drones

MDPI

Article

Audio–Visual Fusion Sim2Real Platform for Anti-UAV Detection and Tracking

Xiaohong Nian [1,†], Haolun Liu [1,†] and Xunhua Dai [2,*]

[1] Department of Automation, Central South University, Changsha 410083, China
[2] School of Computer Science and Engineering, Central South University, Changsha 410083, China
* Correspondence: dai.xh@csu.edu.cn; Tel.: +86-152-1090-5355
† These authors contributed equally to this work.

Highlights

What are the main findings?

- We developed a Sim2Real physics-informed audio–visual simulation platform that integrates physics-based acoustic propagation modeling with dynamic visual environments to simulate complex degradation scenarios like fog and occlusion.
- Experimental validation, conducted through both simulation and physical field tests, demonstrates that the proposed audio–visual fusion strategy extends the tracking coverage of single-modality vision systems from approximately 52% to near-continuous operation, maintaining reliable target awareness even when visual targets are severely obstructed, albeit with a moderate trade-off in average angular precision.

What are the implications of the main findings?

- This platform significantly mitigates the high costs and safety risks associated with real-world data collection by providing a scalable solution for generating precisely annotated multi-modal datasets, thereby accelerating the training and validation of robust fusion algorithms.
- The validated consistency between simulation and physical results confirms the system's capability to bridge the simulation-to-reality gap, offering a critical infrastructure for developing next-generation all-weather Anti-UAV perception systems.

Abstract

To address the escalating security challenges posed by unauthorized Unmanned Aerial Vehicles, this paper presents a Sim2real physics-informed audio–visual fusion simulation platform designed to enhance Counter-Unmanned Aerial Vehicle detection and tracking performance. The proposed method integrates two complementary sensing pipelines: a physics-based acoustic localization system utilizing Time Difference of Arrival principles and a deep learning-driven visual detection framework. To ensure robust surveillance against non-cooperative targets, these pipelines are not only fused through strict spatiotemporal synchronization but also mutually reinforce each other—acoustic data guides visual attention in low-visibility scenarios typical of adversarial intrusions, while visual detections refine acoustic parameter estimation. Building upon prior work in multi-modal perception, we extend the framework to dynamic environments characterized by configurable visual obstructions, including smoke and fog, which frequently compromise conventional optical anti-drone systems. Experiments demonstrate that the fusion system progressively adapts to degraded visual conditions, extending tracking continuity from approximately 50% coverage under vision-only operation to near-continuous target awareness, with a moderate trade-off in average angular precision when acoustic-only segments are included. Physical validation with quadrotor Unmanned Aerial Vehicles confirms the platform's

Check for updates

Academic Editor: San Jiang

Received: 25 January 2026
Revised: 5 March 2026
Accepted: 5 March 2026
Published: 10 March 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) license.

capability to bridge simulation-to-reality gaps. Our results highlight the system's robustness against sensor degradation and its potential to accelerate the development of resilient multisensor Counter-Unmanned Aerial Vehicle systems while reducing dependency on costly field testing.

Keywords: audio–visual fusion; Sim2Real; anti-UAV; simulation platform; TDOA localization

1. Introduction

The security landscape of low-altitude airspace is undergoing a profound transformation, primarily driven by the democratization and rapid capability growth of consumer-grade aerial technology. While Unmanned Aerial Vehicles (UAVs) offer transformative potential for commercial logistics and photography, their weaponization and misuse for unauthorized surveillance present severe asymmetric threats to critical infrastructure and public safety. Incidents involving airspace disruption at major transportation hubs and unauthorized incursions into sensitive government facilities have underscored the critical fragility of traditional perimeter defense mechanisms. Unlike cooperative aircraft that adhere to flight paths and broadcast identification signals, non-cooperative drones often operate with high maneuverability at ultra-low altitudes. These targets effectively exploit urban topography and ground clutter to mask their presence, possessing a minimal Radar Cross-Section (RCS) that renders them nearly invisible to conventional air defense radars designed for larger ballistic targets. However, acquiring sufficient labeled data for such erratic targets in extreme conditions is significantly costly and risky, creating a compelling imperative for Sim2Real paradigms that can synthesize diverse corner cases to bootstrap real-world robustness. Consequently, the establishment of robust Counter-UAV (C-UAV) capabilities has transitioned from a precautionary measure to a critical operational requirement, necessitating the deployment of next-generation detection technologies that are resilient to both environmental complexity and active concealment strategies.

Driven by these security imperatives, the ubiquitous deployment of UAVs has not only intensified the demand for intelligent perception algorithms to ensure legitimate operational safety [1,2], but also precipitated an urgent need for enhanced regulatory frameworks and defensive countermeasures. However, current C-UAV systems primarily depend on single-sensor modalities, typified by vision-based architectures such as the YOLO series and LiDAR [3–5]. These unimodal approaches exhibit inherent vulnerabilities when operating in complex, adversarial environments. Visual surveillance systems, while providing rich texture information, suffer precipitous performance degradation under occlusion or adverse weather conditions, including smoke, dust storms, and dense fog. Furthermore, their reliability is severely compromised during the transition between day and night, or by abrupt illumination changes that result in overexposure or motion blur. Similarly, active sensors like LiDAR face distinct challenges; their signals are susceptible to absorption by low-reflectivity materials (often used in stealth drone coatings) and suffer from significant attenuation in heavy rain. Moreover, active LiDAR pulses expose the detection system to discovery and are vulnerable to electromagnetic interference in urban or adversarial electronic warfare scenarios [6,7].

To address the limitations of unimodal perception, multisensor fusion has emerged as a pivotal strategy in UAV detection tasks [8,9]. The rationale lies in the complementary nature of the modalities: visual systems facilitate high-precision localization and fine-grained classification under ideal illumination but fail drastically in occluded or low-visibility scenarios, such as smoke or fog [10,11]. In contrast, acoustic sensing possesses omnidirectional

detection capabilities and remains functional in visually degraded environments such as total darkness, albeit with a spatial resolution that is governed by the array aperture and the number of sensors [12]. The strategic integration of these two modalities—leveraging the distinct advantages of acoustic robustness against occlusion and visual precision for target verification—is therefore critical for constructing resilient environmental perception systems capable of all-weather surveillance [13–15].

To facilitate the development of these complex fusion systems, simulation platforms provide a cost-effective, scalable, and risk-free solution for C-UAV audio–visual perception research. Constructing real-world datasets for rogue drone detection is often constrained by stringent safety regulations, privacy concerns, and the practical difficulty of reproducing extreme conditions. For example, staging a drone intrusion in severe weather or testing jamming scenarios requires complex logistical approval and poses physical risks. Furthermore, obtaining precise ground truth data—specifically the exact 3D position of the drone relative to sensors—is inherently difficult in outdoor field experiments. By mathematically modeling real-world multi-modal interference scenarios—such as visual occlusions caused by volumetric fog, variable lighting conditions, or acoustic attenuation due to distance and obstacles—simulation platforms significantly mitigate data collection costs. They allow researchers to generate massive, precisely annotated datasets, thereby accelerating the iteration and validation of robust fusion algorithms before physical deployment [16].

In the domain of UAV detection research, the trajectory has shifted decisively towards multi-modal fusion and algorithmic optimization to tackle the challenge of identifying small, fast-moving targets against complex backgrounds. Within the visual domain, deep learning-based object detection frameworks, notably the YOLO series and Faster R-CNN, have evolved to integrate attention mechanisms and Feature Pyramid Networks (FPNs). These architectural innovations are designed to mitigate the small object problem, enhancing the feature representation of distant targets that occupy only a few pixels in the frame [17–19]. In parallel, wavelet transform-based methods have demonstrated effectiveness for detecting small objects in aerial imagery by capturing multi-scale spatial-frequency features, as recently applied to UAV image analysis for infrastructure inspection [20]. Furthermore, temporal analysis techniques, such as optical flow methods, have been employed to improve the dynamic tracking of fast-moving targets by exploiting motion continuity [21]. However, despite these advancements, the adaptability of visual algorithms remains constrained by environmental factors. Abrupt illumination changes, complex background clutter in urban skylines (such as moving vehicles or swaying trees), and adverse weather all degrade visual feature discriminability, frequently leading to high false-positive rates or missed detections in operational settings [22].

Parallelly, acoustic detection algorithms offer a non-line-of-sight detection capability. Traditionally focusing on Mel-frequency cepstral coefficients (MFCCs) and wavelet transforms for spectral signature extraction [23–25], these methods have evolved to integrate deep architectures such as convolutional neural networks (CNNs) and long short-term memory (LSTM) networks. These models are often coupled with microphone array beamforming and time difference of arrival (TDOA) techniques to estimate the spatial origin of the drone based on its distinct rotor noise [26–28]. However, acoustic detection faces its own set of challenges. The efficacy of these systems is frequently compromised by environmental noise interference, particularly from wind gusts and urban traffic which can mask the drone's acoustic signature. Additionally, the high-frequency components of drone noise suffer significant atmospheric attenuation over long distances, reducing the effective detection range compared to visual counterparts [29].

To bridge these individual gaps, Audio–visual fusion has been adopted, employing hierarchical frameworks that range from data-level to decision-level integration [30]. Modern

approaches utilize spatiotemporal synchronization of raw data streams, Transformer-based cross-attention mechanisms for deep feature correlation, and Bayesian-weighted decision-integration to enhance detection reliability in low-light or occluded environments [31,32].

While mainstream simulation platforms such as Gazebo and AirSim establish benchmarks for flight dynamics and visual perception [33,34], they are not specialized for physics-informed acoustic modeling. These general-purpose platforms often simplify complex wave phenomena—such as the Doppler effect, atmospheric absorption, and reverberation—creating a significant Sim2Real gap. Consequently, algorithms validated in such acoustically idealized environments struggle to adapt to physical deployment. To address this limitation, we propose a Sim2Real platform that integrates rigorous acoustic propagation models, thereby aligning the synthetic data distribution with real-world environments.

Figure 1 presents the overall architecture of our Sim2Real platform, encompassing both the simulation environment and the corresponding physical experimental setup. It is important to note that the visual perspectives in this figure are chosen to best highlight the environmental rendering features and hardware deployment, respectively, rather than representing a controlled experimental comparison.

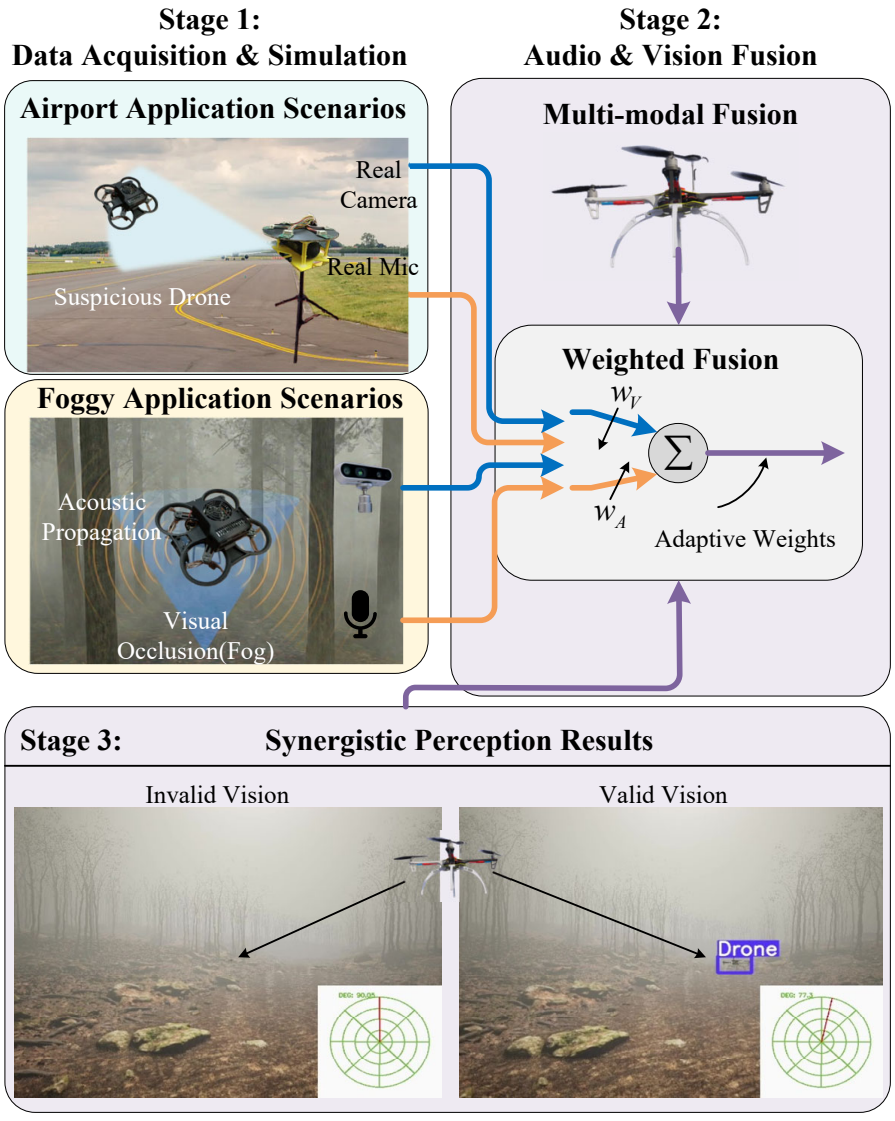


Figure 1. Application scenarios of the proposed Sim2Real audio–visual anti-UAV platform. The simulation side illustrates representative deployment contexts rendered in the Unreal Engine 5.1 environment, while the physical side shows corresponding real-world deployment conditions, highlighting the platform's capability to bridge virtual and physical domains for counter-UAV perception research.

The primary contributions of this work are threefold:

- **Sim2Real Audio–Visual Platform:** Building upon the flight dynamics foundation of RFlySim [35], we develop a specialized simulation ecosystem that introduces a physics-based acoustic propagation engine capable of real-time simulation of atmospheric absorption, geometric spreading, multipath reverberation, and the Doppler effect. Unlike general-purpose simulators (e.g., Gazebo, AirSim) that lack physics-informed acoustic modeling, our platform provides strict audio–visual synchronization and configurable environmental degradation scenarios (fog, smoke, occlusion) for reproducible multi-modal dataset generation.
- **Confidence-Gated Complementary Fusion Strategy:** In contrast to mainstream static feature-level or decision-level fusion approaches, we propose a bidirectional mutual refinement mechanism: acoustic sensing provides directional guidance to steer the visual field of view when visual confidence degrades below a threshold, while reliable visual detections refine the angular estimation of the TDOA-based acoustic subsystem. This dynamic switching ensures continuous tracking coverage under severe environmental degradation.
- **Sim2Real Validation:** Rigorous cross-domain experiments demonstrate consistent behavioral trends between simulation and physical deployment, confirming the platform's transfer capability and the fusion strategy's practical robustness.

2. Multi-Modal Simulation Framework Design

The RFlySim framework [35] provides essential foundational functions, such as high-precision custom kinematic models, 3D environmental modeling, and basic visual/point-cloud sensor simulations. However, it completely lacks acoustic sensor simulation capabilities and the necessary software interfaces for audio-based algorithms. To bridge this critical gap, the simulation platform developed in this study inherits the flight dynamics engine and controller interfaces (SITL/HITL) from RFlySim, and extends it by introducing a dedicated physics-based acoustic simulation layer and comprehensive audio–visual development toolchains built upon Unreal Engine 5.1. Specifically, it represents a comprehensive, physics-informed ecosystem specifically engineered to bridge the gap between theoretical acoustic models and real-world UAV deployment. Unlike general-purpose flight simulators, this platform specializes in advanced acoustic propagation modeling, providing a rigorous testbed for sound-based algorithms, including Sound Source Localization (SSL) and blind source separation. It incorporates a physics-based acoustic engine that accurately emulates complex wave phenomena, encompassing atmospheric absorption, geometric spreading loss, multipath reverberation in cluttered environments, and the Doppler effect induced by high-speed relative motion.

To ensure experimental versatility, the platform features highly customizable sound source configurations and flexible scenario generation. Researchers can precisely define the spectral characteristics of target sound sources, such as specific rotor harmonics, and inject diverse environmental noise profiles—ranging from wind turbulence to urban background noise—to simulate varying Signal-to-Noise Ratios (SNRs). Furthermore, the platform integrates high-precision visual sensor interfaces, enabling the generation of synchronized RGB, depth, and thermal imagery. This strict spatiotemporal alignment between audio and visual data streams is critical for developing and testing audio–visual fusion algorithms, thereby enhancing multi-modal perception capabilities in visually degraded or acoustically noisy environments.

Beyond single-agent capabilities, the architecture supports UAV swarm simulation, facilitating research into multi-agent coordination, distributed perception, and collaborative swarm control strategies. This scalability allows for the validation of complex mission

profiles involving cooperative detection and tracking. Finally, to ensure the deployability of algorithms, the platform provides Software-in-the-Loop (SITL) and Hardware-in-the-Loop (HITL) simulation capabilities. This feature enables researchers to validate algorithms directly on embedded flight control hardware, such as the Pixhawk and PX4 series, assessing computational latency and real-time performance constraints prior to field experiments, thus offering a closed-loop validation pipeline from algorithm design to hardware deployment.

2.1. Auditory Modality Implementation

To ensure faithful acoustic replication, the simulation platform supports the granular assignment of dedicated noise profiles to distinct UAV models. This capability allows researchers to map unique acoustic signatures to specific drone architectures—ranging from small quadcopters to larger fixed-wing aircraft—thereby capturing the spectral nuances inherent to different propulsion systems. By enabling this level of acoustic heterogeneity, the platform significantly improves simulation accuracy and reliability, particularly for research tasks involving target classification and multi-type swarm identification.

Functionally, the platform establishes robust audio resource bindings that logically link specific sound source objects within the 3D virtual environment to their corresponding raw acoustic data files. This spatial association provides the fundamental signal support required for subsequent complex audio rendering processes, ensuring that sound emission is strictly coupled with the kinematic state of the UAV. Furthermore, to streamline the workflow, the system automatically creates usable audio assets by ingesting and converting properly imported audio files. This automated transcoding process ensures that external recordings are optimized for the simulation engine, handling necessary sample rate conversions and format standardization without manual intervention.

Beyond asset management, users possess comprehensive, fine-grained control over a wide array of audio parameters to tailor the simulation to specific experimental requirements. Configurable attributes include playback modes (looping or one-shot triggers), compression settings to balance audio fidelity against memory usage, and global volume scaling factors to calibrate source intensity. Crucially, the platform supports dynamic pitch adjustments, a feature essential for simulating the correlation between motor RPM and frequency shifts during maneuvering. To facilitate debugging and validation, the system also provides a capability for real-time monitoring through an integrated audio player, allowing operators to audit audio assets and verify signal integrity prior to executing full-scale simulations.

2.2. Spatial Audio Processing

In sound source localization algorithm simulation, spatial processing of audio sources plays a critical role. Audio spatialization simulates the positional and directional characteristics of sound in three-dimensional space to replicate real-world acoustic behavior. This process involves two key perceptual dimensions: directional perception and distance perception. The integration of these perceptual dimensions provides a comprehensive representation of a sound source's spatial position relative to the receiver. Specifically, spatial position variations are modeled using three key parameters: interaural time difference (ITD), interaural level difference (ILD), and frequency variation between sensors. ITD serves as a fundamental determinant for sound source localization. When a sound source approaches a sensor, the received sound intensity at that sensor increases significantly, thereby improving localization accuracy.

2.3. Sound Attenuation Analysis

Sound attenuation describes the gradual reduction of sound intensity during propagation due to distance or environmental factors, representing a fundamental principle in acoustics. In sound source localization simulations, accurate attenuation modeling enables realistic representation of acoustic propagation characteristics, providing test conditions that closely approximate real-world scenarios. The distance-dependent intensity attenuation demonstrates distinct frequency characteristics high-frequency sounds attenuate faster while low-frequency components propagate farther. Such modeling allows evaluation of algorithm performance under varying intensity and frequency conditions across different bands.

The platform integrates real-world validation with algorithm optimization capabilities. By incorporating attenuation models, developers can proactively enhance algorithm adaptability for improved real-world performance. The platform offers standardized benchmarking under consistent attenuation conditions while allowing flexible parameter adjustments to evaluate algorithm effectiveness across various propagation scenarios. The platform implements audio rendering through Unreal Engine's built-in Sound Attenuation asset, which dynamically adjusts sound intensity and frequency characteristics based on distance and environmental factors. Figure 2 illustrates the attenuation simulation workflow. For dynamic scenarios, Blueprint scripting enables real-time coordinate updates of moving sources, accurately simulating propagation changes during motion. The platform provides multiple attenuation models, allowing selection of linear attenuation for natural environments or complex nonlinear curves for industrial scenarios. Figure 3 presents typical attenuation curves for reference.

1. **Linear Attenuation:** When employing a linear function, the volume changes at a constant rate with distance. This straightforward attenuation approach is suitable for scenarios requiring uniform variations, such as simulating basic environmental sound effects or close-range source transitions. As the listener approaches or moves away from the sound source, the volume changes are intuitive and easily controllable, making it ideal for applications with fundamental requirements.
2. **Logarithmic Attenuation:** The attenuation model based on logarithmic functions more closely approximates natural attenuation patterns observed in the physical world. Its characteristic feature is pronounced volume changes at close distances and gradual variations at greater distances, enabling more realistic simulation of sound's physical propagation characteristics. This model is particularly suitable for most natural environments, including environmental sound effects and close-range source variations.
3. **Inverse Proportional Attenuation:** The inverse proportional function exhibits a similar trend to logarithmic functions but with greater variation magnitude. This model emphasizes close-range volume changes more significantly while rapidly decreasing volume at greater distances, making it particularly suitable for simulating distant source attenuation in open-world environments, such as modeling long-distance sound propagation.
4. **Custom Attenuation:** Developers can precisely control sound attenuation characteristics through custom attenuation functions to meet specific scenario requirements. For instance, the red curve in Figure 3 demonstrates the effect of a quadratic attenuation function: during the initial stage of increasing distance, sound attenuation occurs rapidly; as it approaches the outer radius, the attenuation rate stabilizes. This non-linear yet smooth attenuation approach is particularly suitable for complex scenarios requiring gradual fade-out with natural transitions, such as in sophisticated audio environments.

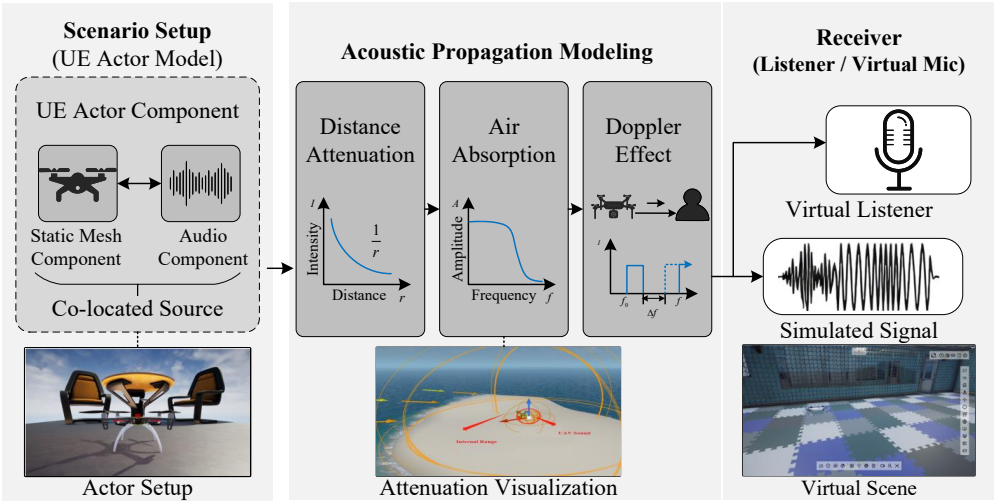


Figure 2. Sound attenuation simulation workflow. The diagram illustrates the processing pipeline from raw audio source binding through distance-dependent attenuation parameter configuration to the final spatially attenuated output rendered by the Unreal Engine Sound Attenuation asset.

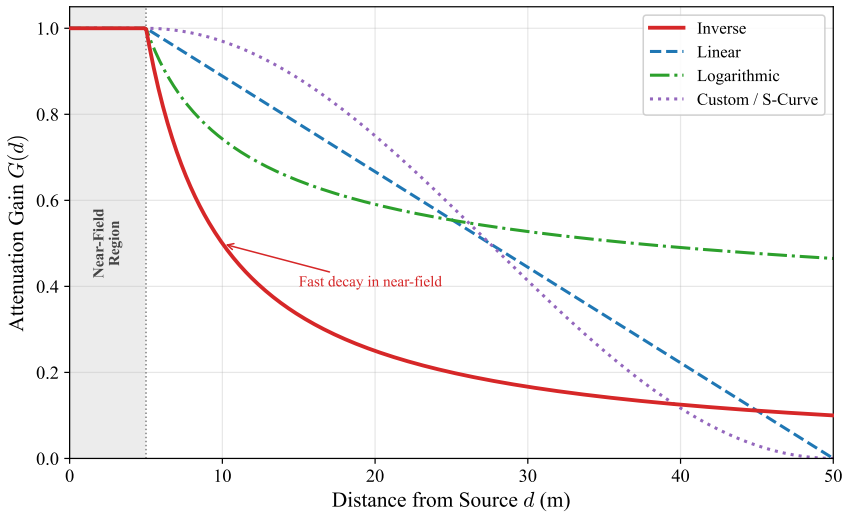


Figure 3. Representative sound attenuation curves as a function of normalized distance. Four attenuation models are compared: linear, logarithmic, inverse proportional, and a custom quadratic function (red curve), demonstrating the configurable intensity roll-off behavior supported by the platform.

2.4. Spatial Reverberation and Obstacle Reflection

Spatial reverberation refers to the acoustic phenomenon resulting from the reflection, refraction, and scattering of sound waves as they interact with environmental objects during propagation. This effect represents a fundamental characteristic of acoustic environments, enabling the simulation of realistic sound propagation behaviors in various spatial contexts. Specifically, this is realized through the Audio Volume actor coupled with the Reverb Effect asset provided by the engine's native audio subsystem. Reverb volumes define specific zones within a scene where predetermined reverberation characteristics are applied when either the sound source or listener enters these areas. Reverb effects provide comprehensive parameter customization, including reverberation decay time, density, diffusion, and gain, allowing for precise control over the acoustic properties of simulated environments.

2.5. Data Transmission Protocol

In simulation environments, audio data output for algorithmic processing is implemented through dedicated interfaces. The platform utilizes the UDP network protocol for

audio data transmission, leveraging its low-latency and real-time performance advantages. The system enables precise device addressing through IP and port configuration, with the loopback address 127.0.0.1 typically employed for local testing and debugging.

The platform provides real-time audio recording capabilities coupled with socket-based data transmission. The transmitter encodes audio at the configured sampling rate for continuous streaming, while the receiver decodes incoming data through designated port monitoring. To optimize performance, developers can adjust sampling rates to balance audio quality against bandwidth requirements. Critical to transmission integrity is maintaining parameter consistency between endpoints, including identical encoding formats, matching sampling rates and consistent bit depth specifications. This synchronization prevents decoding artifacts, playback anomalies, or signal degradation. The UDP socket implementation ensures efficient data transfer while preserving real-time characteristics, as format conversion during transmission would introduce undesirable latency. Through proper parameter configuration and sampling rate optimization, the platform delivers faithful audio data suitable for advanced algorithmic processing. Figure 4 shows the real-time operating status of the platform.

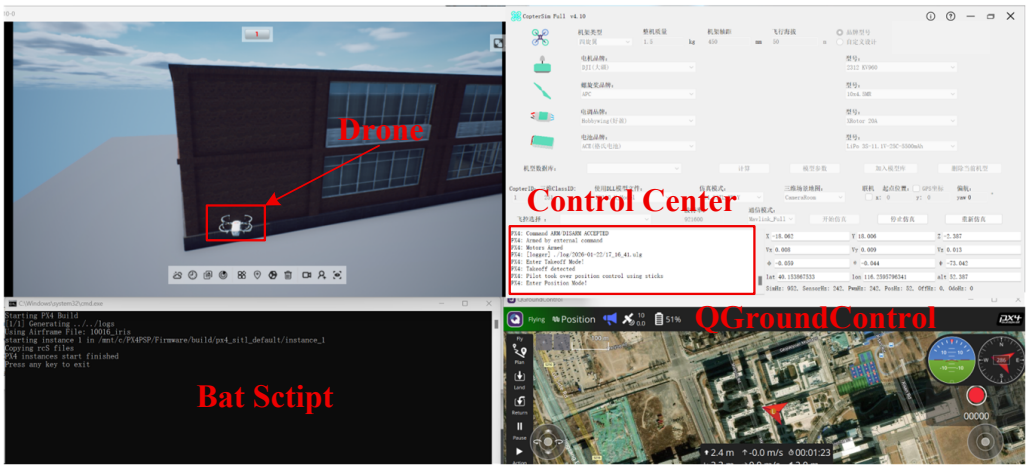


Figure 4. Real-time operating status of the simulation platform. The interface displays the Unreal Engine 5.1 viewport with a UAV navigating a 3D environment, alongside the audio processing panel showing active dual-channel waveform monitoring and network data transmission status via UDP. The Chinese in the picture is the software parameters.

2.6. Visual Simulation Component

The visual simulation system employs physically accurate rendering techniques to achieve realistic environmental simulation. This system facilitates the modeling of diverse scenarios including urban landscapes, natural terrains, and industrial complexes, with precise simulation of diurnal lighting variations, dynamic weather effects, and complex occlusion relationships. The camera simulation module provides multispectral imaging capabilities encompassing visible light, depth, infrared, and semantic segmentation modalities, while supporting configuration of critical optical parameters such as field of view, focal length, and lens distortion. The real-time rendering engine maintains stable high-frame-rate performance even in complex scenes, and enables synchronous multi-sensor data acquisition through standardized interfaces. This platform significantly enhances the development and validation efficiency of UAV environmental perception, object recognition, and autonomous navigation algorithms, while substantially reducing the cost and duration of physical testing. In particular, atmospheric fog is implemented using the engine's Exponential Height Fog component, whose rendering follows the Beer–Lambert extinction law. The Fog Density parameter controls the optical extinction coefficient per

unit distance, enabling the reproduction of visibility conditions ranging from light haze to dense fog.

Collectively, these design choices explicitly target the Sim2Real gap across both modalities. On the visual side, the physically based rendering pipeline, combined with the Beer–Lambert fog model and complex 3D scene layouts, reproduces the material reflections, lighting variations, and atmospheric degradation that challenge real-world detectors. On the acoustic side, the Sound Attenuation asset models distance-dependent absorption, the Reverb Effect asset reproduces environmental reverberation, and configurable ambient noise profiles allow SNR conditions to be matched to physical deployment sites. The quantitative validation in Section 4 confirms that these combined measures effectively narrow the domain gap to an operationally acceptable level. The theoretical foundations for atmospheric sound attenuation in outdoor propagation scenarios are well established in the literature [36].

3. Sound Source Localization Based on Time Delay Estimation

In modern sound source localization technology, TDOA-based localization algorithms are widely applied in fields such as acoustics, surveillance, and robotic navigation due to their efficiency and precision [37]. The core principle of TDOA algorithms involves utilizing the time differences of the same sound source signal received by two or more microphones to estimate the sound source's position through geometric relationships, as illustrated in Figure 5.

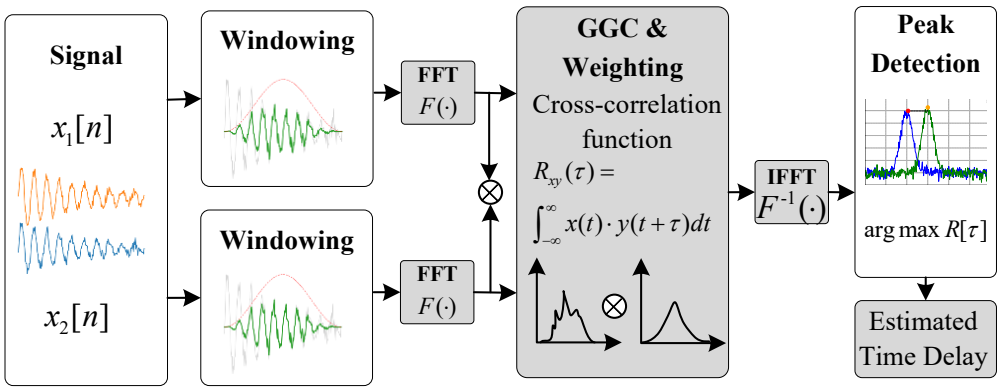


Figure 5. Schematic of time delay estimation using a dual-microphone array. The diagram depicts the geometric relationship between the sound source, the two spatially separated microphones (M_1 and M_2), and the resulting path length difference (Δd) that gives rise to the time difference of arrival (Δt).

3.1. Time Delay Estimation

In the realm of passive acoustic localization, the TDOA serves as a fundamental parameter for estimating the direction and position of a sound source. By analyzing the propagation delay between spatially separated sensors, the geometric locus of the source can be constrained to a hyperboloid. In practical engineering applications, the cross-correlation function is rigorously employed to estimate this time delay. This statistical method quantifies the similarity between two distinct signal streams as a function of the temporal lag applied to one of them, thereby identifying the precise moment of maximum coherence.

Mathematically, the cross-correlation function, denoted as $R_{xy}(\tau)$, measures the displacement of one signal relative to another by integrating their product over time. Given two continuous-time signals $x(t)$ and $y(t)$ received by a sensor pair, the cross-correlation function with respect to the time delay variable τ is defined as:

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} x(t) \cdot y(t+\tau)\, dt. \quad (1)$$

This integral represents a "sliding inner product", where the correlation value indicates the degree of waveform alignment at a specific lag τ. The function $R_{xy}(\tau)$ exhibits a global maximum when the lag τ compensates for the physical propagation delay Δt between the sensors.

Under the assumption of an ideal anechoic environment where $x(t)$ and $y(t)$ originate from a single stationary sound source, the signal received by the second microphone, $y(t)$, can be modeled as a time-shifted and attenuated version of the reference signal $x(t)$. Neglecting amplitude attenuation for the purpose of delay estimation, the relationship is expressed as:

$$y(t) = x(t - \Delta t). \tag{2}$$

Consequently, the objective of the TDOA estimation algorithm is to identify the lag value that maximizes the signal energy overlap. The estimated time delay Δt corresponds to the argument of the peak of the cross-correlation function:

$$\Delta t = \arg\max_{\tau} R_{xy}(\tau). \tag{3}$$

To enable efficient real-time processing, the GCC method [38] exploits the frequency domain. By leveraging the Convolution Theorem, the time-domain cross-correlation is equivalently expressed as the Inverse Fourier Transform of the Cross-Power Spectrum:

$$R_{xy}(\tau) = \mathcal{F}^{-1}(X(f) \cdot Y^*(f)) \tag{4}$$

where $\mathcal{F}^{-1}$ denotes the Inverse Fourier Transform operator, and $X(f)$ and $Y(f)$ represent the frequency spectra of the signals $x(t)$ and $y(t)$, respectively. The term $Y^*(f)$ denotes the complex conjugate of $Y(f)$, ensuring that the phase difference between the signals is correctly captured in the spectral domain. The geometric basis for the subsequent DOA derivation is illustrated in Figure 6.

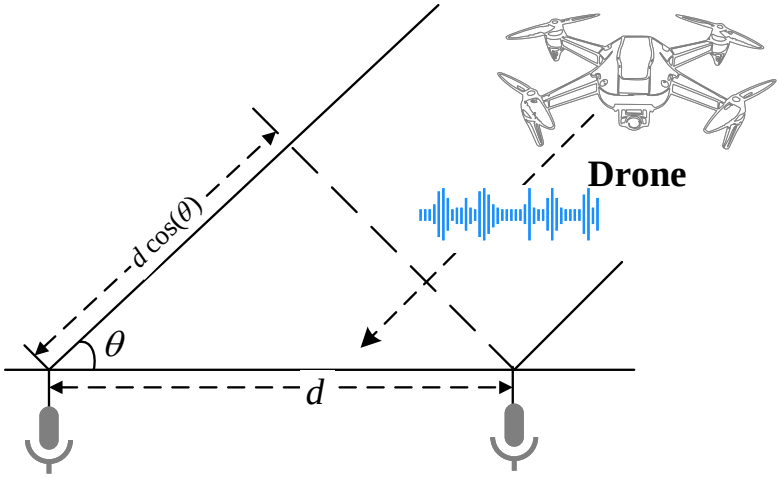


Figure 6. Geometric derivation of the Direction of Arrival (DOA) angle. Under the far-field plane-wave approximation, the azimuth angle θ is computed from the projection of the baseline distance d onto the propagation direction, yielding the closed-form relationship $\Delta d = d\cos\theta$.

3.2. TDOA-Based Acoustic Sensing Principle

To mathematically model the acoustic perception capability within the simulation platform, we adopt a physics-based Time Difference of Arrival (TDOA) framework tailored for a dual-channel sensor configuration. The spatial origin of a sound source is estimated from the temporal disparity of signals reaching two spatially separated sensors.

Let the acoustic source be denoted as S, located at an unknown position in the Euclidean space. The platform is equipped with a linear array consisting of two microphones, M_1 and M_2, separated by a fixed baseline distance d. Assuming the propagation medium is homogeneous with a constant speed of sound c (typically 343 m/s in standard atmospheric conditions), the absolute Time of Arrival (TOA) of the wavefront at the two microphones is denoted as t_1 and t_2, respectively.

In passive detection scenarios, the absolute emission time t_0 of the non-cooperative target is unknown, rendering absolute distance measurement impossible. However, the relative time difference, or TDOA (Δt), can be precisely extracted. Mathematically, this is defined as:

$$\Delta t = t_2 - t_1. \tag{5}$$

Based on the kinematics of sound propagation, this temporal delay corresponds to a path length difference Δd, which represents the additional distance the acoustic wavefront must travel to reach the farther sensor:

$$\Delta d = c \cdot \Delta t. \tag{6}$$

This relationship serves as the fundamental physical constraint used by the simulation engine to generate physically consistent acoustic data, ensuring that the simulated signals strictly adhere to the geometric laws of wave propagation.

3.3. DOA Estimation

The system focuses on estimating the DOA, providing a directional vector that guides the visual sensor's field of view.

To derive the azimuth angle, we employ the far-field approximation. Since the distance between the UAV target and the sensor array is significantly larger than the baseline d, the acoustic rays reaching M_1 and M_2 can be approximated as parallel plane waves.

Under this geometric assumption, the path length difference Δd corresponds to the projection of the baseline d onto the direction of propagation. Let θ denote the azimuth angle of the sound source relative to the array's axis. The geometric relationship is governed by the law of cosines:

$$\Delta d = d \cdot \cos\theta. \tag{7}$$

In practice, the raw dual-channel audio contains broadband spectral noise from wind turbulence, motor harmonics, and sensor self-noise. To ensure robust TDOA estimation, two signal conditioning steps are applied. First, a fourth-order Butterworth band-pass filter with passband [200 Hz, 4000 Hz] is applied to each channel to reject low-frequency and high-frequency noise, improving the SNR within the frequency band of interest. The time delay Δt is then estimated via the Generalized Cross-Correlation with Phase Transform (GCC-PHAT):

$$R_{12}(\tau) = \int_{-\infty}^{+\infty} \frac{X_1(f)\, X_2^*(f)}{|X_1(f)\, X_2^*(f)|}\, e^{j2\pi f\tau}\, df \tag{8}$$

where $X_1(f)$ and $X_2(f)$ are the Fourier transforms of the filtered channels. The PHAT weighting whitens the cross-power spectrum, suppressing narrowband noise and reverberation. The TDOA is obtained as $\Delta t = \arg\max_\tau R_{12}(\tau)$.

The GCC-PHAT estimator maintains reliable performance for in-band SNR above approximately -5 dB, i.e., even when the ambient noise power moderately exceeds that of the target signal. In typical outdoor UAV detection scenarios, the in-band SNR is well above this threshold. Moreover, in the audio–visual fusion framework, when the acoustic

SNR falls below the effective boundary, the system prioritizes visual detections and the acoustic branch serves only as a coarse directional cue for camera steering.

Substituting the TDOA relationship derived in Equation (8), we obtain the closed-form solution for the angle of arrival:

$$\theta = \cos^{-1}\left(\frac{c \cdot \Delta t}{d}\right). \tag{9}$$

Here, the estimated angle θ is constrained to the range $[0^\circ, 180^\circ]$. It is important to note that a linear dual-microphone array inherently exhibits front-back ambiguity. In the context of our audio–visual fusion framework, this ambiguity is resolved either by the directional constraints of the camera's FOV or by the platform's kinematic maneuvers, which effectively synthesize a larger virtual aperture over time.

4. Experimental Setup and Results

To systematically validate the proposed platform and fusion strategy, four experiments are conducted in a paired Sim2Real structure:

1. **Acoustic Localization—Simulation (Section 4.1):** A quadrotor UAV executes uniform circular motion around a virtual sensor array within the UE5.1 environment. This experiment evaluates the TDOA algorithm's angular tracking accuracy across a continuous azimuthal sweep using purely simulated acoustic data.
2. **Acoustic Localization—Physical (Section 4.2):** Physical quadrotor UAVs are deployed on tripods, with one serving as the sound source and the other hosting the microphone array. The source is repositioned at discrete angular intervals to validate the simulation-derived localization performance against real-world acoustic propagation and hardware noise.
3. **Audio–Visual Fusion—Simulation (Section 4.3):** A dynamic tracking scenario is constructed in a simulated foggy woodland, where atmospheric scattering intermittently degrades visual detection. The confidence-gated fusion controller automatically switches between visual and acoustic modalities, demonstrating continuous target tracking under severe visibility loss.
4. **Audio–Visual Fusion—Physical (Section 4.4):** The fusion experiment is replicated in a real indoor environment using physical occlusions (columns and whiteboards). A D435i camera and microphone array are fused on physical hardware to verify that the cooperative switching behavior observed in simulation transfers reliably to the real world.

Experiments 1–2 isolate the acoustic subsystem to characterize its intrinsic accuracy, while Experiments 3–4 evaluate the full audio–visual pipeline. Each simulation–physical pair enables direct Sim2Real comparison under matched conditions.

4.1. Sound Source Localization Simulation

The experimental setup is initialized by configuring the platform's network listening interface, which is bound to a specific IP address and port to receive the raw dual-channel audio stream from the simulation environment at a sampling rate of $f_s = 48{,}000$ Hz. Since the raw audio stream typically arrives in an interleaved format, a pre-processing stage is implemented to decouple the stream into independent left and right channel vectors. This is achieved by extracting alternating samples, which is a prerequisite for computing the cross-correlation function and the subsequent time delay τ. To guarantee robustness and real-time performance, a concurrent multi-threading architecture is adopted. A "Producer-Consumer" model is utilized where the data reception thread continuously populates a circular ring buffer (set to a capacity of 0.2 MB), while the processing thread simultaneously

retrieves data frames for the algorithm. This decoupling mechanism prevents data packet loss caused by algorithmic computation delays, ensuring a continuous and audio stream, as depicted in Figure 7.

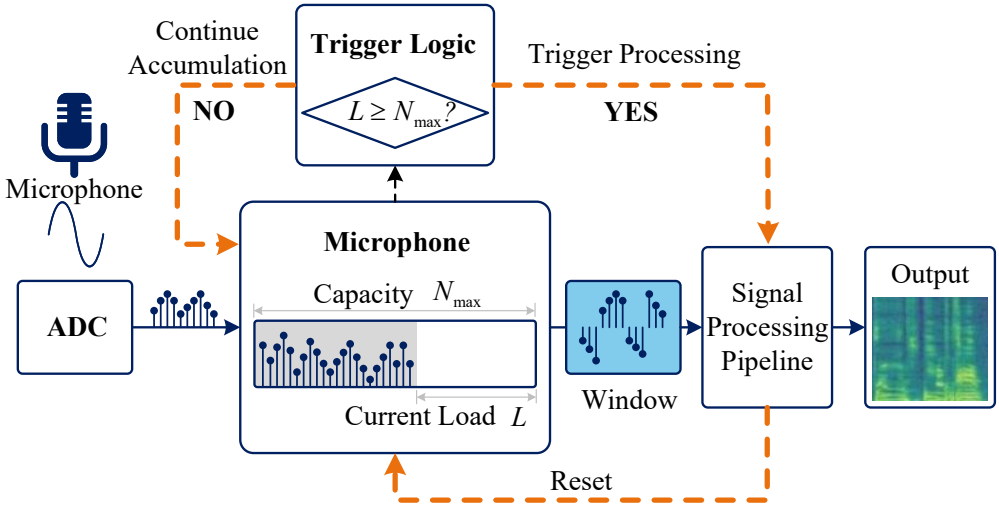


Figure 7. Schematic of the concurrent data buffer architecture. The reception thread continuously writes incoming audio frames into a circular ring buffer, while the processing thread asynchronously retrieves buffered data for real-time TDOA computation, thereby decoupling data acquisition from algorithmic processing to prevent packet loss.

To rigorously validate the localization fidelity, a dynamic trajectory scenario is constructed within the simulation platform. A quadrotor UAV is assigned a pre-planned trajectory of UCM around the sensor array, which acts as the geometric center of the trajectory. This setup allows for a comprehensive evaluation of the algorithm's performance across a continuous range of azimuth angles. As illustrated in Figure 8, the scenario visualization displays the sensor location at the apex of the central cone, with the real-time ground truth azimuth of the UAV displayed via the blue numerical indicator in the upper-left overlay. The specific experimental parameters are configured as follows: the simulated microphone array baseline (spacing) is fixed at $d = 0.5$ m to balance spatial resolution and phase ambiguity; the UAV's angular velocity is stabilized at $\omega = 2°/s$ to minimize Doppler shift interference; and the audio processing window length is set to 0.5 s per iteration to achieve a balance between temporal responsiveness and frequency domain resolution. The simulation engine host is configured with the loopback IP address 127.0.0.1, listening on port 28003.

The estimated trajectory results derived from the TDOA algorithm are presented in Figure 9a. The comparative analysis demonstrates that the calculated azimuth angle exhibits a high degree of correlation with the UAV's actual kinematic trend, confirming the algorithm's ability to track moving sources within the virtual environment. However, a distinct geometric constraint is observed due to the linear topology of the dual-microphone array. The system exhibits inherent "front-back ambiguity," constraining the unique solution space to the range of $[0°, 180°]$. Consequently, when the UAV traverses the rear half-plane (actual angle $\theta > 180°$), the algorithm correctly identifies the time delay magnitude but maps the result to its symmetric conjugate in the front half-plane. This results in the calculated angle appearing as the supplementary angle (or symmetric mirror) of the true azimuth. It should be noted that the dual-microphone configuration is a deliberate design trade-off favoring lightweight on-board deployment over full azimuthal coverage. In practice, this front-back ambiguity is resolved cooperatively within the fusion pipeline. When the visual detector is operative, the camera-based bearing provides an unambiguous

full-circle reference, allowing the system to select the TDOA solution (either θ or $180^\circ - \theta$) that is geometrically consistent with the visual detection sector, effectively using vision as a real-time arbiter for hemisphere selection. In situations where visual detection becomes unavailable due to occlusion or fog, the fusion controller commands the platform to execute a deliberate rotational maneuver; the resulting change in the observed DOA—specifically, whether the estimated angle increases or decreases relative to the known rotation direction—uniquely identifies the true hemisphere of the sound source, thereby resolving the mirror ambiguity without visual input. This cooperative disambiguation strategy ensures that the $[0^\circ,\ 180^\circ]$ geometric constraint does not degrade the system's operational coverage in practice. The quantitative performance is further detailed in Figure 9b, which plots the error curve between the estimated angle (corrected for symmetry) and the ground truth. Crucially, visual detection resolves the ambiguity through both positive and negative confirmation. For instance, if the TDOA estimate indicates 30°, and the visual detector identifies the target within the camera FOV, the estimate is directly confirmed. Conversely, if no target is observed despite an active TDOA reading, the system infers that the source lies in the rear hemisphere, as the absence of a visual detection within the forward-facing FOV constitutes conclusive negative evidence. Under conditions where visual detection fails entirely, the system enters an acoustic-guided rotational search mode, and the resulting change in the TDOA-estimated angle uniquely identifies the true hemisphere.

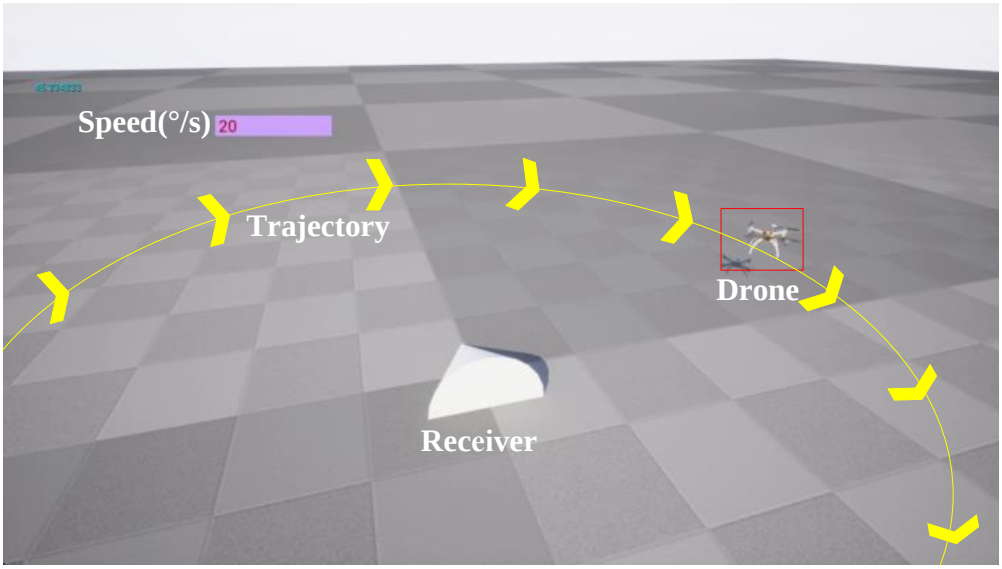
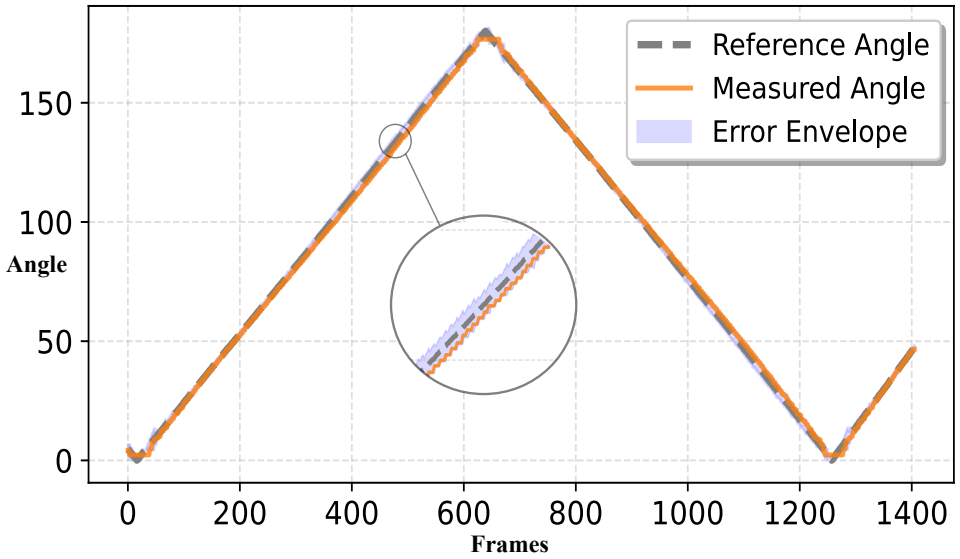
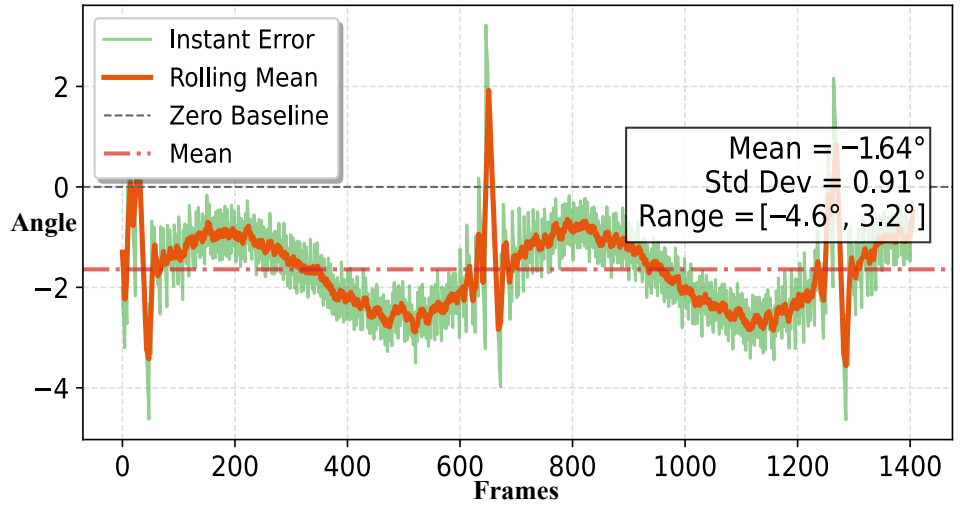


Figure 8. Simulation scenario for the acoustic localization experiment. A quadrotor UAV executes uniform circular motion around the central sensor array. The blue numerical indicator in the upper-left overlay displays the real-time ground-truth azimuth.

By deploying the TDOA-based sound source localization algorithm within the acoustic simulation platform, the relative azimuth of the quadrotor UAV was successfully computed and verified. The results effectively validate the effectiveness of the proposed simulation platform for developing and testing UAV acoustic perception algorithms. Through a rigorous post hoc error analysis, the reliability of the platform has been confirmed, with the observed deviations attributed to a confluence of algorithmic and systemic factors. Specifically, the inherent discrete-time sampling imposes a quantization limit on the TDOA resolution, which inevitably propagates into the angular estimation accuracy. Simultaneously, unavoidable simulated hardware latencies—including buffer jitter and asynchronous data transfer delays—introduce minor temporal misalignments. These artifacts effectively mirror real-world sensor imperfections, thereby further validating the platform's fidelity in replicating practical deployment challenges.



(**a**) Localization Angle



(**b**) Angle Error

Figure 9. Sound source localization results from the simulation experiment. (**a**) Comparison of the TDOA-estimated azimuth trajectory against the ground-truth angle during uniform circular motion, illustrating the front-back ambiguity inherent to the linear dual-microphone array. (**b**) Corresponding angular error over time.

4.2. Sound Source Localization Physical Experiment

As illustrated in Figure 10, the experimental setup consists of two quadrotor UAVs, each mounted on a tripod at a height of 1 m, separated by 2 m. UAV A is equipped with a linear dual-microphone array (baseline $d = 0.5$ m), oriented horizontally along the inter-UAV axis. Both channels are acquired through a single dual-channel audio interface at $f_s = 48{,}000$ Hz, guaranteeing inter-channel synchronization with zero relative time offset. UAV B serves as the sound source with its rotors operating normally to generate realistic in-flight acoustic emissions, while UAV A's rotors remain stationary to isolate the received signal from self-noise. The microphone array is mounted on the physical UAV airframe—rather than on a standalone tripod—because the UAV's onboard flight controller is connected to the optical motion capture system via the local-area network, enabling time-synchronized acquisition of ground-truth position data. The tripod mounting eliminates translational and rotational motion, thereby isolating the acoustic localization performance

from flight dynamics uncertainties in a controlled setting. The relative angular position of UAV B is calculated using the GCC-PHAT TDOA algorithm described in Section 3. In all physical experiments, the ground-truth azimuth was obtained using an optical motion capture system with a positional accuracy of $\delta = 0.1$–0.5 mm. The angular uncertainty of the reference measurement was derived as $\delta_\theta = \arctan(\delta/d)$, where d is the source-to-array distance. At the experimental baseline of $d = 2$ m and worst-case positional error $\delta = 0.5$ mm, this yields $\delta_\theta \approx 0.014°$, which is two orders of magnitude below the reported localization RMSE and therefore has a negligible impact on the evaluation results. During the experiment, the sound source UAV B is moved to different angular positions at 20° intervals. UAV A, equipped with the dual-microphone array, receives the signals emitted by the sound source UAV and uses the collected data to calculate the angular position of the sound source.

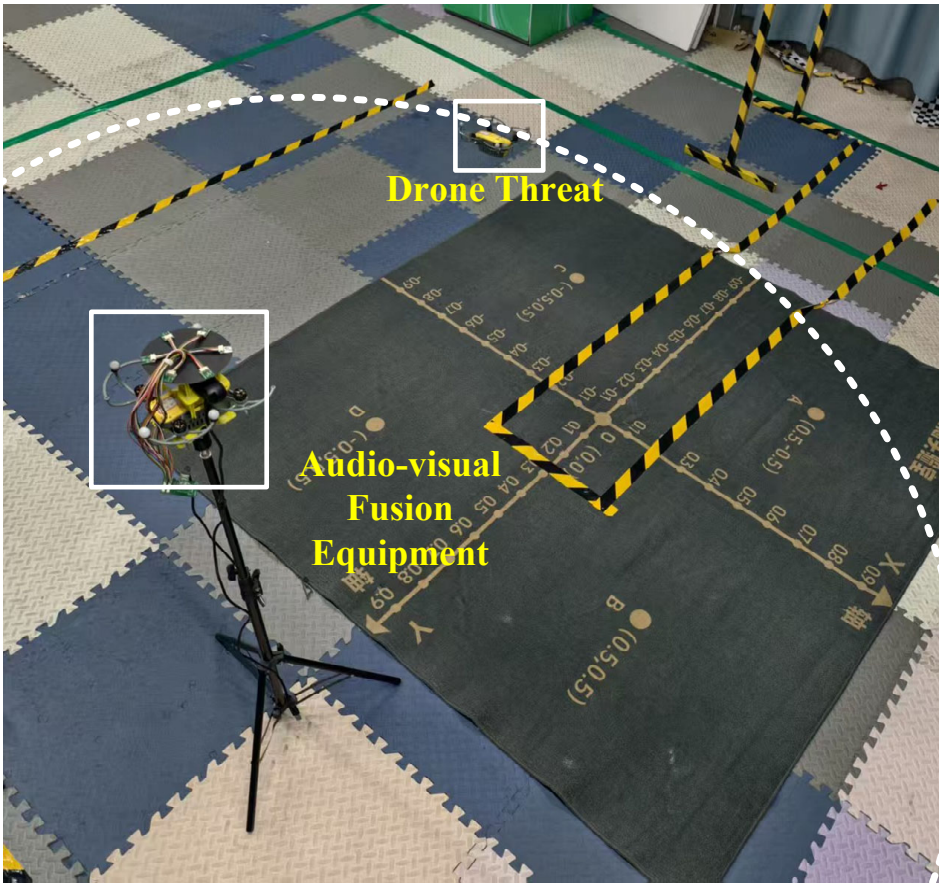
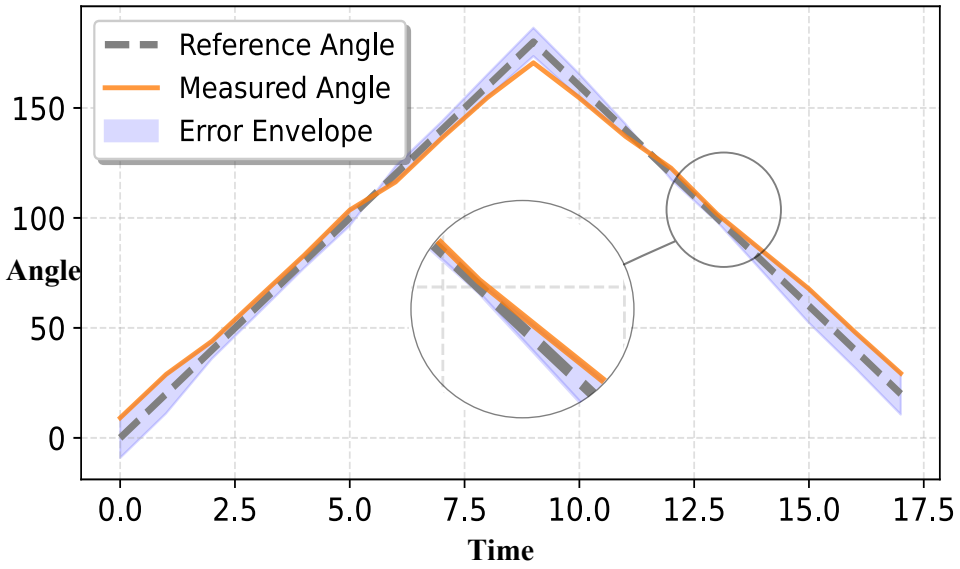


Figure 10. Schematic diagram of the physical experimental setup. Two quadrotor UAVs are mounted on tripods at a height of 1 m and separated by 2 m. UAV A is equipped with a dual-microphone sensor array for signal reception, while UAV B serves as the sound source with active rotors.

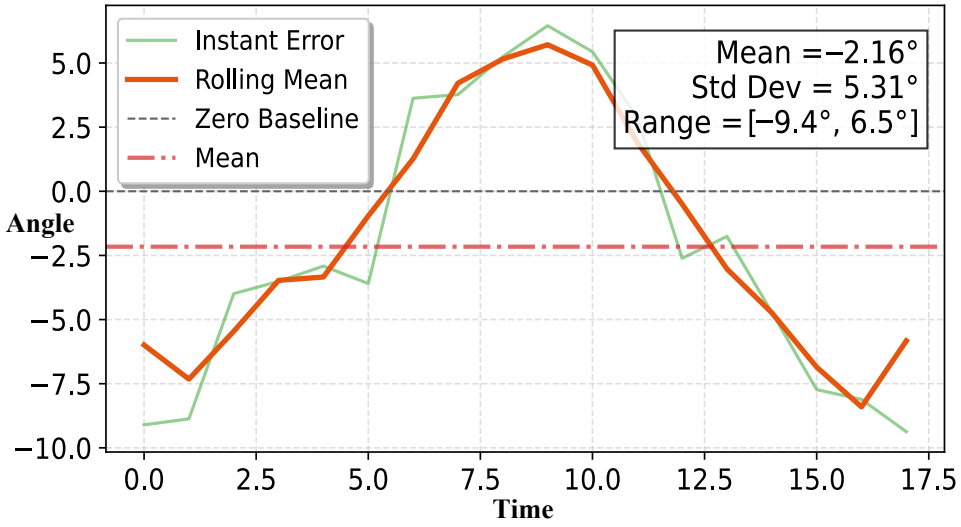
Figure 11a demonstrates the performance of the sound source localization algorithm in practical applications. As can be seen from the figure, the algorithm maintains high accuracy across most angular ranges. However, significant increases in angular error are observed near 0° and 180°. This phenomenon is primarily attributed to the geometric constraint of the linear dual-microphone array: at these end-fire directions, the cosine function in Equation (9) approaches its extrema, where the derivative $|\partial\theta/\partial(\Delta t)|$ tends to infinity, causing small TDOA estimation errors to be amplified into large angular deviations. Formally, error propagation of the DOA estimator $\theta = \arccos(c\,\Delta t/d)$ yields:

$$\delta\theta \approx \frac{c}{d} \cdot \frac{\delta(\Delta t)}{|\sin\theta|} \tag{10}$$

As $\theta \to 0°$ or $180°$, the denominator $|\sin\theta| \to 0$, causing any finite TDOA estimation error $\delta(\Delta t)$ to be amplified without bound. This divergence is an inherent mathematical property of linear array geometry and is independent of rotor interference. As shown in Figure 11b, the error distribution confirms that this boundary effect is the dominant source of localization inaccuracy, highlighting a well-known limitation of linear array geometries in practical deployment scenarios.



(**a**) Localization Angle



(**b**) Angle Error

Figure 11. Physical UAV sound source localization results. (**a**) Estimated azimuth angle versus ground-truth across discrete angular positions at 20° intervals. (**b**) Angular error distribution at each measured position.

4.3. Acoustic-Visual Simulation Experiment

This experiment designs a UAV cooperative detection system tailored for complex meteorological conditions, enhancing environmental adaptability by fusing visual and acoustic sensing modalities. In the simulated foggy woodland scenario, the visual subsystem experiences significant degradation in imaging performance due to atmospheric scattering effects. To address this, the system establishes a detection mechanism based on complementary heterogeneous sensors: the visual perception subsystem is constructed

using the YOLO object detection algorithm, and when the visual detection confidence falls below a preset threshold, the acoustic perception subsystem is automatically activated, compensating for the limitations of visual perception. Experimental validation demonstrates that this fusion strategy effectively overcomes the environmental sensitivity of single sensors, significantly improving the continuous detection capability of UAV targets under dense fog conditions.

Formally, let $c_t \in [0,\ 1]$ denote the YOLO detection confidence score at frame t, and let τ denote the switching threshold. The fusion controller operates as a binary state machine with two states: Visual-Active (S_V) and Acoustic-Active (S_A). The state transition and output selection rules are defined as:

$$S_t = \begin{cases} S_V, & \text{if } c_t \geq \tau \\ S_A, & \text{if } c_t < \tau \end{cases} \tag{11}$$

$$\hat{\theta}_t = \begin{cases} \theta_t^{\text{vis}}, & \text{if } S_t = S_V \\ \theta_t^{\text{aco}}, & \text{if } S_t = S_A \end{cases} \tag{12}$$

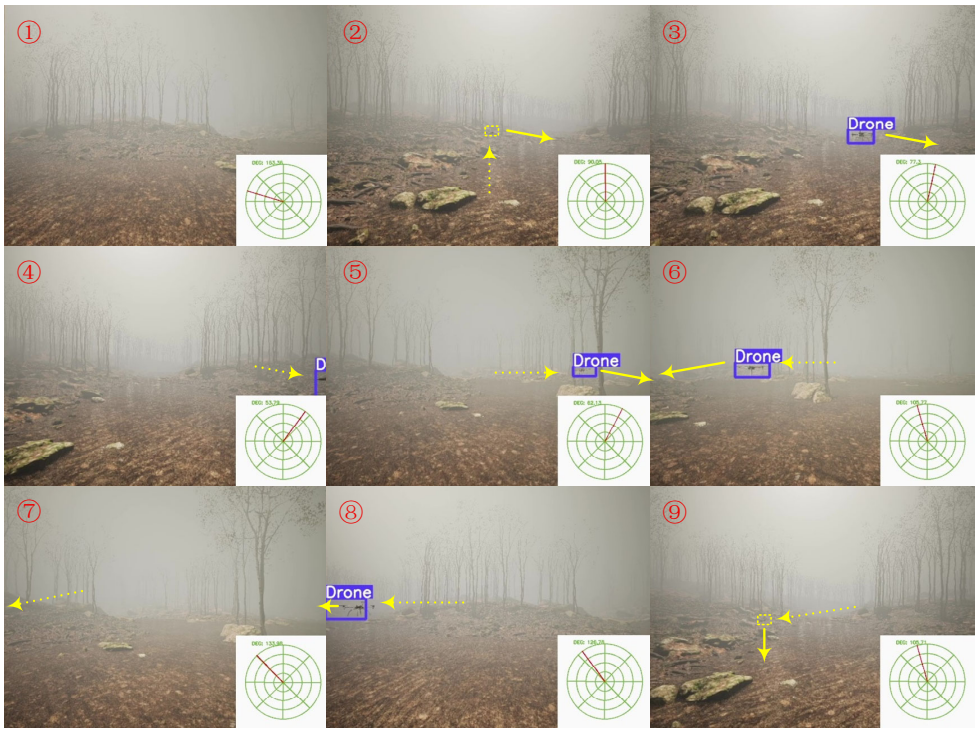
where θ_t^{vis} is the visual bearing derived from the YOLO bounding box center, θ_t^{aco} is the TDOA-based acoustic bearing, and $\hat{\theta}_t$ is the fused output. In the experiments presented in this work, the threshold is set to $\tau = 0.6$.

Regarding sensitivity to this hyperparameter: setting τ too high (e.g., $\tau > 0.7$) causes the system to distrust valid but lower-confidence visual detections, leading to excessive switching into acoustic mode; the resulting frequent mode transitions introduce jitter at the switching boundaries and degrade overall angular precision, as acoustic estimates carry inherently higher noise. Conversely, setting τ too low (e.g., $\tau < 0.3$) causes the system to over-rely on the visual channel, admitting low-confidence detections that may correspond to false positives (e.g., background clutter misidentified as a UAV), thereby introducing erroneous bearing estimates. The value $\tau = 0.6$ was empirically selected to balance these two failure modes, providing stable state transitions while maintaining reliable visual detections.

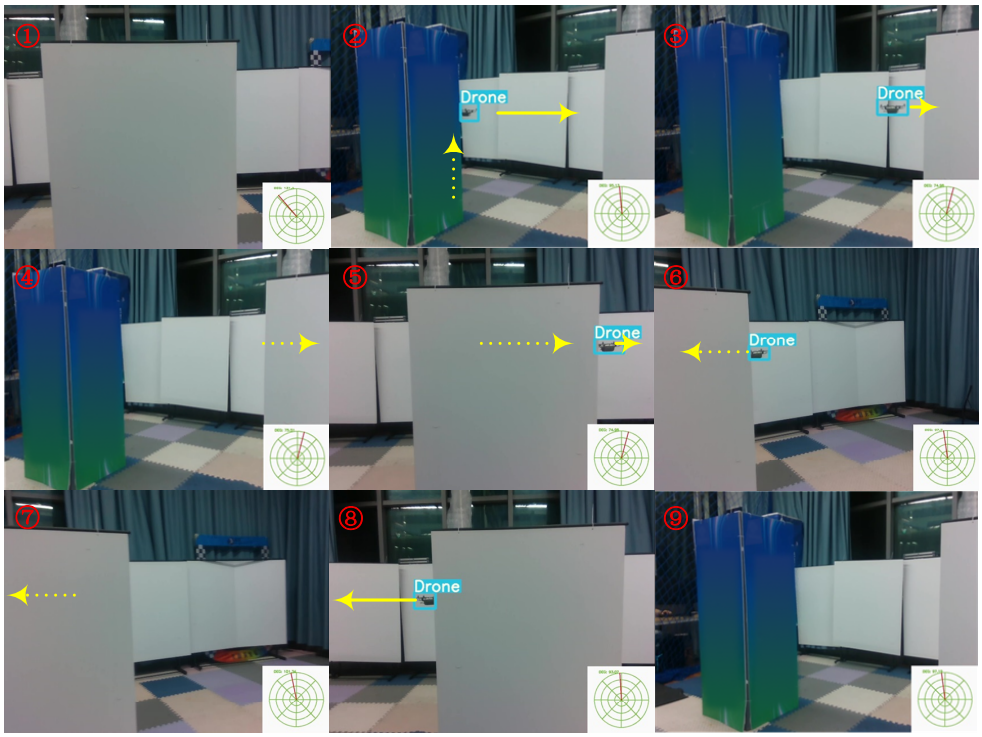
Figure 12a illustrates the dynamic detection process of the simulation experiment: when the target UAV exceeds the visual visibility range, the system automatically activates the acoustic sensor for azimuth detection; when the target enters the visible range, it switches to visual detection mode. In Figure 13, through the spatiotemporal correspondence between the acoustic localization trajectory and the visual detection period (highlighted by shaded areas in the figure), visually reveals the cooperative working mechanism of the dual-modal sensors. Experimental data indicate that in foggy environments with severe visibility fluctuations, the alternating activation and information complementarity of the acoustic-visual sensors effectively maintain the continuity of target tracking, validating the adaptability of the multi-modal fusion strategy to adverse meteorological conditions.

4.4. Acoustic-Visual Physical Experiment

This experiment successfully replicated the detection process of Experiment 3 in a real-world scenario. By fusing sensor data from the microphone array and the D435i camera, a complementary detection mechanism was constructed, and a motion compensation system was employed for path planning of the target UAV. The experimental setup included two types of obstructions—vertical columns and whiteboards—positioned at the starting point and along the flight path of the UAV to simulate detection challenges in complex environments.

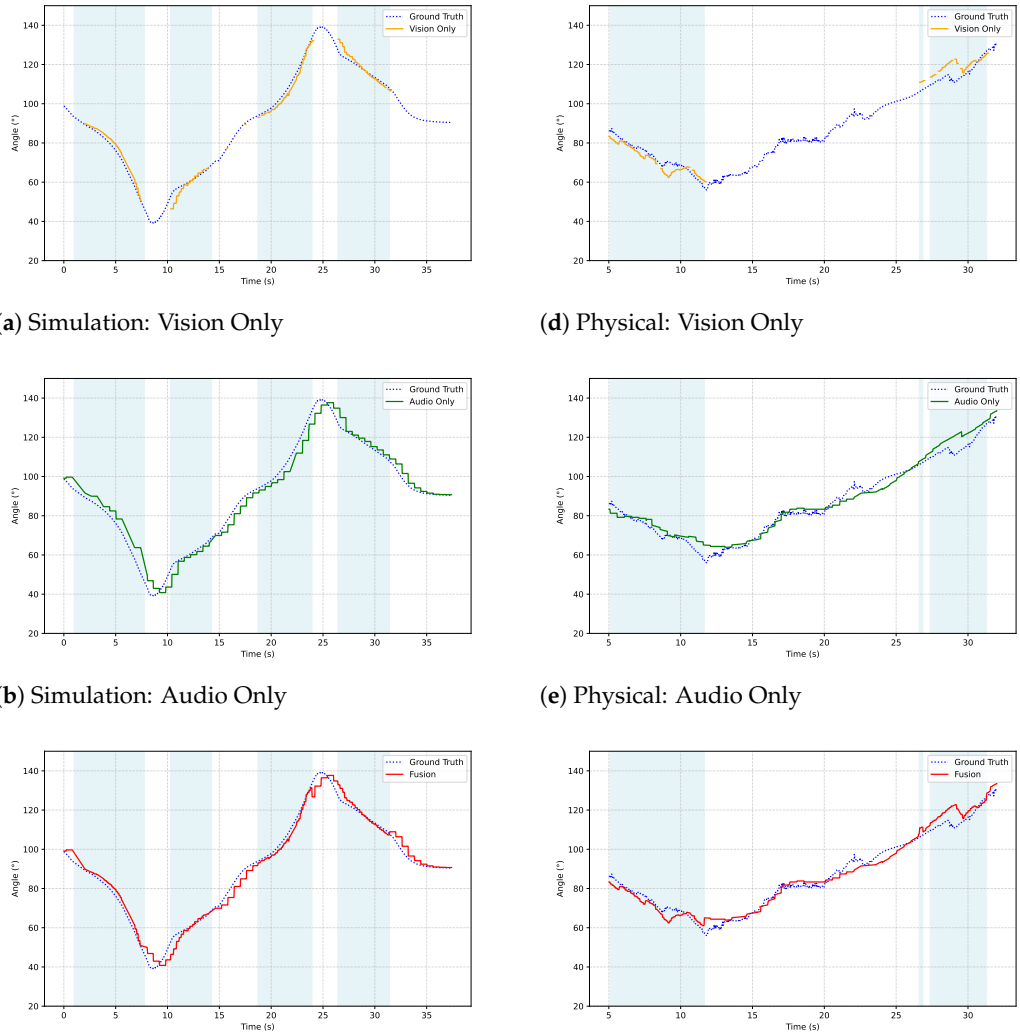


(**a**)



(**b**)

Figure 12. Audio–visual fusion UAV localization. (**a**) Simulation of a low-visibility foggy woodland scenario. The blue box indicates visual detection, while the red pointer displays TDOA-based acoustic azimuth. Yellow lines denote trajectory and prediction. Sequence 1–9 records the engagement: emerging, maneuvering, and receding. The dashed arrow indicates the situation when it is blocked. (**b**) Experimental setup with display board obstacles for evaluating perception under physical occlusion. Frames 1–9 illustrate the sequence: emerging from occlusion, maneuvering, and receding.



(**a**) Simulation: Vision Only

(**d**) Physical: Vision Only

(**b**) Simulation: Audio Only

(**e**) Physical: Audio Only

(**c**) Simulation: Audio–Visual Fusion

(**f**) Physical: Audio–Visual Fusion

Figure 13. Modality ablation comparison of tracking performance. (**a**–**c**): Simulation under foggy woodland conditions. (**d**–**f**): Physical experiment under occlusion. Light blue shading indicates intervals where visual detection was active.

As shown in Figure 12b, during the dynamic detection process, when the target UAV is obstructed, the system automatically activates the auditory subsystem to assist in steering, thereby ensuring the continuity and stability of the detection process. This mechanism effectively addresses the limitations of single-sensor detection in complex environments, significantly enhancing the system's robustness. Experimental results demonstrate that the combined acoustic-visual sensor approach can adapt to various environmental conditions and assist the visual system in achieving precise localization when visual information is insufficient. Furthermore, the consistency between simulation and real-world experimental results indicates that simulation experiments hold significant reference value and can provide reliable theoretical foundations for the design and optimization of practical systems. As shown in Table 1, in Experiment 3 and Experiment 4 where only vision was active, the proportion of time during which the UAV was detected was relatively low due to environmental obstructions or foggy conditions. To compensate for this impact, auditory assistance was introduced, albeit with a slight reduction in accuracy. It is important to recognize the fundamental role of the acoustic modality within the proposed system. Rather than serving to significantly enhance absolute tracking precision—a domain where visual sensors naturally excel—the acoustic sensor acts as a critical auxiliary modality. Its primary

value lies in generating reliable directional cues for initial target discovery and providing a robust fallback for continuous tracking when the visual target is lost due to occlusion or severe weather. Specifically, the lower RMSE of the vision-only system (2.61° in simulation) is computed exclusively over intervals where the visual detector was active (approximately 52% of the duration), while the target was entirely untracked during the remaining periods. In contrast, the audio–visual fusion system leverages acoustic steering to extend tracking coverage to near-continuous operation. Therefore, the mathematically higher RMSE of the fusion system does not indicate inferior performance; instead, it underscores the systemic reliability gained by incorporating acoustic sensing to maintain target awareness during periods of critical visual failure.

Table 1. Modality Ablation Study: Simulation and Physical Experiment Results.

Experiment	RMSE (°)	SD (°)	Tracking Coverage (%)
Simulation (vision-only)	2.61°	1.73°	51.56
Simulation (audio-only)	4.11°	4.11°	100.00
Simulation (audio–visual)	3.28°	3.10°	100.00
Physical (vision-only)	3.93°	2.70°	39.82
Physical (audio-only)	3.66°	3.43°	100.00
Physical (audio–visual)	3.50°	2.83°	100.00

To quantify the platform's reliability, we analyze the consistency between the simulated and physical domains. As detailed in Table 1, the system exhibits isomorphic behavioral trends in both environments. Although the physical experiments show a slightly higher RMSE, this performance degradation is within an acceptable margin. This reality gap is primarily attributed to unmodeled stochastic factors, such as wind gusts affecting the physical drone's stability and non-linear hardware noise in the commercial microphone array. Nevertheless, the fact that the fusion strategy successfully triggered and maintained tracking in both domains validates the platform's Sim2Real transfer capability.

The ablation results in Table 1 and Figure 13 quantitatively reveal the contribution of each modality. The vision-only configuration achieves the lowest RMSE (2.61° in simulation, 3.93° in physical) but provides tracking coverage for only approximately 52% and 40% of the total duration, respectively, due to environmental obstructions and foggy conditions. The audio-only configuration maintains 100% temporal coverage but exhibits higher angular noise (RMSE of 4.11° in simulation, 3.66° in physical). The audio–visual fusion system combines the complementary strengths of both modalities, achieving 100% tracking coverage while maintaining overall RMSE values (3.28° and 3.50°) that are comparable to or lower than those of audio-only operation. These results confirm that each modality provides an indispensable contribution to the system and that the fusion strategy effectively leverages their complementarity.

To substantiate the Sim2Real consistency beyond point-estimate comparisons, we conducted a suite of statistical analyses on the angular error distributions, summarized in Table 2.

Bootstrap resampling ($N = 10{,}000$) confirms narrow confidence intervals (all widths $< 0.5°$), demonstrating statistical stability. Cohen's d effect size indicates that the acoustic Sim2Real gap is negligible ($d = 0.14$), while the visual gap is moderate ($d = 0.64$), attributable to illumination and texture variations not fully replicated by the rendering engine. A percentile-matched Pearson correlation ($r > 0.99$ for both channels) further confirms that the error distribution shapes are nearly identical across domains.

Table 2. Statistical validation of Sim2Real consistency.

Metric	Acoustic	Visual
RMSE	4.11°	3.93°
CI	[3.98°, 4.24°]	[3.73°, 4.13°]
Cohen's *d*	0.14	0.64
Pearson *r*	0.994	0.990
OVL	77.1%	54.4%

4.5. Comparative Quantitative Assessment

To provide a thorough evaluation, we conducted an extensive comparative assessment by introducing five recognized baseline methods—spanning solitary modality tracking and various multisensor fusion techniques—resulting in a 6-method quantitative comparison.

The newly added comparison includes:

- **Audio Methods**: GCC-PHAT, MUSIC, and MVDR.
- **Filtering & Fusion Methods**: Moving Average and Kalman Filter.

To ensure a fair and consistent evaluation, all resulting trajectories—including our proposed method—were processed with an identical median filtering step to reduce measurement outliers. The comprehensive results are summarized in Table 3.

Table 3. Quantitative Comparison of 6 Tracking Methods.

Method	RMSE (°)	MAE (°)	SD (°)	Detection Rate $< 10°$ (%)
Audio Baselines				
GCC-PHAT	5.49	4.69	2.85	96.28
MUSIC	6.90	3.11	6.16	98.21
MVDR	6.88	3.08	6.16	98.21
Multi-Sensor Fusion Baselines				
Moving Average	6.31	5.24	3.52	89.45
Kalman Filter	5.89	4.98	3.15	95.04
Proposed Fusion	5.85	4.95	3.11	94.27

The quantitative results in Table 3 and Figure 14 demonstrate that the proposed confidence-gated fusion strategy achieves a stable performance (RMSE 5.85°). It provides comparable accuracy to established, computationally intensive filtering algorithms such as the Kalman Filter (5.89°), while audio-only methods like GCC-PHAT, MUSIC, and MVDR can achieve low tracking errors when localized accurately (e.g., MAE ∼3.1°), they are inherently more susceptible to environmental reverberation, which occasionally produces multi-path localization errors that penalize the RMSE. The comparative assessment confirms that our proposed method can maintain reliable and robust tracking performance without relying on predefined complex kinematic models. This lightweight characteristic makes it particularly suitable for deployment on resource-constrained anti-UAV platforms.

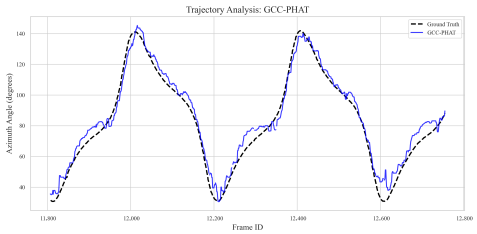
4.6. Discussion

While the proposed platform and fusion strategy demonstrate robust performance, several technical limitations warrant discussion. At the hardware level, the current dual-microphone configuration is restricted to 1D azimuth estimation, lacking the vertical and depth resolution required for full 3D target localization. Furthermore, the microphone baseline is constrained by the physical dimensions of the quadrotor, which inevitably limits the spatial resolution for low-frequency acoustic sources. From an algorithmic perspective, the confidence-gated fusion relies on empirical thresholds which may require adaptive optimization for varying environmental SNR levels. Additionally, the system

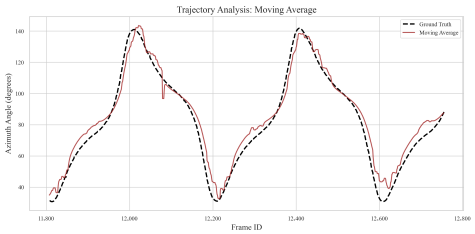
is currently optimized for single-target tracking; robust multi-source separation in dense swarm scenarios remains an open challenge.

Future research will focus on expanding the platform's multi-modal capabilities. This includes modeling dynamic atmospheric constraints (e.g., wind gusts, heavy precipitation) and integrating additional sensors such as LiDAR and Radar to enhance robustness in extreme Sim2Real transitions. We also aim to evolve the fusion logic into a multi-UAV cooperative framework, enabling swarm-level collaborative perception and autonomous navigation in GNSS-denied environments.

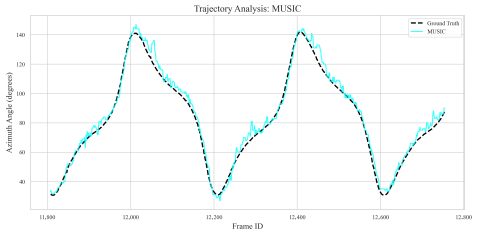
Moreover, a dedicated, component-level validation of the acoustic propagation model—such as comparing simulated versus measured sound pressure levels across distance—was not conducted in this study and remains a direction for future work; however, the end-to-end statistical analysis presented above (Table 2) demonstrates that the acoustic Sim2Real gap is negligible at the task level.



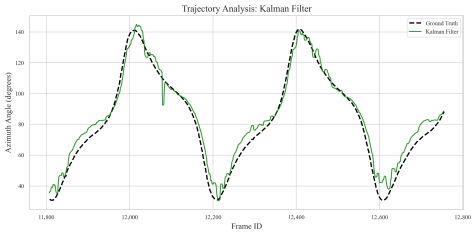
(**a**) Audio: GCC-PHAT



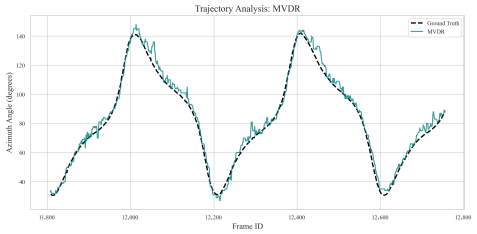
(**d**) Fusion: Moving Average



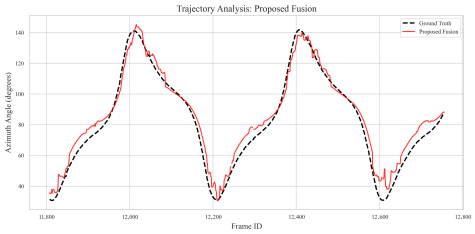
(**b**) Audio: MUSIC



(**e**) Fusion: Kalman Filter



(**c**) Audio: MVDR



(**f**) Fusion: Proposed

Figure 14. Trajectory tracking performance across different baseline methods. The plots are categorized into pure audio high-resolution algorithms and representative multi-sensor fusion algorithms including ours.

5. Conclusions

This paper presents a novel Sim2Real simulation platform designed to provide a physics-informed simulation environment dedicated to UAV acoustic algorithm development. By implementing an advanced acoustic rendering engine that models sound propagation physics, the platform bridges the domain gap between synthetic and physical acoustic environments. Through rigorous cross-validation with physical experiments, we demonstrate not only the platform's effectiveness in implementing TDOA-based sound localization but also the transferability of fusion algorithms to real-world scenarios involv-

ing visual interference. The demonstrated consistency between simulated and physical behaviors establishes this platform as an essential tool for narrowing the reality gap in UAV acoustic sensing research.

The platform extends beyond conventional acoustic simulation by offering unique audio–visual co-simulation capabilities, establishing a unified framework for multi-modal sensor fusion research. This physics-consistent simulation environment enables systematic investigation of cross-modal interaction dynamics, which is particularly valuable for overcoming visual perception degradation. Crucially, the high alignment between simulation and reality significantly reduces development cycles, providing a reliable proxy for algorithm evaluation and ensuring robust deployment to physical hardware with minimal adaptation.

Future work will focus on implementing dynamic weather modeling and moving obstacle interactions to further enhance Sim2Real fidelity under complex atmospheric constraints. Parallel efforts will explore adversarial swarm coordination and GNSS-denied navigation, utilizing the platform to pre-validate robust cross-modal perception models before field deployment, thereby establishing a closed-loop pipeline from virtual training to physical reality. To facilitate reproducibility and community adoption, the simulation scenes and audio rendering scripts have been released at https://github.com/RflySim/RflySimAudio (accessed on 4 March 2026).

Author Contributions: Conceptualization, X.N., H.L. and X.D.; methodology, X.N., H.L. and X.D.; software, X.D.; validation, X.N. and H.L.; formal analysis, X.N., H.L. and X.D.; investigation, X.N. and H.L.; resources, X.D.; data curation, X.N. and H.L.; writing—original draft preparation, X.N.; writing—review and editing, X.N., H.L. and X.D.; visualization, X.N. and H.L.; supervision, X.D.; project administration, X.D.; funding acquisition, X.D. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (Grant No. 62406345), and in part by the Natural Science Foundation of Hunan Province (Grant No. 2025JJ50341).

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Mohsan, S.A.H.; Othman, N.Q.H.; Li, Y.; Alsharif, M.H.; Khan, M.A. Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends. *Intell. Serv. Robot.* **2023**, *16*, 109–137. [CrossRef]
2. Shakhatreh, 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]
3. Jiang, P.; Ergu, D.; Liu, F.; Cai, Y.; Ma, B. A Review of Yolo algorithm developments. *Procedia Comput. Sci.* **2022**, *199*, 1066–1073. [CrossRef]
4. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
5. Gee, T.; James, J.; Van Der Mark, W.; Delmas, P.; Gimel'farb, G. Lidar guided stereo simultaneous localization and mapping (SLAM) for UAV outdoor 3-D scene reconstruction. In *Proceedings of the 2016 International Conference on Image and Vision Computing New Zealand (IVCNZ)*; IEEE: New York, NY, USA, 2016; pp. 1–6.
6. Rosa, J.; Basiri, M. Cooperative Audio-Visual System for Localizing Small Aerial Robots. In *Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*; IEEE: New York, NY, USA, 2019; pp. 6064–6069.
7. Su, Y.; Wang, T.; Shao, S.; Yao, C.; Wang, Z. GR-LOAM: LiDAR-based sensor fusion SLAM for ground robots on complex terrain. *Robot. Auton. Syst.* **2021**, *140*, 103759. [CrossRef]
8. Peng, J.; Zhang, P.; Zheng, L.; Tan, J. UAV positioning based on multi-sensor fusion. *IEEE Access* **2020**, *8*, 34455–34467. [CrossRef]

9. Weishi, C.; Yifeng, H.; Xianfeng, L. Survey on application of multi-sensor fusion in UAV detection technology. *Mod. Radar* **2020**, *42*, 15–29.
10. Zitar, R.A.; Al-Betar, M.; Ryalat, M.; Kassaymeh, S. A review of UAV visual detection and tracking methods. *arXiv* **2023**, arXiv:2306.05089. [CrossRef]
11. Broughton, G.; Majer, F.; Rouček, T.; Ruichek, Y.; Yan, Z.; Krajník, T. Learning to see through the haze: Multi-sensor learning-fusion system for vulnerable traffic participant detection in fog. *Robot. Auton. Syst.* **2021**, *136*, 103687. [CrossRef]
12. Blanchard, T.; Thomas, J.H.; Raoof, K. Acoustic localization and tracking of a multi-rotor unmanned aerial vehicle using an array with few microphones. *J. Acoust. Soc. Am.* **2020**, *148*, 1456–1467. [CrossRef]
13. Busset, J.; Perrodin, F.; Wellig, P.; Ott, B.; Heutschi, K.; Rühl, T.; Nussbaumer, T. Detection and tracking of drones using advanced acoustic cameras. In *Proceedings of the Unmanned/Unattended Sensors and Sensor Networks XI; and Advanced Free-Space Optical Communication Techniques and Applications*; SPIE: Bellingham, Germany, 2015; Volume 9647, pp. 53–60.
14. Christnacher, F.; Hengy, S.; Laurenzis, M.; Matwyschuk, A.; Naz, P.; Schertzer, S.; Schmitt, G. Optical and acoustical UAV detection. In *Proceedings of the Electro-Optical Remote Sensing X*; SPIE: Bellingham, Germany, 2016; Volume 9988, pp. 83–95.
15. Kheirandish, M.; Yazdi, E.A.; Mohammadi, H.; Mohammadi, M. A fault-tolerant sensor fusion in mobile robots using multiple model Kalman filters. *Robot. Auton. Syst.* **2023**, *161*, 104343. [CrossRef]
16. Hentati, A.I.; Krichen, L.; Fourati, M.; Fourati, L.C. Simulation tools, environments and frameworks for UAV systems performance analysis. In *Proceedings of the 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*; IEEE: New York, NY, USA, 2018; pp. 1495–1500.
17. Hu, Y.; Wu, X.; Zheng, G.; Liu, X. Object detection of UAV for anti-UAV based on improved YOLO v3. In *Proceedings of the 2019 Chinese Control Conference (CCC)*; IEEE: New York, NY, USA, 2019; pp. 8386–8390.
18. Liu, Y.; Yang, F.; Hu, P. Small-object detection in UAV-captured images via multi-branch parallel feature pyramid networks. *IEEE Access* **2020**, *8*, 145740–145750. [CrossRef]
19. Chen, Z.; Fan, K.; Ye, J.; Xu, Z.; Wei, Y. A Lightweight Multi-Module Collaborative Optimization Framework for Detecting Small Unmanned Aerial Vehicles in Anti-Unmanned Aerial Vehicle Systems. *Drones* **2025**, *10*, 20. [CrossRef]
20. Zheng, Y.; Shcherbakova, G.; Rusyn, B.; Sachenko, A.; Volkova, N.; Kliushnikov, I.; Antoshchuk, S. Wavelet transform cluster analysis of UAV images for sustainable development of smart regions due to inspecting transport infrastructure. *Sustainability* **2025**, *17*, 927. [CrossRef]
21. Chao, H.; Gu, Y.; Napolitano, M. A survey of optical flow techniques for UAV navigation applications. In *Proceedings of the 2013 International Conference on Unmanned Aircraft Systems (ICUAS)*; IEEE: New York, NY, USA, 2013; pp. 710–716.
22. Ivanov, L.I.; Obukhova, N.A.; Baranov, P.S. Review of modern UAV detection algorithms using methods of computer vision. In *Proceedings of the 2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*; IEEE: New York, NY, USA, 2020; pp. 322–325.
23. Yang, B.; Matson, E.T.; Smith, A.H.; Dietz, J.E.; Gallagher, J.C. UAV detection system with multiple acoustic nodes using machine learning models. In *Proceedings of the 2019 Third IEEE International Conference on Robotic Computing (IRC)*; IEEE: New York, NY, USA, 2019; pp. 493–498.
24. Shi, L.; Ahmad, I.; He, Y.; Chang, K. Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments. *J. Commun. Netw.* **2018**, *20*, 509–518. [CrossRef]
25. Medaiyese, O.O.; Ezuma, M.; Lauf, A.P.; Guvenc, I. Wavelet transform analytics for RF-based UAV detection and identification system using machine learning. *Pervasive Mob. Comput.* **2022**, *82*, 101569. [CrossRef]
26. Utebayeva, D.; Almagambetov, A.; Alduraibi, M.; Temirgaliyev, Y.; Ilipbayeva, L.; Marxuly, S. Multi-label UAV sound classification using Stacked Bidirectional LSTM. In *Proceedings of the 2020 Fourth IEEE International Conference on Robotic Computing (IRC)*; IEEE: New York, NY, USA, 2020; pp. 453–458.
27. Manamperi, W.; Abhayapala, T.D.; Zhang, J.; Samarasinghe, P.N. Drone audition: Sound source localization using on-board microphones. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 508–519. [CrossRef]
28. Hoshiba, K.; Komatsuzaki, I.; Iwatsuki, N. Proposal of practical sound source localization method using histogram and frequency information of spatial spectrum for drone audition. *Drones* **2024**, *8*, 159. [CrossRef]
29. Jekateryńczuk, G.; Piotrowski, Z. A survey of sound source localization and detection methods and their applications. *Sensors* **2023**, *24*, 68. [CrossRef]
30. 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]
31. Alla, I.; Olou, H.B.; Loscri, V.; Levorato, M. From sound to sight: Audio-visual fusion and deep learning for drone detection. In Proceedings of the 17th ACM Conference on Security and Privacy in Wireless and Mobile Networks, Seoul, Republic of Korea, 27–29 May 2024; pp. 123–133.
32. Yang, Y.; Yuan, S.; Yang, J.; Nguyen, T.H.; Cao, M.; Nguyen, T.M.; Wang, H.; Xie, L. AV-FDTI: Audio-visual fusion for drone threat identification. *J. Autom. Intell.* **2024**, *3*, 144–151. [CrossRef]

33. Haridevan, A.D.; Kang, J.; Yuan, M.; Shan, J. ROS2-Gazebo Simulator for Drone Applications. In *Proceedings of the 2024 International Conference on Unmanned Aircraft Systems (ICUAS)*; IEEE: New York, NY, USA, 2024; pp. 1232–1238.
34. Bondi, E.; Dey, D.; Kapoor, A.; Piavis, J.; Shah, S.; Fang, F.; Dilkina, B.; Hannaford, R.; Iyer, A.; Joppa, L.; et al. Airsim-w: A simulation environment for wildlife conservation with UAVs. In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, San Jose, CA, USA, 20–22 June 2018; pp. 1–12.
35. Dai, X.; Ke, C.; Quan, Q.; Cai, K.Y. RFlySim: Automatic test platform for UAV autopilot systems with FPGA-based hardware-in-the-loop simulations. *Aerosp. Sci. Technol.* **2021**, *114*, 106727. [CrossRef]
36. Bass, H.E.; Sutherland, L.C.; Zuckerwar, A.J.; Blackstock, D.T.; Hester, D.M. Atmospheric absorption of sound: Further developments. *J. Acoust. Soc. Am.* **1995**, *97*, 680–683. [CrossRef]
37. Liaquat, M.U.; Munawar, H.S.; Rahman, A.; Qadir, Z.; Kouzani, A.Z.; Mahmud, M.P. Localization of sound sources: A systematic review. *Energies* **2021**, *14*, 3910. [CrossRef]
38. Liu, H.; Zhang, X.; Li, P.; Yao, Y.; Zhang, S.; Xiao, Q. Time delay estimation for sound source localization using CNN-based multi-GCC feature fusion. *IEEE Access* **2023**, *11*, 140789–140800. [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.