensuring algorithmic consistency throughout the process.

Then, construct a binary time–frequency masking matrix based on logarithmic energy in the time–frequency domain:

$$M_{TF}(f, t) = I[\log|X'_i(f, t)| < T_{\text{energy}}(f)] \quad (9)$$

where $M_{TF}(f, t)$ is the binary time–frequency masking matrix; $I[\cdot]$ is the indicator function; $T_{\text{energy}}(f)$ is the frequency-dependent energy threshold.

Determine the energy threshold using the percentile method:

$$T_{\text{energy}}(f) = \text{percentile}(\log|X'_i(f, :)|, p) \quad (10)$$

where $X'_i(f, :)$ denotes the sequence of time–frequency coefficient magnitudes across all time frames at a fixed frequency f , and $\log|X'_i(f, :)|$ represents the corresponding one-dimensional sequence of logarithmic energy values for all frames at that frequency. The function $\text{percentile}(\cdot, p)$ is a percentile calculation operator used to determine the p -th percentile value of the sequence as the frequency-dependent energy threshold. In this paper, the percentile coefficient $p = 95$, verified as optimal through ablation experiments.

Finally, apply the constructed time–frequency masking matrix to the time–frequency domain coefficients:

$$x''_i(t) = \text{ISTFT}[X'_i(f, t) \cdot M_{TF}(f, t)] \quad (11)$$

where $X'_i(f, t) \cdot M_{TF}(f, t)$ denotes the element-wise masking operation on the time–frequency coefficients, and $x''_i(t)$ represents the final enhanced time-domain signal of the i -th channel after the two-stage denoising process incorporating Wiener filtering and

time–frequency masking. The inverse short-time Fourier transform (ISTFT) parameters are consistent with the STFT parameters defined previously, ensuring distortion-free reconstruction of the time-domain signal.

2.4. Multichannel Signal Processing

To improve UAV recognition signal-to-noise ratio (SNR) and accuracy, this study employs a six-element microphone array synchronous acquisition method.

Step 1: Multichannel Averaging Processing

Assume the signal received by each channel is

$$x_i(t) = s(t) + n_i(t), i = 1, 2, \dots, N \quad (12)$$

where $x_i(t)$ is the acquired time-domain noisy signal of the i -th channel, $s(t)$ is the UAV source signal, and $n_i(t)$ is the noise component of the i -th channel.

The signal after multichannel averaging is

$$x_{avg}(t) = \frac{1}{N} \cdot \sum_{i=1}^N x_i(t) = s(t) + \frac{1}{N} \cdot \sum_{i=1}^N n_i(t) \quad (13)$$

where $x_{avg}(t)$ is the averaged time-domain signal of the 6 microphone channels, and N is the number of microphone channels ($N = 6$ in this paper).

Step 2: Channel Synchronization Calibration

Due to distance differences between the six-microphone array and UAV acoustic source, the cross-correlation method is used to synchronize and calibrate the 6 channels:

$$\tau_i = \arg \max_{\tau} R_{0i}(\tau) = \arg \max_{\tau} \int x_0(t) \cdot x_i(t + \tau) dt \quad (14)$$

where τ_i is the time delay of the i -th channel relative to the reference channel (channel 0), $\arg \max_{\tau}(\cdot)$ denotes the function that finds the value of τ maximizing the argument, $R_{0i}(\tau)$ is the cross-correlation function between the reference channel signal and the i -th channel signal, $x_0(t)$ is the time-domain signal of the reference channel, and $x_i(t + \tau)$ is the shifted signal of the i -th channel.

Based on the calculated time delay τ_i , time alignment is performed on the signal of each channel, with the alignment formula expressed as

$$x_i(t) = x_i(t + \tau_i) \quad (15)$$

where $x_i(t)$ denotes the time-domain signal of the i -th channel after time alignment, and $x_i(t + \tau_i)$ on the right-hand side of the equation represents the original signal of the i -th channel shifted by τ_i along the time axis. Through this operation, all channel signals are synchronized with the reference channel, eliminating the effect of time delays caused by distance differences.

It should be noted that the six-channel configuration is an engineering design adopted in this study rather than a strict mathematical requirement for the recognition model. The CNN-LSTM classifier can operate on acoustic features extracted from fewer channels, but the multichannel array provides spatial diversity and SNR improvement through synchronization and averaging. For synchronized averaging with approximately uncorrelated channel noise, the theoretical array gain can be roughly estimated as $10 \log_{10}(N)$, where N is the number of microphones. Accordingly, three, four, six, and eight microphones correspond to approximate gains of 4.8 dB, 6.0 dB, 7.8 dB, and 9.0 dB, respectively. Increasing the number of microphones can improve spatial diversity and noise averaging, but it

also increases hardware cost, calibration difficulty, synchronization error accumulation, and deployment size. Conversely, using only one or two microphones reduces system complexity but loses most of the multichannel averaging gain and weakens robustness to local discharge noise and placement variation.

From a practical perspective, one microphone is sufficient only for single-channel acoustic classification after denoising, whereas at least three to four microphones are preferable for useful multichannel enhancement because they provide basic spatial diversity and a theoretical gain of approximately 4.8–6.0 dB. The six-microphone circular array used in this study provides about 1.8 dB additional theoretical gain over a four-microphone configuration while avoiding the limited marginal benefit and higher deployment burden of larger arrays such as eight or more microphones. The six-channel configuration was selected as an engineering compromise among denoising performance, synchronization reliability, hardware cost, array size, and field deployment convenience.

2.5. Mel-Spectrogram Calculation

The UAV noise samples after multichannel preprocessing remain time-domain signals, which cannot be directly utilized for feature learning in the UAV recognition model. Therefore, this paper employs the mel-spectrogram for feature extraction. The core objective is to transform time-domain signals into time–frequency representations that better describe the spectral and temporal characteristics of UAV acoustic signals, thereby emphasizing critical frequency information of UAV flight noise and improving robustness to noise interference [11,12]. The mel-spectrogram feature extraction process for sampled signals is described as follows.

The mel-spectrogram was selected because it preserves two-dimensional time–frequency structures while emphasizing perceptually and physically relevant low-frequency components. UAV rotor sounds usually contain fundamental frequencies, harmonic ridges, and temporal modulation patterns. Compared with MFCCs, which compress the spectral envelope into cepstral coefficients, mel-spectrograms retain richer local time–frequency textures for CNN-based spatial feature extraction and sequential patterns for LSTM-based temporal modeling. This representation is therefore suitable for the proposed CNN-LSTM framework, in which the CNN branch extracts local frequency-energy patterns and the LSTM branch models the temporal evolution of UAV acoustic features.

In addition, logarithmic compression and normalization are applied to the mel-spectrogram to reduce the influence of amplitude variation and improve the stability of model training under different SNR conditions. Therefore, mel-spectrogram features provide a balanced representation between acoustic interpretability, robustness, and compatibility with deep learning models.

Step 1: Mel Scale Conversion

The core formulas for converting between Mel scale and linear frequency are as follows:

$$\text{Mel}(f) = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \quad (16)$$

$$f = 700 \cdot \left(10^{\frac{\text{Mel}}{2595}} - 1\right) \quad (17)$$

where f denotes the linear frequency, and $\text{Mel}(f)$ represents the corresponding Mel scale value for frequency f .

Step 2: Mel Filter Bank Construction

This paper constructs M triangular filters on the Mel scale. First, the maximum Mel value corresponding to the highest frequency is calculated:

$$\text{Mel}(f_{\max}) = 2595 \cdot \log_{10}(1 + \frac{f_{\max}}{700}) \quad (18)$$

where $f_{\max}$ is the highest sampling frequency.

Subsequently, the Mel-scale center frequencies are converted to linear frequencies:

$$f_m = 700 \cdot (10^{\frac{\text{Mel}(m)}{2595}} - 1), m = 1, 2, \dots, M \quad (19)$$

where f_m is the linear center frequency of the m -th Mel filter, and $\text{Mel}(m)$ is its corresponding Mel scale value.

Next, the weight of the m -th filter at the k -th frequency bin is calculated:

$$H_m(f_k) = \begin{cases} \frac{f_k - f_{m-1}}{f_m - f_{m-1}}, & \text{if } f_{m-1} \leq f_k \leq f_m \\ \frac{f_{m+1} - f_k}{f_{m+1} - f_m}, & \text{if } f_m < f_k \leq f_{m+1} \\ 0, & \text{otherwise} \end{cases} \quad (20)$$

where f_k is the center frequency of the k -th frequency bin, and f_{m-1} and f_{m+1} are the linear center frequencies of the adjacent filters, respectively.

Step 3: Mel-Spectrogram Calculation and Preprocessing

This paper adopts a 2048-point Fast Fourier Transform (FFT) to compute the mel-spectrogram, perform logarithmic compression, and apply normalization. The steps are as follows:

First, the mel-spectrogram is calculated:

$$M(m, t) = \sum_{k=0}^{K-1} |X(f_k, t)|^2 \cdot H_m(f_k) \quad (21)$$

where $|X(f_k, t)|^2$ represents the energy spectrum of the k -th frequency bin at time t .

Second, logarithmic compression is performed using the following formula:

$$M_{\log}(m, t) = \log(1 + \gamma \cdot M(m, t)) \quad (22)$$

where $\gamma = 1$ is the compression coefficient, and $M_{\log}(m, t)$ is the compressed energy value.

Finally, the compressed features are normalized to eliminate the influence of different dimensions and improve the stability of model training. The normalization formula is

$$M_{\text{norm}}(m, t) = \frac{M_{\log}(m, t) - \mu}{\sigma} \quad (23)$$

where $M_{\log}(m, t)$ is the logarithmic compressed mel-spectrogram energy value, μ is the mean of all mel-spectrogram features in the training set, and σ is the standard deviation of all mel-spectrogram features in the training set. These statistics are obtained from the training set to ensure that the normalized features follow a distribution with zero mean and unit variance.

2.6. CNN-LSTM Hybrid Network

CNNs extract spatial features and bidirectional Long Short-Term Memory (Bi-LSTM) model temporal relationships. Hybrid architectures have demonstrated excellent performance in acoustic recognition [12,13]. Attention mechanisms can emphasize key features and improve discriminative ability [14]. This paper adopts a CNN-LSTM hybrid network for UAV noise identification based on mel-spectrogram features. The CNN branch

extracts spatial features from the mel-spectrogram, the LSTM branch models temporal features, and the attention mechanism enhances key features. Finally, the classifier completes the UAV type recognition. The specific workflow and core parameters are as follows:

2.6.1. CNN Branch

The CNN branch is responsible for extracting spatial features from the normalized mel-spectrogram $M_{norm}(m, t)$, such as frequency distribution and energy aggregation characteristics. The network structure is shown in Table 1. The parameters and functional descriptions of each layer are as follows:

Table 1. CNN network structure.

Layer	Input	Kernel	Output	Activation
Conv Block 1	$1 \times 128 \times 251$	$64 \times 3 \times 3$	$64 \times 64 \times 125$	ReLU
Max Pooling	$64 \times 64 \times 125$	2×2	$64 \times 64 \times 125$	-
Conv Block 2	$64 \times 64 \times 125$	$128 \times 3 \times 3$	$128 \times 32 \times 62$	ReLU
Max Pooling	$128 \times 32 \times 62$	2×2	$128 \times 32 \times 62$	-
Conv Block 3	$128 \times 32 \times 62$	$256 \times 3 \times 3$	$256 \times 16 \times 31$	ReLU
Max Pooling	$256 \times 16 \times 31$	2×2	$256 \times 16 \times 31$	-

As shown in Table 1, the input dimension “ $1 \times 128 \times 251$ ” corresponds to the mel-spectrogram feature dimensions, with 1 channel, 128 mel bins, and 251 time steps. The kernel parameter “ $64 \times 3 \times 3$ ” indicates 64 convolutional kernels of size 3×3 for extracting local spatial features. The ReLU activation function introduces non-linearity to enhance the network’s expressive ability. The max pooling layer preserves critical features, reduces dimensionality, and decreases computational complexity. The CNN branch finally outputs a $256 \times 16 \times 31$ spatial feature map, which is flattened and fed into the LSTM branch.

2.6.2. LSTM Branch

The LSTM branch is responsible for modeling the temporal correlation of mel-spectrogram features. A bidirectional LSTM structure is adopted, which can simultaneously capture forward and reverse temporal features of sequential signals. The two hidden layers enhance deep temporal feature extraction, and dropout layers are used to prevent overfitting. Finally, the temporal feature sequence at each time step is output to provide input for the attention mechanism. The specific parameters are shown in Table 2.

Table 2. Core LSTM parameters.

Key Parameter	Parameter Value/Description
Network Type	Bidirectional LSTM
Input Size	(batch, 31, 4096) (batch size $\times$ time steps $\times$ feature dimensions)
Hidden Layer Size	256 (number of neurons in a single LSTM unit)
Number of Layers	2 layers (enhances deep temporal feature extraction)
Dropout Coefficient	0.3 (prevents overfitting)
Output Size	(batch, 31, 512) (forward 256-dim + backward 256-dim)

As shown in Table 2, the LSTM branch receives the flattened features (4096 dimensions) output from the CNN branch, which are reorganized into a temporal feature sequence according to time steps. The input dimension is (batch, 31, 4096), where batch is the training batch size; 31 is the number of time steps, consistent with the time dimension of the mel-spectrogram; and 4096 is the feature dimension of each time step, obtained by flattening the 256×16 output of the CNN branch.

2.6.3. Attention Mechanism

An attention mechanism is introduced to assign different weights to the temporal feature sequence generated by the LSTM branch, thereby emphasizing time step features that are critical for UAV recognition and suppressing irrelevant features. The specific calculation process is as follows:

Step 1: Attention Weight Calculation

$$\alpha_t = \text{softmax}(W \cdot h_t + b) \quad (24)$$

where α_t is the attention weight at the t -th time step, satisfying $\sum_{t=1}^T \alpha_t = 1$. W is the weight matrix, b is the bias vector, h_t is the output feature of the LSTM branch at the t -th time step (512 dimensions); and the $\text{softmax}(\cdot)$ function normalizes the weights to the $[0, 1]$ interval to achieve weight assignment.

Step 2: Context Vector Calculation

$$c = \sum_{t=1}^T \alpha_t \cdot h_t \quad (25)$$

where c is the context vector output by the attention mechanism, representing the weighted sum of temporal features across all time steps. $T = 31$ is the number of time steps, consistent with the LSTM input time steps. $\alpha_t \cdot h_t$ is the weighted value of features at the t -th time step. The context vector c synthesizes key information from temporal features and is input to the subsequent classifier to complete recognition.

As described above, the attention weight of features at each time step is calculated to assign higher weights to time step features critical for UAV recognition while suppressing interference from irrelevant features. The context vector obtained through weighted summation achieves effective fusion of temporal features, improving subsequent classification accuracy.

2.6.4. Classifier and Training Parameters

The classifier adopts a fully connected layer structure, mapping the context vector c output from the attention mechanism to UAV type probabilities. The key parameters are shown in Table 3.

Table 3. Classifier core parameters.

Key Parameter	Parameter Value/Description
Network Structure	FC1(512→256) → FC2(256→128) → FC3(128→8)
Activation Function	FC1/FC2: ReLU; FC3: softmax (outputs class probabilities)
Dropout Coefficient	FC1: 0.5; FC2: 0.3 (prevents overfitting)

As described in Table 3, the classifier receives the 512-dimensional context vector output from the attention mechanism, and gradually compresses the feature dimension through 3 fully connected layers. The ReLU activation function maintains non-linear expression, and dropout layers further suppress overfitting. Finally, the softmax activation function maps features to the probability distribution of various UAV types, where the type with the maximum probability is recognized as the UAV.

During network training, the cross-entropy loss function is used to measure the difference between model predictions and true labels. Network parameters are updated through the Adam optimizer to minimize the loss function, ensuring model convergence and improving recognition accuracy. Reasonable learning rate and weight decay are set to balance training speed and model generalization ability.

2.7. Parameter Setting and Sensitivity Discussion

The proposed framework contains several key hyperparameters in the denoising, feature extraction, and classification stages. These parameters were fixed after offline selection and were not manually adjusted for individual test samples. Table 4 summarizes the main hyperparameter settings used in this study.

Table 4. Key hyperparameter settings used in the proposed framework.

Parameter	Value Used	Role
STFT window/hop	512/256	Time–frequency resolution
Initial noise frames	10	Noise PSD estimation
MCRA smoothing factor	0.98	PSD smoothing
Wiener gain floor	0.1	Avoid over-suppression
Masking percentile	95	Impulse suppression threshold
Mel FFT length	2048	Mel-spectrogram
Mel bins	128	Feature resolution
Dropout	0.3/0.5	Overfitting control

As shown in Table 4, these hyperparameters mainly affect the balance between noise suppression and UAV acoustic feature preservation. The denoising STFT window and hop length determine the tradeoff between frequency resolution and time localization. A larger window can improve harmonic resolution but may weaken the localization of impulsive discharge noise, whereas a smaller window can capture impulses more effectively but may reduce the resolution of UAV harmonic structures. The number of initial noise frames affects the stability of noise power spectral density estimation. Too few frames may lead to unstable estimation, while too many frames may include UAV components and cause over-suppression.

The MCRA smoothing factor controls the update speed of the signal power spectrum. A lower smoothing factor responds faster to noise variation but may distort stable UAV harmonic components, whereas a higher value improves stability but reduces adaptability. The Wiener gain floor is introduced to avoid excessive attenuation. If the gain floor is too low, useful UAV acoustic components may be removed; if it is too high, residual gas discharge noise may remain. The masking percentile controls the suppression of high-energy impulsive components. A lower threshold may remove useful UAV features, while a higher threshold may leave impulse interference.

For feature extraction, the 2048-point FFT and 128 mel bins provide sufficient frequency resolution while keeping the model input size manageable. The dropout rates in the LSTM and fully connected layers are used to reduce overfitting under the available dataset scale.

Overall, the selected parameter values were used as a fixed configuration to maintain stable performance under different SNR conditions without manual adjustment during testing.

After the offline parameter setting stage, the complete recognition pipeline operates automatically for each input sample. Specifically, the system automatically performs multichannel synchronization, noise power spectrum estimation, adaptive Wiener gain calculation, time–frequency masking, mel-spectrogram extraction, and CNN-LSTM classification. Therefore, the proposed method does not require manual parameter adjustment during the testing or inference stage. Nevertheless, the current implementation still uses fixed empirical parameters determined before deployment. In practical environments with rapidly changing discharge intensity or unseen noise types, fully adaptive threshold learning and online parameter optimization may further improve autonomy.

2.8. Overall Algorithm Flow

The overall implementation flow of the UAV acoustic recognition algorithm proposed in this paper is shown in Figure 2, which includes four main stages: signal acquisition and preprocessing, noise reduction processing, feature extraction, and model inference.

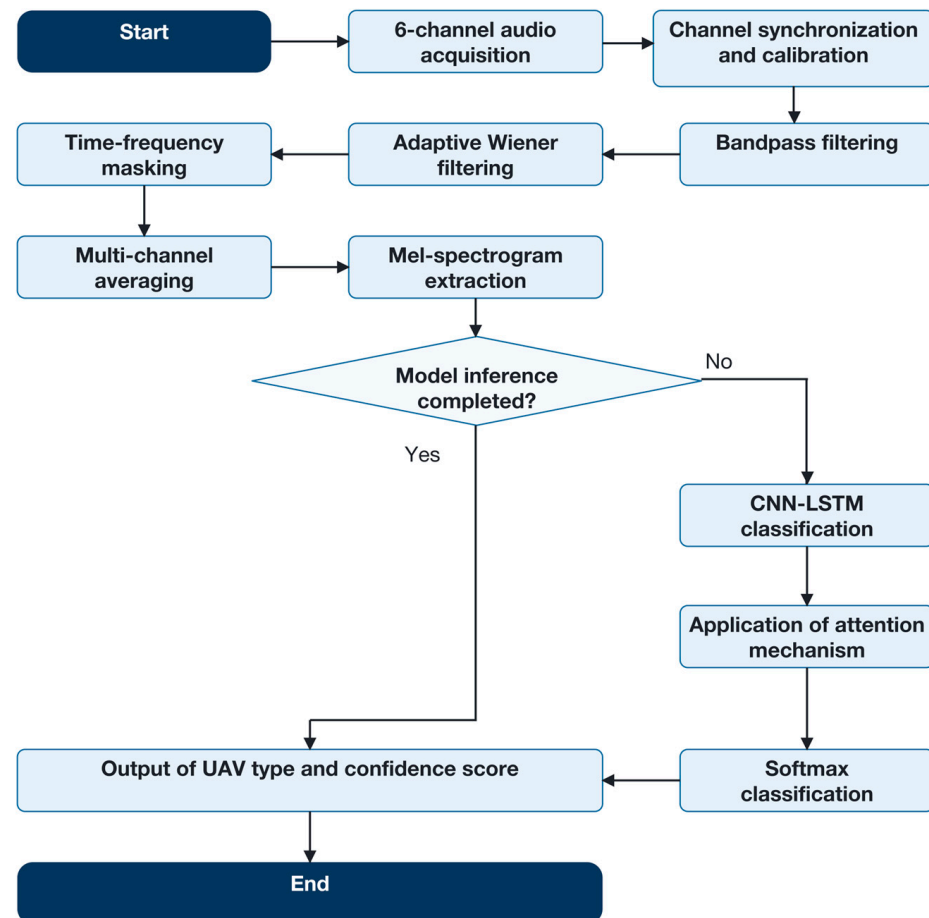


Figure 2. Audio signal processing flowchart.

As shown in Figure 2, the overall algorithm of the UAV acoustic recognition system proposed in this paper consists of five main stages: signal acquisition, signal preprocessing, noise reduction processing, feature extraction, and model inference. The Wiener filtering, time–frequency masking, and multichannel averaging methods can significantly improve the signal-to-noise ratio (SNR) of UAV audio signals, while the CNN-LSTM hybrid network can enhance the recognition accuracy under low-SNR conditions.

3. Experimental Results

3.1. Experimental Conditions

The experiments used six quadcopters, one hexacopter, and one octocopter for testing. Table 5 shows the structural parameters of each UAV. UAVs 1–6 were, in sequence, the Phantom 4, Mavic 2 Pro, Inspire 1, Faith 2, R500 RTK, and ZD550 quadrotor drones. Their manufacturers were DJI (Shenzhen, China), DJI (Shenzhen, China), DJI (Shenzhen, China), CFLYAI (Shenzhen, China), WALKERA (GuangZhou, China), and YH (Dongguan, China), respectively. UAV 7 was the LONGXIANG (Wuhan, China) F550 hexacopter drone, and UAV 8 was the DJI (Shenzhen, China) Spreading Wings S1000 octocopter drone.

Table 5. Structural parameters of UAVs.

Model	Rotor Number	Size
UAV 1	4	350 mm
UAV 2	4	355 mm
UAV 3	4	559 mm
UAV 4	4	280 mm
UAV 5	4	460 mm
UAV 6	4	552 mm
UAV 7	6	560 mm
UAV 8	8	1045 mm

The experiments employed a YD-6 six-channel microphone array (YUNDEA, Shenzhen, China) to collect UAV noise signals. The microphones were arranged in a circular array with a radius of 30 cm and element spacing of 15 cm. The software environment was built on Ubuntu 20.04 LTS with PyTorch 2.0.1 (CUDA 11.8) as the deep learning framework.

Three experimental scenarios are designed to evaluate the performance of the proposed method. Scenario 1 is a low-noise environment with background noise < 30 dB, used for collecting baseline data. Scenario 2 is a high-voltage laboratory, simulating power-facility environments to test anti-interference robustness. Scenario 3 is a mixed environment containing wind noise and other interferences, used to verify the recognition effectiveness under complex environmental conditions.

Therefore, the evaluation is not limited to low-noise or low-variability noise conditions. The low-noise scenario is used only as a baseline for comparison, whereas the high-voltage laboratory scenario is designed to represent gas discharge interference in power-facility environments. The mixed environment further introduces wind noise and other background interferences to examine the robustness of the method under more variable acoustic conditions. In addition, five SNR levels from +10 dB to −10 dB are used to simulate different noise intensities, including extremely low-SNR conditions.

3.2. Data Scale

For each UAV model, 100 audio clips were collected under each SNR condition, each clip lasting 5 s, with a total of 2400 samples. They were divided into a training set (1920 samples), validation set (240 samples), and test set (240 samples) at an 80%, 10%, 10% ratio. Five SNR conditions were set: −10, −5, 0, +5, +10 dB.

Although this dataset is sufficient for validating the proposed framework under the designed experimental conditions, we acknowledge that a larger dataset would further strengthen the generality of the conclusions. Future studies will include more UAV types, more flight states, longer recordings, and more real power-facility noise conditions to further evaluate the generalization ability of the proposed method.

3.3. Evaluation Metrics

Denoising effectiveness is comprehensively compared through SNR difference, spectral fidelity, and Perceptual Evaluation of Speech Quality (PESQ) scores. Recognition performance is evaluated through accuracy, precision, recall, F1 score, and confusion matrix. Five benchmark methods including MFCC-SVM, MFCC-HMM, CNN-ResNet, LSTM, and CRNN were selected as baseline methods.

3.4. Denoising Performance

3.4.1. SNR Improvement

Figure 3 presents the comparison of noise reduction effects under different SNR conditions. As shown in the figure, the proposed two-layer noise reduction method achieves an

SNR improvement of 12.3 dB, which is significantly higher than single-layer Wiener filtering (6.8 dB), time–frequency masking (7.5 dB), and band-pass filtering (3.2 dB). The spectral fidelity reaches 0.89, avoiding speech distortion caused by noise reduction; the PESQ score reaches 2.87, representing a 57.69% improvement compared to the original signal.

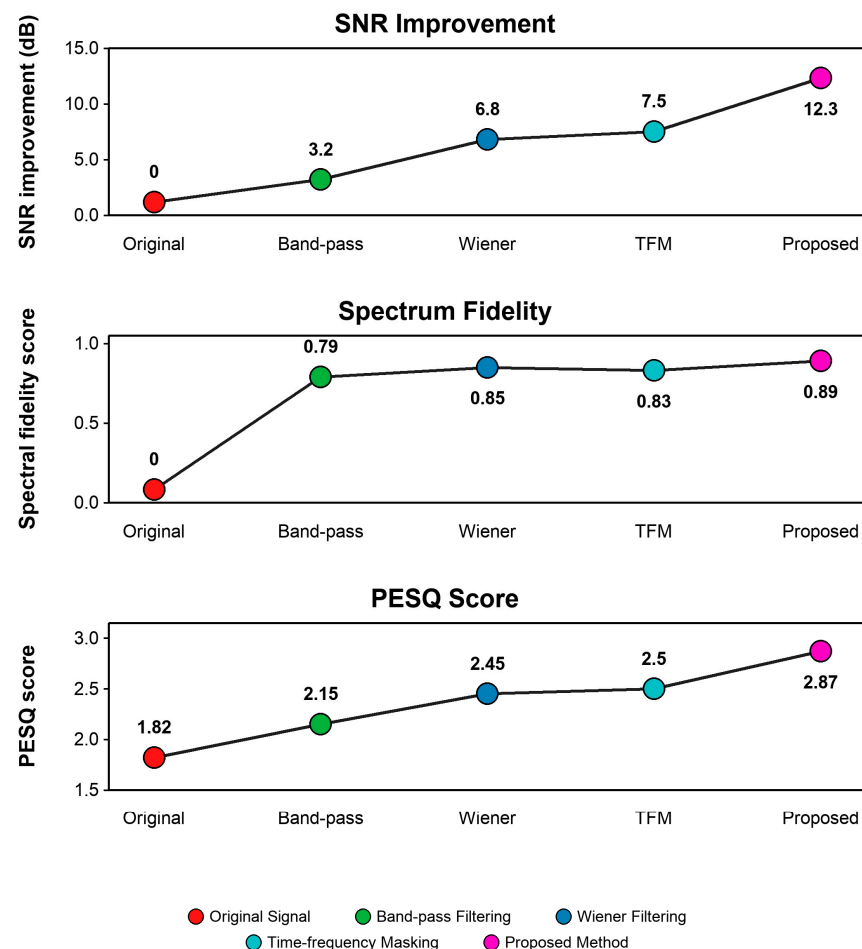


Figure 3. Comparison of noise reduction effects.

Figure 4 illustrates the progressive noise reduction process from the original signal contaminated by gas discharge noise to the final enhanced signal. (a) Original signal at -10 dB SNR, showing significant impulse interference and high background noise levels. (b) Adaptive Wiener filtering, in which impulse noise is partially suppressed and background noise is reduced. (c) Time–frequency masking processing, in which random impulses are effectively removed. (d) Two-layer noise reduction processing, in which the signal is further smoothed. Using the pure UAV signal as a reference, the proposed two-layer noise reduction method achieves a high impulse noise suppression ratio and signal detail retention rate.

As shown in Figures 3 and 4, the SNR improvement of two-layer noise reduction far exceeds single methods. Compared with band-pass filtering, Wiener filtering, and time–frequency masking, the SNR improvements are 9.1 dB, 5.5 dB, and 4.8 dB, respectively. The spectral fidelity is also the highest, with the PESQ score increasing from 1.82 to 2.87.

The proposed two-layer denoising method achieves a 12.3 dB SNR improvement, which is significantly higher than that achieved with single-layer Wiener filtering (6.8 dB), time–frequency masking (7.5 dB), and band-pass filtering (3.2 dB). Spectral fidelity reaches 0.89, avoiding speech distortion caused by denoising; the PESQ score reaches 2.87, a 57.69% improvement over the original signal.

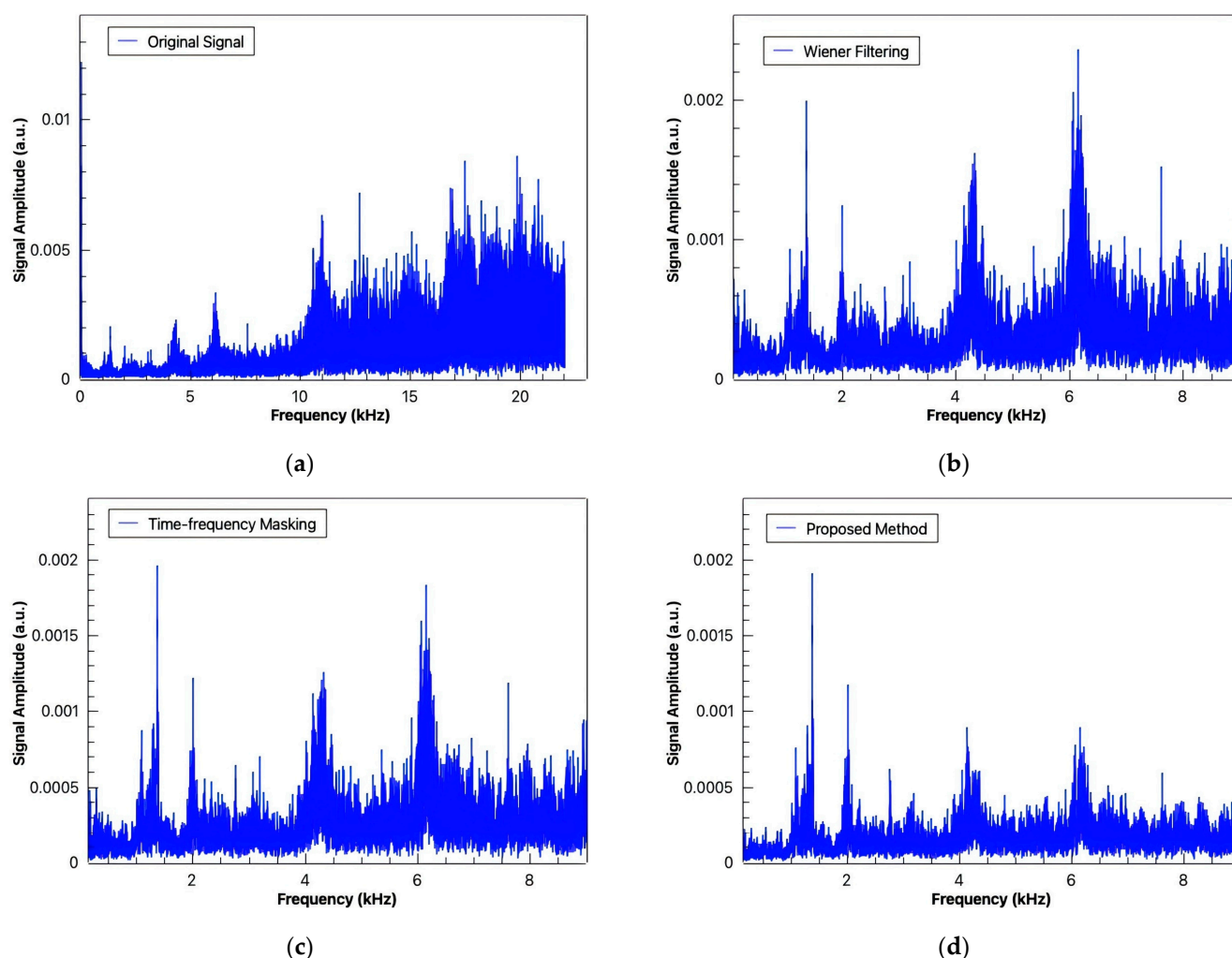


Figure 4. Signal processing results. (a) Raw signal with gas discharge noise; (b) adaptive Wiener filtering; (c) time–frequency masking; (d) two-stage denoising.

3.4.2. Analysis of the Noise Reduction Process

Table 6 shows the noise reduction process metrics. The results indicate that single non-linear noise reduction methods such as Wiener filtering and time–frequency masking cannot solve the problem of core feature loss of the signal. With increasing noise reduction intensity, the signal time-domain energy, spectral peak, and harmonic structure all exhibit varying degrees of loss. Time–frequency masking even causes damage to the harmonic structure, failing to simultaneously meet the dual requirements of deep noise reduction and high feature fidelity.

Table 6. Noise reduction process metrics.

Processing Stage	Time-Domain Energy Ratio	Spectral Peak Retention Rate	Harmonic Structure Integrity
Original Signal	1.000	100%	Complete
After Band-Pass Filtering	0.872	92%	Complete
After Wiener Filtering	0.654	88%	Complete
After Time–Frequency Masking	0.583	85%	Slight Damage
After Two-Layer Noise Reduction	0.567	90%	Restored

In UAV noise recognition scenarios, it is necessary to not only improve the signal SNR through deep noise reduction but also completely preserve core recognition features

such as signal spectrum and harmonics. The two-layer noise reduction method proposed in this paper adapts to this core requirement, providing high-quality input features for subsequent recognition models and effectively improving the accuracy and robustness of the recognition model.

3.5. Recognition Performance

3.5.1. Recognition Accuracy Under Different SNR Conditions

The recognition performance of the proposed method is evaluated under five SNR conditions ranging from +10 dB to −10 dB. Figure 5 presents the recognition accuracy data under different SNR conditions.

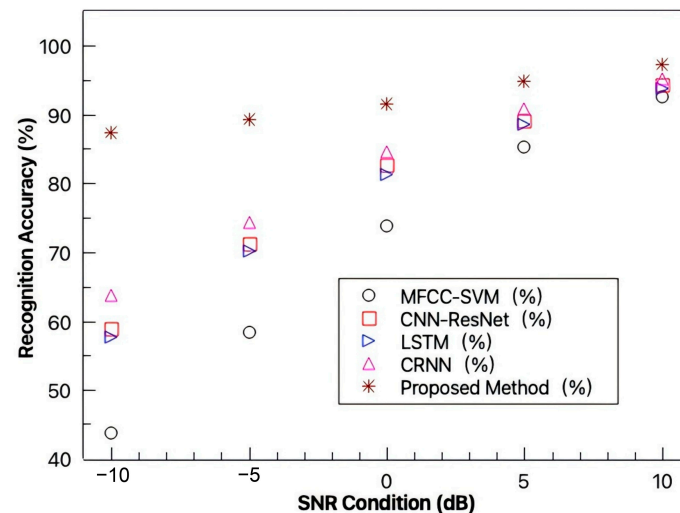


Figure 5. Comparison of recognition accuracy.

The data in Figure 5 shows that the proposed method achieves the best performance under all SNR conditions. Under ideal conditions at +10 dB, the accuracy of the proposed method reaches 97.2%, which is 2.1% higher than CRNN, the method with the highest recognition rate among other approaches. As SNR decreases, the performance of traditional methods and deep learning baselines drops sharply, while the proposed method maintains high accuracy. Under extremely low-SNR conditions at −10 dB, the accuracy of the proposed method is 87.4%, far exceeding other recognition methods. Compared with MFCC-SVM, it improves by 43.7 percentage points, and compared with CRNN, the method with the highest recognition rate, it improves by 23.7 percentage points.

3.5.2. Robustness Analysis Across SNR Levels

The recognition accuracy drop rate when the SNR condition changes from +10 dB to −10 dB is shown in Table 7. The data in Table 7 indicate that the drop rate of the proposed method is only 9.8%, which is lower than the 31–49% observed for other methods [12,15]. The above results demonstrate that the proposed method has good robustness in strong noise environments.

Table 7. Recognition accuracy drop rate.

Method	Drop Rate from +10 dB to −10 dB
MFCC-SVM	−48.8%
CNN-ResNet	−35.4%
LSTM	−36.1%
CRNN	−31.4%
Proposed Method	−9.8%

3.6. Ablation Study

To evaluate the performance contribution of each core component in the proposed method and verify the effectiveness and synergistic gain of the hybrid network architecture and each noise reduction module, systematic ablation experiments were conducted under extremely low-SNR conditions of -10 dB. Using UAV type recognition accuracy as the core evaluation metric, the experimental results are shown in Figure 6. Here, BPM denotes band-pass filtering method, WF denotes Wiener filtering method, and TFM denotes time–frequency masking method.

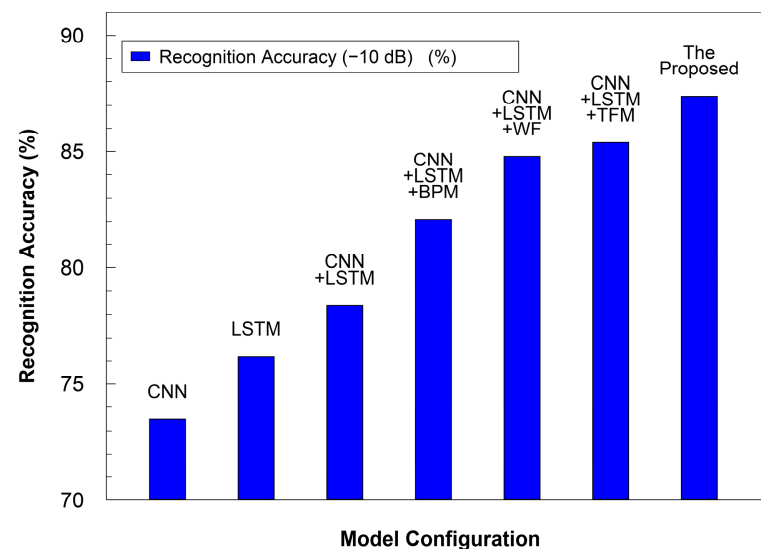


Figure 6. Ablation experimental results.

The data in Figure 6 indicates that the CNN-LSTM hybrid architecture outperforms single CNN or LSTM methods, demonstrating synergistic effects between spatial feature extraction and temporal modeling in target recognition. Compared with the CNN+LSTM method without noise reduction, band-pass filtering improves recognition accuracy by 3.7%, the Wiener filtering method improves it by 6.3%, and the time–frequency masking method improves by 6.9%. Furthermore, the proposed two-layer noise reduction method outperforms single-stage noise reduction strategies, achieving a 2.7% higher recognition accuracy than Wiener filtering, 2.1% higher than time–frequency masking, and 5.3% higher than band-pass filtering (82.1%). The above results demonstrate that the proposed two-layer noise reduction strategy has a synergistic gain effect, which can deeply suppress noise while maximally preserving the core recognition features of UAV signals, ultimately achieving optimal recognition performance under extremely low-SNR scenarios.

3.7. UAV Category Recognition Performance

To evaluate the fine-grained recognition ability of the proposed method for different types of UAVs, precision, recall, and F1-score are used as core evaluation metrics for classification tasks. Quantitative testing and analysis of the recognition performance for eight types of typical UAVs are conducted under full SNR conditions. The results of each metric for recognition performance of each UAV type are shown in Table 8.

From the overall test results, the proposed method achieves an average precision of 88.4%, average recall of 86.8%, and average F1-score of 0.876 in the fine-grained recognition task for eight UAV types, demonstrating excellent overall recognition performance.

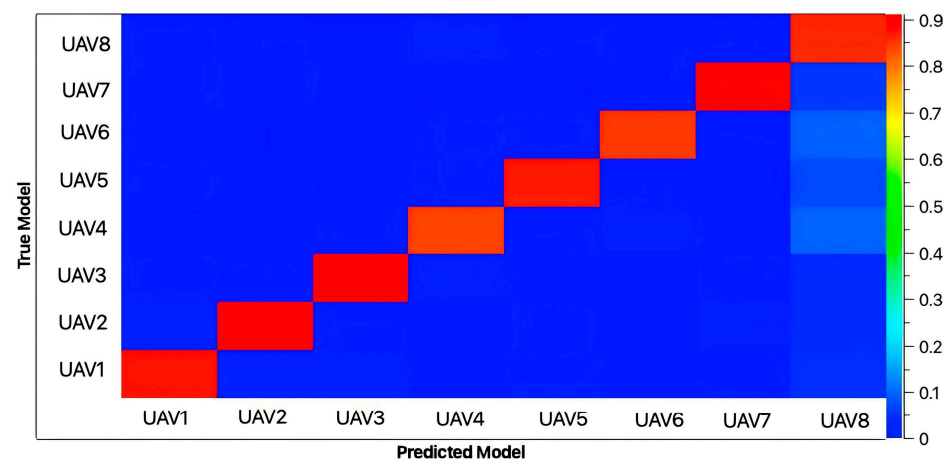
Table 8. UAV type recognition metrics.

UAV Type	Precision	Recall	F1-Score
UAV 1	89.2%	87.5%	0.883
UAV 2	91.5%	89.7%	0.906
UAV 3	87.3%	85.8%	0.865
UAV 4	85.7%	84.2%	0.849
UAV 5	88.9%	87.1%	0.880
UAV 6	86.4%	85.0%	0.857
UAV 7	90.1%	88.6%	0.893
UAV 8	87.8%	86.3%	0.870
Average	88.4%	86.8%	0.876

From the differential characteristics of single-model recognition performance, there are certain distribution differences in recognition performance among UAVs with different structures and purposes. The proposed method maintains approximately 90% high-precision recognition for most UAV models, with only slight performance degradation on complex multi-rotor structure models such as six-rotor (UAV 7) and eight-rotor (UAV 8) models. However, it still fully meets the application requirements of UAV type recognition in scenarios such as actual transmission line inspection.

3.8. Confusion Matrix Analysis

To reveal the category recognition characteristics and misclassification patterns of the model, full-category recognition confusion matrix analysis is conducted under the -10 dB SNR test scenario. The 2D heatmap is shown in Figure 7, and specific experimental data from the confusion matrix analysis are presented in Table 9.

**Figure 7.** Confusion matrix heatmap.

The confusion matrix heatmap in Figure 7 shows the classification results for eight UAV categories. The diagonal values represent the correct classification ratio, ranging from 84.2% (UAV 4) to 89.7% (UAV 3).

As shown in Table 9, the diagonal elements represent the correct classification ratios for each UAV category under the -10 dB SNR condition. The values on the diagonal range from 0.842 to 0.897, indicating that the proposed method maintains stable classification performance even under strong gas discharge noise. The off-diagonal elements represent misclassification ratios between different UAV categories. Most off-diagonal values remain low, suggesting that only limited confusion occurs among the eight UAV types. Relatively higher misclassification ratios mainly appear between acoustically similar UAVs, which may have comparable rotor numbers, sizes, or harmonic structures. This result further

demonstrates that the proposed denoising and CNN-LSTM recognition framework can preserve discriminative acoustic features under extremely low-SNR conditions.

Table 9. Confusion matrix of eight UAV categories under the -10 dB SNR condition. Rows indicate true UAV categories, columns indicate predicted UAV categories, diagonal values indicate correct classification ratios, and off-diagonal values indicate misclassification ratios.

Types	UAV 1	UAV 2	UAV 3	UAV 4	UAV 5	UAV 6	UAV 7	UAV 8
UAV 1	0.875	0.021	0.015	0.008	0.012	0.009	0.010	0.050
UAV 2	0.018	0.888	0.012	0.007	0.010	0.008	0.015	0.042
UAV 3	0.012	0.010	0.897	0.015	0.009	0.007	0.012	0.038
UAV 4	0.008	0.007	0.012	0.842	0.010	0.015	0.009	0.097
UAV 5	0.010	0.009	0.008	0.012	0.871	0.010	0.008	0.072
UAV 6	0.009	0.008	0.007	0.010	0.012	0.850	0.010	0.094
UAV 7	0.010	0.012	0.009	0.008	0.007	0.010	0.886	0.058
UAV 8	0.012	0.010	0.008	0.015	0.010	0.012	0.010	0.863

3.9. Scenario Generalization Ability

To verify the generalization ability and scenario robustness of the proposed method in different practical application environments, cross-scenario generalization tests are conducted in three typical scenarios: an outdoor environment, indoor environment, and new environment with random noise perturbations. Using recognition accuracy as the core evaluation metric, the recognition stability of the model under non-training distribution scenarios is quantitatively assessed [28,29]. The results are shown in Table 10.

Table 10. Cross-scenario generalization results.

Test Scenario	Recognition Accuracy	Accuracy Change
Laboratory Scenario	87.4%	-
Outdoor Environment	89.7%	+2.3%
Indoor Environment	91.3%	+4.0%
New Environment	87.8%	+0.4%

As can be seen from Table 10, the proposed method achieves a recognition accuracy of 87.4% under the training scenario. In completely new, unknown environments that did not participate in training, the model still maintains a high accuracy of 87.8%, showing no significant performance degradation compared with the laboratory scenario. In two typical open environments, indoor and outdoor, recognition accuracy further increases to 91.3% and 89.7%, representing improvements of 4.0% and 2.3%, respectively.

Overall, in various unseen environments, the model recognition accuracy remains stable within the 87–91% range without significant performance degradation, fully demonstrating that the proposed two-layer noise reduction and CNN-LSTM hybrid recognition method can effectively suppress environmental noise and background interference, possessing excellent cross-scenario generalization ability, strong scenario robustness, and engineering practical value.

4. Discussion

The aforementioned experimental data analysis demonstrates that in scenarios with strong gas discharge noise from electrical facilities, the performance of traditional identification methods exhibits significant shortcomings, while the proposed method in this paper demonstrates excellent anti-noise capability and identification performance. The traditional MFCC-SVM method is severely affected by gas discharge noise and shows

limited identification accuracy under strong noise conditions, making it difficult to meet practical monitoring requirements. In contrast, the method proposed in this paper achieves an identification accuracy of 87.4% under the same extreme noise environment, with a significant performance improvement, effectively solving the core problem of identification performance degradation under strong noise conditions.

For the unique gas discharge noise in electrical power sites, this paper designs a two-layer denoising method combining adaptive Wiener filtering and time–frequency masking, achieving an optimal balance between targeted denoising and feature preservation. Among them, Wiener filtering can accurately suppress stationary background noise, effectively preserving the harmonic structure of UAV acoustic signals and avoiding feature distortion caused by excessive denoising. Time–frequency masking can specifically remove random pulse interference, maintain the temporal continuity of acoustic signals, and compensate for the deficiency of single Wiener filtering in suppressing pulse noise and the limited effect of a single masking method in processing stationary noise. The aforementioned signal processing methods provide an effective technical path for low-cost and high-reliability UAV acoustic monitoring at electrical power sites.

The computational complexity and practical tradeoff of the proposed method are further discussed as follows.

The computational complexity and practical tradeoff of the proposed method are further discussed as follows. The computational cost of the proposed framework is mainly introduced by multichannel STFT/ISTFT, adaptive Wiener gain estimation, time–frequency masking, mel-spectrogram extraction, and CNN-LSTM inference. The CNN-LSTM classifier is more computationally complex than traditional MFCC-SVM methods because it jointly models local time–frequency patterns and temporal dependencies. However, this additional cost is introduced to improve recognition robustness under extremely low-SNR conditions. The experimental results show that, at -10 dB SNR, the proposed method achieves substantially higher recognition accuracy than traditional and single-stage baseline methods. Therefore, the method represents a tradeoff between computational complexity and robustness, which is acceptable for power-facility monitoring scenarios where recognition reliability is prioritized.

From the performance comparison across the full signal-to-noise ratio range, the proposed method outperforms other existing similar models under all tested signal-to-noise ratios, and the performance advantage becomes more significant as the SNR decreases. Among them, under the strong noise condition of -10 dB, the proposed method improves accuracy by 23.7 percentage points compared with current methods. From the perspective of performance stability, the accuracy decline rate of the proposed method from $+10$ dB to -10 dB is only 9.8%, far lower than the 31–49% decline rate of traditional methods, indicating that this method has extremely strong adaptability to changes in noise intensity and can stably adapt to complex environments with large fluctuations in noise intensity at electrical power sites. Cross-scenario generalization experiments further verify the engineering applicability of the method. In outdoor, indoor, and entirely new unknown environments that did not participate in training, the model identification accuracy remains stable at 87–91%, effectively avoiding the problem of repeated model training for different electrical scenarios and reducing engineering deployment costs.

It should be emphasized that the evaluation in this study is not limited to low-noise or low-variability noise conditions. The low-noise scenario is used only as a baseline for comparison. In contrast, the high-voltage laboratory scenario is designed to simulate gas discharge interference in power-facility environments, and the mixed environment introduces wind noise and other background interferences to examine the robustness of the proposed method under more variable acoustic conditions. In addition, the experiments

cover five SNR levels from +10 dB to −10 dB, including the extremely low-SNR condition of −10 dB. The cross-scenario tests in outdoor, indoor, and new environments with random noise perturbations further indicate that the evaluation includes unseen and variable acoustic conditions rather than only low-noise cases.

In summary, the core advantages of the proposed method can be summarized into the following three aspects. First, the anti-noise robustness is prominent. Through two-layer denoising and multichannel collaborative enhancement, it effectively suppresses gas discharge noise interference at electrical power sites, maintaining high identification accuracy even in extreme strong noise environments. Second, the identification accuracy is excellent, outperforming existing benchmark methods across the full signal-to-noise ratio range, especially with significant performance improvement in strong noise scenarios, enabling accurate recognition of different UAV models. Third, the generalization capability is strong, with stable performance in cross-scenario and cross-model tests, adapting to diverse monitoring environments and UAV types at electrical power sites.

Regarding the model architecture, Transformer- and Conformer-based architectures, as well as more specialized convolutional neural networks, are promising alternatives for acoustic recognition. However, these models usually require larger labeled datasets and higher computational resources to fully exploit their modeling capacity and to avoid overfitting. In the present study, the dataset scale is limited, and the main objective is to verify the effectiveness of adaptive time–frequency denoising and multichannel feature fusion under strong gas discharge noise. Therefore, a CNN-LSTM architecture was selected to balance local time–frequency feature extraction, temporal dependency modeling, and deployment feasibility. The CNN branch extracts local frequency-energy patterns from mel-spectrograms, while the Bi-LSTM branch models the temporal evolution of UAV acoustic features. In addition, CNN-ResNet, LSTM, and CRNN baselines were included for comparison. Lightweight Transformer, Conformer, or more specialized acoustic CNN architectures will be considered in future work after larger-scale data are available.

Although the proposed method has achieved good results in UAV acoustic identification under gas discharge noise environments of electrical facilities, there are still the following limitations combined with practical engineering application requirements. First, it has a certain dependence on prior information. In complex scenarios with dynamically changing noise types and severe fluctuations in discharge intensity, prior information is difficult to obtain accurately, affecting the adaptive capability of the method. Second, the generalization range is limited. Currently, the method is mainly optimized for common corona-discharge and spark discharge noise at electrical power sites, with insufficient adaptability to extreme discharge types such as arc discharge, and the identification performance for unseen niche UAV models needs improvement. These issues will be addressed in subsequent research through adaptive threshold learning and online noise estimation to reduce dependence on manually set prior information, and enhance model generalization capability through data augmentation, further expanding the applicable scenarios of the method and promoting its engineering implementation and large-scale application in UAV monitoring of electrical facilities.

5. Conclusions

This paper addresses the engineering challenge of UAV acoustic recognition performance degradation caused by strong gas discharge noise around power facilities. The core conclusions are as follows:

Corona-discharge wideband stationary noise and spark discharge random impulse noise in high-voltage power transmission and transformation sites overlap in frequency bands with UAV operation noise. The laboratory spectra measured under 8 kV, 14 kV, and

17 kV applied-voltage conditions show voltage-dependent acoustic energy distributions over the 0–20 kHz frequency range. These discharge components can mask UAV rotor harmonic features and destroy the temporal continuity of UAV acoustic signals, thereby making UAV acoustic feature extraction difficult.

A two-stage denoising and multichannel collaborative signal processing method adapted to power sites has been constructed. Through adaptive Wiener filtering to suppress stationary background noise and time–frequency masking to remove random impulse interference, the method achieves efficient denoising while preserving the integrity and fidelity of UAV acoustic features. Quantitatively, the denoising module achieves a 12.3 dB SNR improvement, a spectral fidelity score of 0.89, and a PESQ score of 2.87, indicating that useful acoustic features can be retained while strong discharge noise is reduced.

A CNN-LSTM hybrid recognition model integrated with an attention mechanism has been proposed, which balances spatial and temporal feature extraction. Across five SNR levels from +10 dB to −10 dB, the proposed method maintains stable recognition performance. Under the extremely low-SNR condition of −10 dB, it achieves 87.4% recognition accuracy, and the accuracy decline from +10 dB to −10 dB is only 9.8%, much lower than the 31–49% decline observed for baseline methods.

The proposed method contributes to acoustics by combining adaptive time–frequency denoising, multichannel feature fusion, and deep temporal–spatial modeling for robust recognition under severe nonstationary noise. It also contributes to aerospace and power infrastructure applications by providing a passive, low-cost, and deployable acoustic monitoring solution for UAV intrusion detection around substations and overhead transmission lines.

Author Contributions: Conceptualization, C.W., Z.Z. and G.Y.; methodology, C.W., W.K. and M.Z.; software, C.W.; validation, C.W., W.K. and M.Z.; data curation, C.W.; writing—original draft preparation, C.W.; writing—review and editing, Z.Z., W.K. and M.Z.; visualization, C.W.; funding acquisition, G.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Innovation and Entrepreneurship Training Program for College Students of the Ministry of Education of the People’s Republic of China under Grant No. 202410700048.

Institutional Review Board Statement: Not applicable. The study did not involve human participants or human-related biological samples, and ethical review approval was not required.

Informed Consent Statement: Not applicable. This study does not involve any human participants, human samples or identifiable personal data.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to containing commercially sensitive information.

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

References

1. Shakhathreh, H.; Sawalmeh, A.H.; Al-Fuqaha, A.; Dou, Z.; Almaita, E.; Khalil, I.; Othman, N.S.; Khreishah, A.; Guizani, M. Unmanned aerial vehicles (UAVs): A survey on civil applications and key research challenges. *IEEE Access* **2019**, *7*, 48572–48634. [\[CrossRef\]](#)
2. Khan, M.A.; Menouar, H.; Eldeeb, A.; Abu-Dayya, A.; Salim, F.D. On the detection of unauthorized drones-techniques and future perspectives: A review. *IEEE Sens. J.* **2022**, *22*, 11439–11455.
3. Lykou, G.; Moustakas, D.; Gritzalis, D. Defending airports from UAS: A survey on cyber-attacks and counter-drone sensing technologies. *Sensors* **2020**, *20*, 3537. [\[PubMed\]](#)
4. Frid, A.; Ben-Shimol, Y.; Manor, E.; Greenberg, S. Drones detection using a fusion of RF and acoustic features and deep neural networks. *Sensors* **2024**, *24*, 2427. [\[CrossRef\]](#) [\[PubMed\]](#)
5. Ding, R.; Yang, L.; Shan, Z.; Wang, Y.; Ren, H.; Liu, L.; Yu, L.; Liu, B. Drone detection and recognition based on fiber-optic EFPI acoustic sensor and CNN-LSTM network model. *IEEE Sens. J.* **2024**, *24*, 28835–28843.

6. Li, Y.; Li, Q.; Meng, W.; Tang, Q.; Li, L.; Hu, J.; Jiang, X. Finite Element Method Assisted Audible Noise Detection for Overhead Line Conductors Using the Cage Experiment. *Energies* **2023**, *16*, 4564. [\[CrossRef\]](#)
7. Cinoglu, B.; Durak, U.; Karakoc, T.H. Utilizing Mel-Frequency Cepstral Coefficients for Acoustic Diagnostics of Damaged UAV Propellers. *Int. J. Aviat. Sci. Technol.* **2024**, *05*, 79–89. [\[CrossRef\]](#)
8. 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. [\[CrossRef\]](#)
9. Glüge, S.; Nyfeler, M.; Aghaebrahimian, A.; Ramagnano, N.; Schüpbach, C. Robust Drone Detection and Classification Using Convolutional Neural Networks in Low SNR Environments. *IEEE J. Radio Freq. Identif.* **2024**, *8*, 821–830. [\[CrossRef\]](#)
10. Fu, Y.; He, Z. Radio Frequency Signal-Based Drone Classification with Frequency Domain Gramian Angular Field and Convolutional Neural Network. *Drones* **2024**, *8*, 511. [\[CrossRef\]](#)
11. Azizi, K.; Sayahi, N.; Nejabat, E. Convolutional Neural Network Classifier for Unmanned Aerial Vehicles Detection and Identification Using Mel-Frequency Spectrograms. *J. Eng.* **2025**, *1*, e70135.
12. Kolamunna, H.; Dahanayaka, T.; Li, J.; Seneviratne, S.; Thilakarathne, K.; Zomaya, A.Y.; Seneviratne, A. DronePrint: Acoustic Signatures for Open-set Drone Detection and Identification with Online Data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* **2021**, *5*, 20. [\[CrossRef\]](#)
13. Liu, B.; Wang, H.; Yang, L.; Shang, J.; Fan, H.; Yu, L.; Shan, M.; Xiao, S. A Differential Fiber-Optic FPI Acoustic Sensor for UAV Detection With CNN-LSTM. *IEEE Sens. J.* **2026**, *26*, 15293–15301. [\[CrossRef\]](#)
14. Li, J.; Zhao, J.; Ren, J.; Gao, X.; Li, Z. A Multi-Scale Integrated Learning Model with Attention Mechanisms for UAV Audio Signal Detection. *Signal Image Video Process.* **2025**, *19*, 344. [\[CrossRef\]](#)
15. Dumitrescu, C.; Minea, M.; Costea, I.M.; Chiva, I.C.; Semenescu, A. Development of an Acoustic System for UAV Detection. *Sensors* **2020**, *20*, 4870. [\[CrossRef\]](#) [\[PubMed\]](#)
16. Pan, W.; Fan, K.; Zhong, J.; He, Q.; Huang, J. Research on Sound Recognition of UAV Based on Spherical Microphone Arrays. In Proceedings of the 2025 IEEE 20th Conference on Industrial Electronics and Applications (ICIEA), Yantai, China, 3–6 August 2025.
17. Nakadai, K.; Kumon, M.; Okuno, H.G.; Hoshiba, K.; Wakabayashi, M.; Washizaki, K.; Ishiki, T.; Gabriel, D.; Bando, Y.; Morito, T.; et al. Development of Microphone-Array-Embedded UAV for Search and Rescue Task. In Proceedings of the 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Vancouver, BC, Canada, 24–28 September 2017.
18. Yousif, S.T.; Mahmmoud, B.M. Speech enhancement algorithms: A systematic literature review. *Algorithms* **2025**, *18*, 272. [\[CrossRef\]](#)
19. Jadda, A.; Prabha, I.S. Speech enhancement via adaptive Wiener filtering and optimized deep learning framework. *Int. J. Wavelets Multiresolut. Inf. Process.* **2023**, *21*, 2250032.
20. Karamyan, D.S. Adaptive noise cancellation for robust speech recognition in noisy environments. *Proc. Yerevan State Univ. Phys. Math.* **2024**, *58*, 22–29. [\[CrossRef\]](#)
21. Williamson, D.S.; Wang, D. Time-Frequency Masking in the Complex Domain for Speech Dereverberation and Denoising. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1492–1501. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Taha, T.M.F.; Wajid, S.K.; Hussain, A. Speech enhancement based on adaptive noise cancellation and particle swarm optimization. *J. Comput. Sci.* **2019**, *15*, 691–701. [\[CrossRef\]](#)
23. Ullah, R.; Wuttisittikulkij, L.; Chaudhary, S.; Parnianifard, A.; Shah, S.; Lbrar, M.; Wahab, F. End-to-end deep convolutional recurrent models for noise robust waveform speech enhancement. *Sensors* **2022**, *22*, 7782. [\[CrossRef\]](#) [\[PubMed\]](#)
24. He, T.; Hou, J.; Chen, D. Multimodal UAV Target Detection Method Based on Acousto-Optical Hybridization. *Drones* **2025**, *9*, 627. [\[CrossRef\]](#)
25. Liu, Z.; Zou, Y.; Hu, Z.; Xue, H.; Li, M.; Rao, B. Research on Multi-Modal Fusion Detection Method for Low-Slow-Small UAVs Based on Deep Learning. *Drones* **2025**, *9*, 852. [\[CrossRef\]](#)
26. Yuan, S.; Yang, Y.; Nguyen, T.H.; Nguyen, T.M.; Yang, J.; Liu, F.; Li, J.; Wang, H.; Xie, L. MMAUD: A Comprehensive Multi-Modal Anti-UAV Dataset for Modern Miniature Drone Threats. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024.
27. Lei, H.; Gadgil, R.; Amgothu, S.K.; Kar, D. UAV Audio Detection and Identification Using Short-Time Fourier Transform Spectrograms with Deep Learning Models. In Proceedings of the 2025 International Conference on Unmanned Aircraft Systems (ICUAS), Charlotte, NC, USA, 14–17 May 2025.
28. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1345–1359. [\[CrossRef\]](#)
29. Zhuang, F.; Qi, Z.; Duan, K.; Xi, D.; Zhu, Y.; Zhu, H.; Xiong, H.; He, Q. A comprehensive survey on transfer learning. *Proc. IEEE* **2021**, *109*, 43–76. [\[CrossRef\]](#)

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.