



TRIDENT: Tri-modal Real-time Intrusion Detection Engine for New Targets

Ildi Alla, Selma Yahia, Valeria Loscri

► To cite this version:

Ildi Alla, Selma Yahia, Valeria Loscri. TRIDENT: Tri-modal Real-time Intrusion Detection Engine for New Targets. Computers & Security, 2025. <hal-05441894>

HAL Id: hal-05441894

<https://hal.science/hal-05441894v1>

Submitted on 5 Jan 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

TRIDENT: Tri-modal Real-time Intrusion Detection Engine for New Targets

Ildi Alla*, Selma Yahia*, Valeria Loscri*

*Inria Lille-Nord Europe, Lille, France

{ildi.alla, selma.yahia, valeria.loscri}@inria.fr

Abstract

The increasing availability of drones and their potential for malicious activities pose significant privacy and security risks, necessitating fast and reliable detection in real-world environments. However, existing drone detection systems often struggle in real-world settings due to environmental noise and sensor limitations. This paper introduces TRIDENT, a tri-modal drone detection framework that integrates synchronized audio, visual, and RF data to enhance robustness and reduce dependence on individual sensors. TRIDENT introduces two fusion strategies—Late Fusion and GMU Fusion—to improve multi-modal integration while maintaining efficiency. The framework incorporates domain-specific feature extraction techniques alongside a specialized data augmentation pipeline that simulates real-world sensor degradation to improve generalization capabilities. A diverse multi-sensor dataset is collected in urban and non-urban environments under varying lighting conditions, ensuring comprehensive evaluation. Experimental results show that TRIDENT achieves 96.89% accuracy in real-world recordings and 83.26% in a more complex setting (augmented data), outperforming unimodal and dual-modal baselines. Moreover, TRIDENT operates in real-time, detecting drones in just 6.09 ms while consuming only 75.27 mJ per detection, making it highly efficient for resource-constrained devices. The dataset and code have been released to ensure reproducibility¹.

Keywords

real-time UAV detection, multi-modal sensor fusion, audio-visual-RF data, edge computing, security and privacy surveillance.

1 Introduction and Motivation

The proliferation of Unmanned Aerial Vehicles (UAVs), commonly known as drones, has introduced transformative capabilities across various sectors, including package delivery, infrastructure inspection, and environmental monitoring [19, 22]. This rapid advancement has also intensified serious privacy vulnerabilities [41, 54]. Modern drones, equipped with high-resolution cameras and advanced sensors, have been involved in unauthorized surveillance. Reports have documented drones hovering near residential areas in the UK, capturing private footage of residents and celebrities without consent [9, 55]. Similar intrusions have been reported in U.S. neighborhoods, where drones have disrupted private gatherings, raising concerns over covert surveillance and data theft [1, 13]. Beyond privacy issues, drones have also been weaponized, with their unauthorized use in restricted airspace posing significant threats to national security [11, 53]. Their ability to operate

at low altitudes, maneuver discreetly, and evade conventional security systems highlights the urgent need for advanced detection technologies to reliably identify and mitigate these evolving threats.

Current drone detection systems generally fall into single-sensor and multi-sensor approaches. Single-sensor methods, including acoustic, visual, and Radio Frequency (RF)-based techniques, are simple to deploy but highly sensitive to environmental interference. Acoustic systems struggle in noisy urban settings, leading to false detections or missed targets [7, 45]. Similarly, visual-based methods, such as Convolutional Neural Networks (CNNs), perform well in controlled conditions but become unreliable in low-light or occluded settings [27, 36, 38]. RF-based systems, which analyze UAV communication signals, are prone to interference in congested areas, where overlapping signals can mask drone transmissions [20, 34]. Multi-sensor systems aim to overcome these limitations by integrating multiple modalities. Dual-modality approaches, such as audio-visual or RF-audio fusion, improve detection by combining complementary information [21, 29]. However, they remain unreliable in challenging environments where noise, obstructions, or overlapping RF signals compromise both sensing modalities [28]. Tri-modality systems enhance detection by increasing robustness and capturing a richer set of information [28, 33, 52].

Despite their advantages, current tri-modal approaches fall short in addressing the complexities of real-world drone detection due to three primary challenges. **First**, the datasets used to train and evaluate these systems often do not adequately represent the complexity of the real world. Many are collected in controlled settings—such as open fields or designated test areas—with minimal environmental variability [46, 52]. These conditions do not reflect the diverse noise profiles, visual clutter, and RF interference typical of urban and semi-urban environments [33]. Furthermore, these datasets generally lack synchronized multi-modal degradation, where disruptions across different sensing modalities occur simultaneously. This limitation hinders model generalization, reducing reliability in unpredictable operational deployments. **Second**, prevailing fusion techniques employed in tri-modal systems are often insufficiently sophisticated for robust multi-modal integration. Existing systems often rely on basic fusion strategies, such as simple averaging or concatenation of sensor output at the decision level [46, 52]. These rudimentary techniques fail to fully leverage the rich, complementary information inherent in heterogeneous sensor data and lack the capacity for deep cross-modal feature interaction and dynamic adaptation to varying sensor reliability under different environmental conditions. **Third**, comprehensive data augmentation strategies, crucial to improving model robustness and generalization, are conspicuously absent in existing tri-modal drone detection research [15, 29]. Real-world drone detection systems must operate reliably despite significant environmental variations, including fluctuating

¹<https://github.com/TRIDENT-2025/TRIDENT>

noise levels, variable lighting conditions, occlusions from urban structures, and diverse RF interference. The lack of comprehensive data augmentation in current methodologies severely limits their ability to adapt to these unseen, challenging environments.

This paper introduces TRIDENT, an effective tri-modal UAV detection framework designed to overcome the limitations of existing systems and enable robust, real-time drone detection in diverse and complex environments. TRIDENT addresses the above critical challenges through a comprehensive approach that integrates novel data acquisition, enhanced pre-processing and augmentation techniques, and sophisticated fusion techniques, ensuring greater adaptability and reliability in real-world scenarios.

Ethics: All experiments were conducted under the necessary permissions and do not raise any ethical concerns.

Summary of Contributions

- We propose TRIDENT, a novel framework for real-time drone detection designed for practical deployment. TRIDENT ensures robust detection by integrating data from audio, visual, and RF sensors, implementing two distinct fusion strategies: Late Fusion, which efficiently combines final predictions from unimodal models, and GMU Fusion, which integrates intermediate-layer features. Both fusion approaches are optimized for energy-efficient model architectures, making TRIDENT well-suited for privacy-preserving security applications with limited computational resources.
- We introduce an audio feature extraction scheme, a specialized pre-processing method for RF signals, and a frame-by-frame analysis approach for visual data. Further, we design a targeted data augmentation pipeline that applies synchronized, modality-consistent noise levels—both low and high—across inputs within each sample. This pipeline provides worst-case real-world degradation and measures the lower bound TRIDENT accuracy. Crucially, these augmentations are used *exclusively at test time*, while all models are trained on real, unaugmented data.
- We collect a 10 GB multi-sensor dataset, a key enabler for advancing research in robust drone detection. This dataset is uniquely distinguished by its diverse real-world collection settings, spanning both complex urban and less-obstructed non-urban locations. Furthermore, data acquisition was conducted under varied lighting conditions, including daylight and sunset scenarios, to capture the dynamic challenges of real-world deployments.
- Our experimental results demonstrate the strong performance advantages of TRIDENT. Specifically, our Late Fusion approach achieves an accuracy of 96.89% for tri-modal fusion in real-world data and 83.26% in high noisy augmented data. Beyond its high accuracy, this approach operates in real time, requiring only 6.09 ms per detection in its optimized configuration. Moreover, energy consumption analysis on a resource-constrained device highlights its efficiency, consuming as little as 75.27 mJ per detection.

2 Background on Drone Detection

Technical Challenges. As discussed in Section 8, existing drone detection techniques struggle in real-world environments, where model generalization issues, real-time processing constraints, and sensor-specific degradations impact detection accuracy and efficiency. Our work directly addresses the following key technical challenges inherent in real-world UAV detection:

• **Environmental Noise and Interference.** Drone detection models must operate reliably in *noisy* and *complex environments*, where multiple factors degrade sensor performance. Acoustic-based detection is affected by traffic, human speech, wind, and wildlife, which can overlap with drone sounds, leading to false negatives and reduced precision [12, 40]. Visual-based detection struggles with occlusions from buildings and trees, as well as lighting variations like shadows, glare, and low light, which reduce accuracy [6, 31]. RF-based detection, which analyzes drone communication patterns, faces significant interference in the crowded 2.4 GHz ISM band, where Wi-Fi, Bluetooth, and other wireless devices operate [23, 39]. This high-density RF environment complicates the reliable isolation of drone signals from background noise, further challenging detection accuracy.

• **Model Generalization.** Machine learning models for drone detection often struggle to generalize when deployed in unfamiliar environments that differ from their training data. Data augmentation, such as adding synthetic noise, is commonly used to improve generalization but often fails to fully bridge the gap between idealized training conditions and real-world complexities [56]. One major challenge is *overfitting to synthetic noise*. Models trained on artificially generated noise often learn its specific characteristics, making them highly sensitive to these distortions while failing to adapt to the diverse and unpredictable noise variations encountered in real-world settings. This rigidity in learned patterns results in a significant performance drop when exposed to unseen noise conditions during deployment. Another limitation is the *incomplete representation of real-world degradations*. Many augmentation techniques modify only one sensor modality at a time, overlooking the synchronized degradations that impact multiple sensors simultaneously. For example, in urban environments, visual occlusions from buildings or trees often coincide with acoustic masking from traffic or construction noise.

• **Real-time Processing Constraints.** Effective drone detection for security and privacy applications demands *real-time* or *near real-time performance*, ensuring timely responses to unauthorized drone activity. However, achieving this speed is challenging due to two primary factors. First, deep learning models, while offering high detection accuracy, require substantial computational power, often creating bottlenecks that lead to delays—particularly on resource-constrained edge devices. Second, multi-sensor data synchronization and fusion latency further hinder real-time performance. Integrating data from heterogeneous sensors in real-time requires precise temporal alignment and efficient fusion mechanisms. Without effective synchronization, delays in processing can degrade the system’s responsiveness, limiting its reliability in operational scenarios.

3 Threat Model

The increasing accessibility and sophistication of drone technology have created a tangible and evolving threat landscape for both individual privacy and organizational security. This work considers a realistic threat model involving a ground-based drone detection station safeguarding a defined area from unauthorized UAV activities. We posit an adversary seeking to exploit drone technology for malicious purposes, characterized by the following realistic knowledge and capabilities, denoted as $\mathcal{K} = \{I, L, C, T, S\}$:

- **I - Drone Identification:** The adversary knows specific drone *types* and *models* commonly used for illicit activities, as these are often commercially available and easily accessible.
- **L - Target Locations:** The adversary identifies specific *target locations* for surveillance or intrusion, such as private properties, restricted airspace, or sensitive facilities, using publicly available maps and aerial imagery.
- **C - Drone Capabilities:** The adversary has a baseline understanding of common drone *capabilities*, such as flight range, endurance, payload capacity, and sensor specifications, which are easily accessible from public technical documentation.
- **T - Flight Trajectories:** The adversary can plan sophisticated *flight trajectories*, including low-altitude paths to evade radar detection and autonomous navigation to reduce RF communication. Knowledge of flight patterns and restricted zone entry/exit points can be derived from public airspace information and observations.
- **S - Communication Signals:** The adversary knows the typical *operating frequency bands* for drones, such as the 2.4 GHz and 5.8 GHz ISM bands, identifiable through publicly available information.

With this knowledge, $\mathcal{K}$, adversaries can engage in privacy-invasive activities, such as collecting private images and videos, intercepting communications, or exploiting flight paths to infiltrate restricted zones. This threat model underscores the need for robust detection systems to protect individual privacy and organizational security.

4 Dataset Collection

4.1 Measurement Locations

Data collection was conducted in two primary locations: a non-urban area (stadium) and an urban area, as illustrated in Figure 1. These locations were carefully selected to account for varying environmental complexities, noise levels, and potential interference.



Figure 1: Measurement locations during data collection.

Non-Urban Area. Data was gathered in a stadium surrounded by trees and grass, offering a controlled environment with minimal urban interference. Background noise was moderate, primarily from nearby roads and occasional human or wildlife activity, making it ideal for baseline measurements.

Urban Area. Measurements were conducted 3.2 km from an airport in a busy urban setting with high noise levels from aircraft, traffic, and electronic devices. This environment introduced challenges such as overlapping audio frequencies, visual distractions, and RF interference, evaluating the system's performance in complex real-world conditions.

4.2 Drones and Sensing Equipment

In our study, we employed a comprehensive sensing setup to analyze and detect drones under various conditions. Specifically, we focused on two drone models: DJI Mini 2 [16] and DJI Mini 3 Pro

[17]. These models were selected due to their distinct characteristics, including differences in sound profiles, size, transmission range, and flight speed, enabling a robust evaluation of the detection system in diverse operational scenarios. The data collection process relied on audio, video, and RF sensors into our experimental setup (Figure 2). Each sensor type was selected to complement the others, ensuring precise and synchronized data acquisition across scenarios. The specifications of the devices are detailed in the Appendix A.

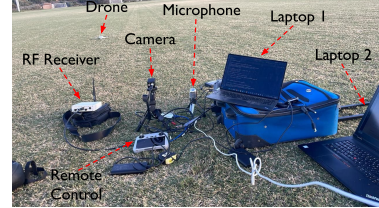


Figure 2: Experimental setup for synchronizing sensor data.

- ① **Audio Sensor.** The Samson Meteor condenser microphone was used to capture drone acoustic signatures. It features a cardioid polar pattern with a sensitivity of -33 dB/Pa and records at 44.1 kHz sampling rate with 16-bit resolution [43].
- ② **Video Sensor.** The Marshall CV-505 camera was selected for its compact design and high resolution [32]. It captures 10-second video clips at 640×640 resolution with 30 frames per second (fps).
- ③ **RF Sensor.** The USRP B210 software-defined radio (SDR), equipped with a VERT2450 antenna, was configured to operate in the 2.4 GHz ISM band. The system recorded in-phase and quadrature (I/Q) data at 55 MS/s, with a center frequency of 2.4415 GHz.

4.3 Experimental Conditions

Data was collected under diverse environmental and operational conditions to assess the system's robustness and real-world applicability, ensuring reliability even when some sensors were partially degraded or less effective.

Light Conditions. Data was collected during *daylight* and *sunset* to evaluate the system's performance under different lighting conditions. These scenarios tested the video sensor's effectiveness under varying visibility and assessed how the system adapts when the video sensor is less effective, relying more on audio and RF data for detection.

Noise Levels. Experiments were conducted in environments with varying noise intensities. A *stadium setting* provided moderate background noise from vehicles and people, while an *urban area* introduced higher noise levels from aircraft and dense activity, challenging the system's ability to isolate drone signals.

Environmental Factors. Data was collected in both line-of-sight (LOS) and non-line-of-sight (NLOS) conditions to evaluate the impact of visual and physical obstructions on sensor performance. In LOS conditions, the drone had a clear path to the sensors, ensuring optimal performance for video, audio, and RF detection. In NLOS conditions, obstacles like trees or walls partially or fully blocked the drone, affecting all sensors to varying degrees.

4.4 Data Collection Procedure

The data collection process was designed to capture high-quality sensor data across diverse scenarios. Locations were carefully selected based on environmental characteristics such as open spaces, interference sources, and background noise levels. Optimal sensor

placements and drone flight paths were determined to ensure comprehensive coverage. A shared timestamp system synchronized all sensors to maintain seamless data alignment.

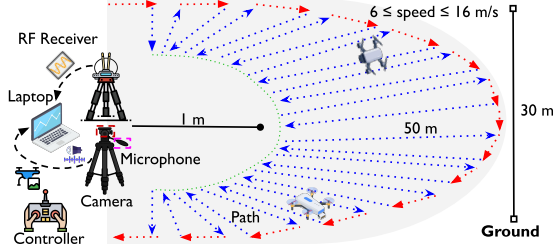


Figure 3: Data acquisition method from drones.

Drone flights included a variety of scenarios to replicate real-world conditions. Hovering scenarios maintained fixed altitudes between 1 and 30 meters (Figure 3) to ensure stable data collection. Linear flights simulated common applications like surveillance and delivery by following straight paths at varying speeds. Horizontal movements covered distances from 1 to 50 meters, capturing how proximity affects sensor data. Agile maneuvers introduced complexity through sharp turns, circular patterns, and other dynamic movements. These scenarios were repeated under both daylight and sunset conditions, as well as in urban and non-urban settings, incorporating both LOS and NLOS configurations—where the drone operated behind obstacles like walls or trees—to evaluate sensor performance across varying environmental and lighting conditions.

5 Sensor Fusion for Drone Detection

5.1 Data Pre-processing

① **Data Segmentation.** For our sensor fusion model, we segmented data into 0.25-second intervals to enhance real-time processing and ensure precise synchronization across audio, video, and RF modalities. Each audio segment corresponds to seven frames of video and one frame of RF spectrograms, creating a comprehensive dataset for analysis. This approach balances temporal resolution and computational efficiency while aligning with the characteristics of drone signatures from three sensors.

② **Audio Pre-processing.** We use Mel Frequency Cepstral Coefficients (MFCCs) to extract features crucial for drone detection, leveraging their robustness in capturing the spectral characteristics of audio signals. MFCCs are particularly effective in modeling the unique acoustic signatures of drones, as supported by prior research [4, 7, 40]. Each audio segment is divided into 40 frames, with 40 MFCCs computed per frame. The extraction process, as shown in Figure 4, begins with windowing to minimize boundary discontinuities and ensure smooth transitions between frames. Fast Fourier Transform (FFT) is then applied to analyze the frequency components of the signal. A Mel filter bank focuses on critical auditory frequencies, and logarithmic scaling models loudness perception. Finally, a Discrete Cosine Transform (DCT) decorrelates the energy distribution across the Mel filters, resulting in 40 MFCCs that effectively capture the acoustic signature of drones.

③ **Video Pre-processing.** We began by decomposing the recorded video clips into individual frames, converting the continuous video stream into a discrete sequence of images for frame-by-frame analysis. Each extracted frame was resized to a uniform resolution

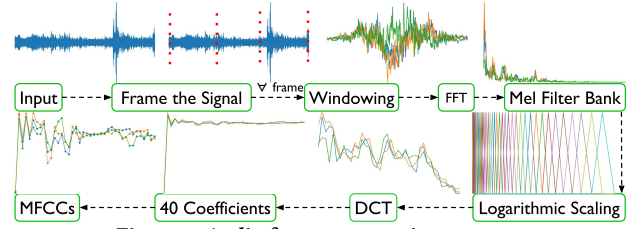


Figure 4: Audio feature extraction process.

of 112×112 pixels, ensuring consistent input dimensions and reducing computational complexity for subsequent deep learning models. Furthermore, to capture short-term temporal dynamics within the video data while maintaining computational efficiency, we employed a frame stacking approach. Specifically, consecutive sequences of seven frames were stacked along a new dimension to form a single input sample for the video-based unimodal models.

② **RF Pre-processing.** For RF data, raw I/Q samples are converted into spectrograms, ensuring a high-resolution representation of the temporal and spectral characteristics essential for drone detection. The transformation employs the Short-Time Fourier Transform (STFT), defined as:

$$S(t, f) = \sum_{n=0}^{N-1} x[n] \cdot w[n-t] \cdot e^{-j2\pi f n/N}, \quad (1)$$

where $x[n]$ is the I/Q data, $w[n]$ is the Hamming window function, t is the time index, and f is the frequency. Each spectrogram is treated as an image and automatically resized to 112×112 pixels.

5.2 Data Augmentation

① **Audio Augmentation.** We applied a range of audio augmentation techniques to evaluate the system's robustness against noise, signal variability, and environmental factors. These techniques included adding *background noise* to simulate conditions where drone sounds might be obscured by other noises, such as traffic, wind, or human activity. *Harmonic distortions* were introduced to emulate non-linear variations in microphone response, while *pitch shifting* modified the frequency to account for changes in drone motor speed or Doppler effects. *Clicks* were generated to represent sudden disturbances, such as electrical noise or short sound bursts caused by interference. *Mono conversion* was applied to simulate audio input from a single-channel sensor type that is different from multi-channel one. Finally, *volume scaling* adjusted audio levels to reflect varying recording distances and sound intensities. The effects of these augmentations are shown in Figure 5. These transformations enriched the dataset by simulating real-world acoustic variations, testing the system's resilience under challenging conditions.

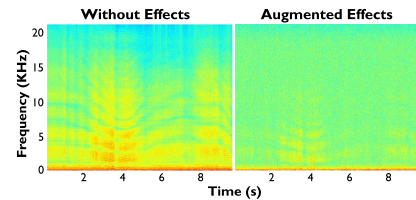


Figure 5: T-F representation of drone sound without/with background noise and volume scaling effects.

The impact of these audio augmentations on the original signal was evaluated using the Mel Cepstral Distortion (MCD) metric

[42]. MCD quantifies the spectral dissimilarity between the original and augmented audio, providing a quantitative measure of the introduced noise and distortion. It is calculated using the following formula in decibels (dB):

$$MCD = \frac{10\sqrt{2}}{\ln(10)} \cdot \frac{1}{T} \sum_{t=1}^T \sqrt{\sum_{m=1}^{n_{\text{coeffs}}} (MFCC_{\text{real}}(m, t) - MFCC_{\text{aug}}(m, t))^2}, \quad (2)$$

where $MFCC_{\text{real}}(m, t)$ and $MFCC_{\text{aug}}(m, t)$ represent the m -th MFCC coefficient at time frame t for the real and augmented audio signals, respectively. T is the total number of time frames, and n_{coeffs} is the number of MFCC coefficients. Lower MCD values indicate minimal distortion, while higher values reflect greater spectral dissimilarity between the original and augmented signals.

② **Visual Augmentation.** Several augmentation techniques were applied to the video data to enhance the dataset's diversity and better reflect real-world conditions. These techniques included adding *random noise* to replicate sensor artifacts by introducing pixel-level distortions, representing degraded image quality. *Horizontal flipping* changed the frame orientation to account for variations in perspective, while *rotation* introduced angle shifts to simulate different viewpoints. *Color jitter* adjusted brightness and contrast to recreate diverse lighting conditions, such as overexposure or shadows. *Gaussian blur* softened the images to mimic motion blur or out-of-focus visuals. Finally, *salt-and-pepper noise* added random white and black pixels to simulate transmission errors or sensor defects, degrading visual clarity. As shown in Figure 6, these transformations effectively diversified the dataset by incorporating a range of challenging conditions.



Figure 6: Drone image without/with Gaussian blur and color jitter.

The impact of these visual augmentations on the original data was evaluated using the Structural Similarity Index (SSIM) [2]. SSIM provides a perceptual similarity metric between the original and augmented video frames, with values ranging from 0 to 1, where 1 indicates perfect similarity and lower values signify increasing distortion. The SSIM calculation is given by Eq 3:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \quad (3)$$

where μ_x, μ_y represent the luminance of the original image (x) and the augmented image (y), σ_x and σ_y denote their contrast, and σ_{xy} represents their structural similarity. The constants C_1 and C_2 stabilize the division to avoid numerical instabilities.

③ **RF Augmentation.** Two distinct data collection scenarios were designed: *in-the-wild* (ITW) and *over-the-air* (OTA). In ITW scenario, RF signals were recorded in a controlled laboratory using a USRP B210, capturing real-world interference from Wi-Fi devices and IoT sensors operating in the 2.4 GHz band. This dataset realistically represents RF environments where multiple devices share the same frequency spectrum. In the OTA scenario, drone signal transmissions were recorded in an outdoor environment, capturing real-world

RF signals with natural interference. For increased interference, a USRP B200 Mini was configured as a jammer, injecting artificial noise into the 2.4 GHz band. Gaussian and uniform noise were digitally generated and transmitted at different bandwidths (e.g., 10 MHz and 20 MHz), simulating multiple overlapping transmissions.

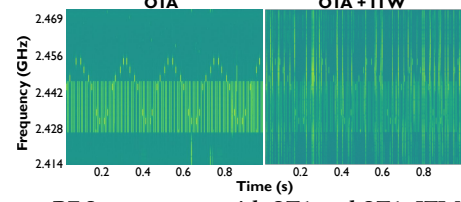


Figure 7: RF Spectrograms with OTA and OTA+ITW cases.

We enhanced the dataset's complexity and diversity by combining ITW and OTA data through additive operations. This process involved mixing drone communication signals collected in the OTA scenario with laboratory-collected noise from the ITW scenario. As shown in Figure 7, OTA scenarios display distinct drone communication patterns, while OTA+ITW scenarios show these patterns increasingly obscured by background noise. This combination reflects real-world challenges and enables a comprehensive evaluation of detection robustness.

The level of noise introduced during RF data augmentation was quantified using the Signal-to-Noise Ratio (SNR) in dB [44]. SNR measures the signal power relative to the noise power, effectively indicating the level of interference added. It is calculated as:

$$SNR = 10 \cdot \log_{10} \left(\frac{P_{\text{signal}}}{P_{\text{noise}}} \right), \quad (4)$$

where P_{signal} represents the power of the drone communication signal and P_{noise} denotes the power of the injected noise. Lower SNR values indicate higher levels of noise relative to the signal.

5.3 Unimodal Feature Extraction

5.3.1 *Models for Audio Feature Extraction.* The pre-processing step of raw audio data generates 2-dimensional vectors with a shape of (1, 1600) from various audio recordings. These 2D vectors are then utilized to extract features relevant to UAV detection. This feature extraction process can be systematically achieved using a ML model. CNNs have proven to be highly effective in learning features from audio, visual, and RF data. For our specific use case, we employ two different state-of-the-art CNN architectures, with further customization to accommodate the unique nature and shape of our audio data. In the following, we detail the architectural specifications and adjustments that we have made for each model: (i) **LeNet-based Architecture.** As detailed in Table 1, this architecture is composed of four feature extraction blocks – each block includes a convolutional (Conv) layer, a batch normalization (BN) layer, and a max pooling (MaxPool) layer. The number of feature maps in these blocks ranges from 8 to 32, which helps to balance the model's complexity and computational efficiency, making it both compact and lightweight. To accommodate the Conv2D layers in our CNN, we perform a dimension reshaping of the audio vector. Initially, the audio vector has dimensions of (1, 1600). We reshape this vector to a size of (1, 40, 40). This transformation is crucial as it enables the use of 2D convolutional layers, which are better suited for extracting spatial hierarchies and intricate patterns from the data compared to 1D convolutional layers.

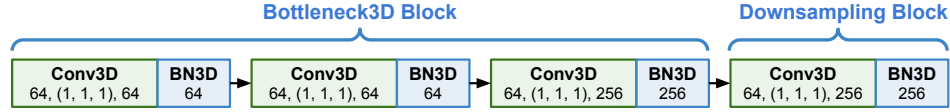


Figure 8: The architecture breakdown of the bottleneck3D and downsampling layers for 'Block_1' in Table 3.

Table 1: Architecture specifications of audio-LeNet.

Block	Neural Layers Operations	Feature map	Activation
Block_1	Conv(3x5x5), BN(8), MaxPool(2)	8	ReLU
Block_2	Conv(8x3x3), BN(16), MaxPool(2)	16	ReLU
Block_3	Conv(16x3x3), BN(24), MaxPool(2)	24	ReLU
Block_4	Conv(24x3x3), BN(32), MaxPool(2)	32	ReLU
Classifier	Fully-Connected(32, 2)	32	Sigmoid()

(ii) **VGG-based Architecture.** We employ a more complex and over-parameterized backbone for audio feature extraction, utilizing the VGGNet architecture [47], which was originally designed for visual data processing. Specifically, we adopt a VGG-19 variant to process audio data in the same manner as previously explained for the LeNet model. The backbone of VGG-19 depicts a wide and deep architecture composed of 16 convolutional layers with 3x3 kernels, sometimes followed by max pooling layers. The number of feature maps in these layers ranges from 64 to 512, allowing the model to capture a broad spectrum of audio features with high precision. We modify the classifier part of the VGG-19 architecture to optimize the model for audio data. Specifically, we remove the two fully-connected layers originally included in the VGG-19 model and replace them with a single, less parameterized fully-connected layer. This adjustment not only reduces the model's complexity but also enhances processing time and facilitates seamless training. Our final audio-VGG model architecture is detailed in Table 2. To simplify the reading, we group the layers by blocks, each containing two convolutional layers with specifications on the number and spatial dimensions of their kernels.

Table 2: Architecture specifications of audio-VGG.

Block	Neural Layers Operations	Feature map	Activation
Block_1	Conv(1x3x3), Conv(64x3x3), MaxPool(2)	64, 64	ReLU
Block_2	Conv(64x3x3), Conv(128x3x3), MaxPool(2)	128, 128	ReLU
Block_3	Conv(128x3x3), Conv(256x3x3)	128, 256	ReLU
Block_4	Conv(256x3x3), Conv(256x3x3), MaxPool(2)	256, 256	ReLU
Block_5	Conv(256x3x3), Conv(512x3x3)	256, 512	ReLU
Block_6	Conv(512x3x3), Conv(512x3x3)	512, 512	ReLU
Block_8	Conv(512x3x3), Conv(512x3x3), MaxPool(2)	512, 512	ReLU
Classifier	Fully-Connected(512, 2)	512	Sigmoid()

5.3.2 Models for Visual and RF Feature Extraction. Visual data are pre-processed and segmented into sequences of frames featuring the presence/absence of the drone. In parallel, RF data (i.e., spectrograms derived from I/Q data transformation) are also organized into frame-like inputs—typically one spectrogram 'frame' per segment. Since we're more interested in detecting the presence/absence of the drone, we consider a binary classification task where labels are assigned to both video and RF segments based on the underlying frames. Formally, let Y be the binary label assigned to a video or RF sequence, and y_i be the binary label of the i -th frame:

$$Y = \max(y_1, y_2, \dots, y_n), \quad (5)$$

where $i = 1, 2, \dots, n$, and n is the total number of frames. The max function returns 1 if *at least one* frame is labeled 1, and 0 otherwise.

The pre-processing of video and RF frames results in 4D vectors with a shape of $X_{\text{shape}} = (N, C, H, W)$, where N is the number of frames, C is the number of color channels per frame, and H and W are the height and width, respectively. By incorporating the N

dimension for stacking a set of frames, both temporal and spatial dependencies can be learned from the sequences. Unlike single-frame classification, which uses 3D vectors, video classification requires specific ML models with filters extended into an additional dimension to train on both spatial and temporal data. Therefore, we utilize 3D CNN models featuring convolutional filters in three-dimensional space. We have selected two of the most efficient and effective state-of-the-art CNN models for visual tasks: MobileNet-v2 and ResNet-10 [24, 26]. Since these models were originally proposed for single-frame classification, we applied a further transformation to map all their layers in the three-dimensional space. Further details on the architectural specifications of each model are provided below:

(i) **ResNet-based Architecture.** ResNet models are highly successful architectures for learning visual features [25]. We adopt a lightweight ResNet-10 variant with fewer weights and a simpler design, incorporating both bottleneck and downsampling blocks. As detailed in Table 3, the backbone of ResNet-10 consists of 10 convolutional layers, organized into bottleneck3D and downsample blocks with 3x3x3 and 1x1x1 kernels, followed by batch normalization layers, as illustrated in Figure 8. The number of feature maps in these layers ranges from 64 to 2048, enabling the model to capture a wide range of video features with high precision. In the classifier block, we apply spatio-temporal average pooling with a kernel size of $(N, 4, 4)$, where the number of frames is $N = 7$. This pooling layer performs downsampling along the spatial dimensions (depth, height, and width) by averaging the values over the pooling window (i.e., $(7, 4, 4)$) for each channel of the input. For RF data, $N = 1$, so the same pooling simply reduces the spatial dimensions. Finally, a fully connected layer processes the output of the pooling operation and produces a single classification score for the N frames, which is then translated into a predicted label for the entire video sequence.

Table 3: Architecture specifications of ResNet-10.

Block	Neural Layers Operations	Fmap	Activation
Conv_1	Conv2D(3x7x7), MaxPool(3), BN(64)	64	ReLU
Block_1	Bottleneck3D(64, 256), Downsample(64, 256)	256	ReLU
Block_2	Bottleneck3D(256, 512), Downsample(256, 512)	512	ReLU
Block_3	Bottleneck3D(512, 1024), Downsample(512, 1024)	1024	ReLU
Block_4	Bottleneck3D(1024, 2048), Downsample(1024, 2048)	2048	ReLU
Classifier	AvgPooling3D($N, 4, 4$), Fully-Connected(2048, 2)	2	Sigmoid()

(ii) **MobileNet-based Architecture.** MobileNet models were initially introduced to enhance processing efficiency, particularly on resource-constrained hardware devices, while maintaining high performance on visual tasks [26]. As such, MobileNet variants are lightweight and compact, with fewer trainable weights and simpler operations, making them ideal for real-time applications. Consequently, to meet the real-time requirements of our UAV detection systems, we utilize the MobileNet architecture to learn visual features from video and RF sequences. We have further modified the original MobileNet by transforming the MBConv2D blocks into MBConv3D blocks to handle the processing of 4D vectors that represent sequences. Table 4 outlines the architectural specifications of the MobileNet3D model. It features a sequence of MBConv3D

layers, which are composed of a typical Conv3D layer followed by batch normalization and Rectified Linear Unit (ReLU) activation. The classifier block involves a spatio-temporal average pooling with a kernel size of (28, 1, 1) and a fully connected layer with sigmoid activation. For the single-frame RF input, the temporal dimension N is 1.

Table 4: Architecture specifications of MobileNet.

Block	Neural Layers Operations	Fmap	Activation
Conv_1	Conv2D(3x3x3x3), BN(32)	32	ReLU
Block_1	MBConv3D(32, 64), MBConv3D(64, 128),	128	ReLU
Block_2	MBConv3D(128, 128), MBConv3D(128, 256)	256	ReLU
Block_3	MBConv3D(256, 256), MBConv3D(256, 512)	512	ReLU
Block_4	MBConv3D(512, 512), MBConv3D(512, 512),	512	ReLU
Block_5	MBConv3D(512, 512), MBConv3D(512, 512)	512	ReLU
Block_6	MBConv3D(512, 512), MBConv3D(512, 1024)	1024	ReLU
Block_7	MBConv3D(1024, 1024)	1024	ReLU
Classifier	AvgPooling3D(N, 1, 1), Fully-Connected(256, 2)	2	Sigmoid()

5.4 Fusion Strategies

The fusion techniques are key components of the TRIDENT framework, as shown in Figure 9. In this pipeline, audio, visual, and RF data are collected, pre-processed, and passed through modality-specific models (LeNet or VGG-19 for audio, ResNet-10 or MobileNet for video and RF). The resulting unimodal features or decisions are then combined via GMU or Late Fusion to yield the final drone prediction. The following subsections describe the fusion strategies.

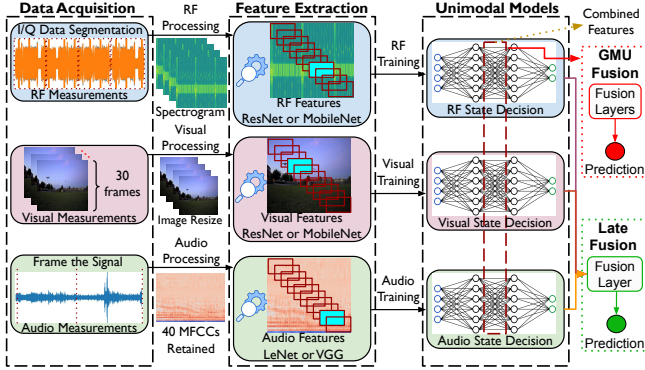


Figure 9: TRIDENT Framework.

5.4.1 Late Fusion. Late Fusion is one straightforward way to fuse the output predictions calculated by each unimodal model. The computations involving each unimodal model should be performed independently, and only their final output predictions (i.e., the output of the classifier block) are considered for the fusion. We note that the output predictions are also probability distributions produced by unimodal models in the case of classification. Thus, Late Fusion is the most interpretable method, operating on the highest level of unimodal output data. Algorithm 1 reports details of the steps involved in Late Fusion. Let y_{Aud} , y_{Vis} , y_{RF} be the output predictions (i.e., probabilities) of absence/presence of UAV computed by audio, video and RF unimodal models, respectively. Given these outputs, Late Fusion operates as follows:

Step ①: Output predictions are combined using normalized fusion weights, α , β , and γ , which determine the contribution of each output prediction to the final prediction. Fusion weights are learned during training based on the importance of each modality and are normalized as follows:

$$\alpha' = \frac{\alpha}{\alpha + \beta + \gamma}, \beta' = \frac{\beta}{\alpha + \beta + \gamma}, \gamma' = \frac{\gamma}{\alpha + \beta + \gamma}. \quad (6)$$

Step ②: The final output prediction, y , is obtained by applying a weighted sum on the the unimodal output predictions y_{Aud} , y_{Vis} , and y_{RF} using normalized weights as follows:

$$y = \alpha' y_{Aud} + \beta' y_{Vis} + \gamma' y_{RF}. \quad (7)$$

Algorithm 1 Late Fusion for audio-video data.

Require: Input predictions y_{Aud} , y_{Vis} , y_{RF}

Require: Fusion weights α, β, γ

- 1: Normalize the fusion weights: $\alpha' = \frac{\alpha}{\alpha + \beta + \gamma}, \beta' = \frac{\beta}{\alpha + \beta + \gamma}, \gamma' = \frac{\gamma}{\alpha + \beta + \gamma}$
- 2: Compute the fused prediction: $y = \alpha' y_{Aud} + \beta' y_{Vis} + \gamma' y_{RF}$
- 3: **return** y

5.4.2 Gated Multimodal Unit (GMU) Fusion. GMU Fusion is designed to effectively fuse information from multiple modalities by learning each contribution to the final prediction. Let x_{Aud} , x_{Vis} , and x_{RF} , be the inputs of the GMU Fusion. Typically, these inputs are selected from the output intermediate feature maps of the unimodal models. Since the last output feature map (the input feature map of the classifier layer) usually encapsulates high-level features that are semantically meaningful, we use them as inputs to the GMU Fusion operation. Hence, contribution scores will be learned and assigned to each modality based on high-abstracted and meaningful output feature maps. Algorithm 2 details the working mechanism of the GMU Fusion.

Algorithm 2 GMU Fusion for audio-video data.

Require: Input vectors x_{Aud} , x_{Vis} , x_{RF}

- 1: Compute the gating vector: $g = \sigma(W_g [x_{Aud}, x_{Vis}, x_{RF}] + b_g)$
- 2: Compute the transformed vectors:
- 3: $y_1 = \tanh(W_y^1 x_{Aud} + b_y^1)$
- 4: $y_2 = \tanh(W_y^2 x_{Vis} + b_y^2)$
- 5: $y_3 = \tanh(W_y^3 x_{RF} + b_y^3)$
- 6: Compute the fused representation: $y = g_1 \odot y_1 + g_2 \odot y_2 + g_3 \odot y_3$
- 7: **return** y

Step ①: A *gating vector* g is computed from the inputs x_{Aud} , x_{Vis} , and x_{RF} using a sigmoid activation function. The gating vector g serves as a control knob for the contribution of each modality to the joint information from the fused data. Formally, the computation of the gated vector is defined as follows:

$$g = \sigma(W_g [x_{Aud}, x_{Vis}, x_{RF}] + b_g), \quad (8)$$

where σ is the sigmoid function, W_g are the weights, and b_g is the bias term. The concatenation of the input vectors x_{Aud} , x_{Vis} , and x_{RF} ensures that the gating mechanism considers learned features from three modalities (i.e., audio, video, and RF features).

Step ②: Each of the inputs x_{Aud} , x_{Vis} , and x_{RF} independently undergoes a linear transformation followed by a $\tanh$ activation function. This step produces the transformed vectors y_1 , y_2 , y_3 :

$$y_1 = \tanh(W_y^1 x_{Aud} + b_y^1), \quad (9)$$

$$y_2 = \tanh(W_y^2 x_{Vis} + b_y^2), \quad (10)$$

$$y_3 = \tanh(W_y^3 x_{RF} + b_y^3), \quad (11)$$

where, W_y^1 , W_y^2 , and W_y^3 are the weights, and b_y^1 , b_y^2 , and b_y^3 are the bias terms for the transformations of x_{Aud} , x_{Vis} , and x_{RF} .

Step ③: A final fused output y is computed by combining the transformed vectors y_1 , y_2 , and y_3 based on the gating vector $g =$

$[g_1, g_2, g_3]^T$. The combination is performed via an element-wise multiplication $\odot$, using g to weigh each modality’s contribution:

$$y = g_1 \odot y_1 + g_2 \odot y_2 + g_3 \odot y_3. \quad (12)$$

This equation ensures that the unimodal features from audio, video, and RF are fused in a controlled manner, based on the learned contribution scores stored in the gating vector g .

6 Results and discussion

6.1 Experimental Setup

6.1.1 Hardware Characteristics. The training of the unimodal and multi-modal models was performed on a single GPU, the GeForce GTX 1080 Ti, built on the Pascal architecture, featuring 584 CUDA cores, 11 GB of GDDR5X VRAM, and a 352-bit memory interface, offering a memory bandwidth of 484 GB/s. The training settings included the ADAM optimizer, LRCOSINEANNEALING for learning rate scheduling within the range of 0.001 to 0.01, and a budget of 20 epochs. After training, the DNNs were run on the Jetson Orin Nano module, powered by a 6-core ARM Cortex-A78AE 64-bit CPU. The system is equipped with 8 GB of 128-bit LPDDR5 memory and a 1024-core NVIDIA Ampere architecture GPU with 32 Tensor Cores.

6.1.2 Dataset Overview and Distribution. This study utilizes a multi-modal dataset integrating audio, video, and RF sensor data, categorized into two classes: drone (class 0) and no drone (class 1). Each data sample is temporally synchronized, comprising a 0.25-second audio segment, a stack of 7 video frames, and a single RF spectrogram. The audio data captures drone flight sounds while differentiating them from general environmental noise. The video data includes drone images at varying distances, along with non-drone objects such as clouds, helicopters, birds, and buildings. The RF data consists of I/Q samples, later transformed into spectrograms to analyze drone-specific communication patterns. A detailed breakdown of the dataset structure, including sample sizes and recording parameters, is provided in Appendix B.

Table 5: Detailed data composition across training, validation, and testing phases.

Data type	Class	Files	Audio	Visual	RF	Duration (s)
Train	Drone (0)	118	4720	33040	4720	2120
	No-drone (1)	94	3760	26320	3760	
Validation	Drone (0)	20	800	5600	800	320
	No-drone (1)	12	480	3360	480	
Test	Drone (0)	21	840	5880	840	330
	No-drone (1)	12	480	3360	480	
Total		277	11080	77560	11080	2770

A structured distribution is implemented to ensure robust model development, with 77% of the dataset allocated for training, 11% for validation, and 12% for testing. The training and validation phases rely on real, non-augmented data, allowing models to learn drone detection features without augmentation effects. In contrast, the test dataset includes noise augmentation applied synchronously across all three modalities within each sample, simulating real-world conditions where sensor degradation can affect all data streams simultaneously. The complete dataset composition across training, validation, and testing phases is summarized in Table 5. The trained models are first evaluated on the entire test dataset to establish a baseline performance. For a more granular assessment, the test dataset (1320 samples) is further categorized based

on lighting conditions (daylight, sunset) and location types (urban, non-urban), as presented in Table 6.

Table 6: Test dataset breakdown by environmental conditions.

Environmental Condition		Lighting		Location	
		Daylight	Sunset	Urban	Non-Urban
Class	Drone (0)	360	480	160	680
	No-drone (1)	320	160	200	280
Total samples per condition		680	640	360	960

6.1.3 Evaluation Scenarios. We evaluated the system’s robustness using two distinct scenarios based on the level of noise distortion introduced in the augmented data: scenario 1 (low noise) and scenario 2 (high noise). Noise was added to each data sample from the audio, visual, and RF sensors in the test set, while the model was trained exclusively on real data to avoid biasing the learning process and ensure a fair evaluation. The augmentation levels were carefully selected to simulate real-world noise conditions, ensuring the test set remains both realistic and representative of practical scenarios. The distortion in each scenario was quantified using specific metrics: MCD (dB) for audio, SSIM for video, and SNR (dB) for RF data, as described in Section 5.2 and summarized in Table 7.

Table 7: Quantification of augmentation levels for noisy augmented scenarios.

Scenario	Audio (MCD) [42]	Visual (SSIM) [2]	RF (SNR) [44]
1. Low Noise	3.54 dB	0.9597	15-20 dB
2. High Noise	11.82 dB	0.7271	5-10 dB

The low noise scenario represents minimal distortion, where audio, visual, and RF data maintain relatively good quality. In contrast, the high noise scenario reflects significant distortion, simulating harsh real-world conditions with substantial noise and interference across all modalities.

6.2 Unimodal Detection Analysis

This section evaluates the performance of each modality—audio, video, and RF—independently to highlight their strengths and limitations under varying levels of environmental noise and distortion. The evaluation was conducted on real data (non-augmented) and in two noisy augmented scenarios (scenario 1 and scenario 2).

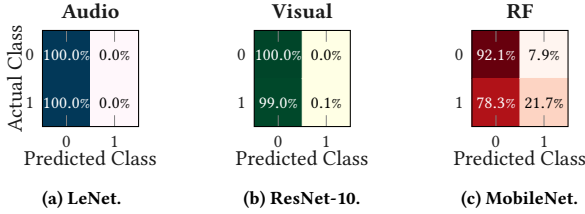
Audio-based Detection. LeNet and VGG-19 architectures were evaluated for audio-based detection. In real-world data, LeNet outperformed VGG-19, achieving 90.00% accuracy compared to 63.64% (Table 8). This superior performance is attributed to LeNet’s ability to extract relevant spectral features, resulting in higher precision, recall, and F1-score. It also demonstrated greater efficiency with a 1.17 ms detection time, compared to 2.13 ms for VGG-19. In noisy augmented data, both models showed a decline in accuracy. LeNet remained relatively robust in scenario 1, achieving 82.5% accuracy, but dropped to 63.64% in the more challenging scenario 2, revealing its sensitivity to high noise. The confusion matrix for audio-LeNet in scenario 2 (Figure 10a) shows a 100% True Positive Rate (TPR) but also a 100% False Positive Rate (FPR), misclassifying all non-drone instances as drones, rendering it ineffective under extreme noise. VGG-19 consistently performed poorly across all conditions, maintaining a 63.64% accuracy in both scenarios.

Video-based Detection. Visual detection performed well in real data, with ResNet-10 (96.97% accuracy) outperforming MobileNet (86.14%), as shown in Table 8. ResNet-10’s deeper architecture and

Table 8: Unimodal model performance under real and augmented noisy data.

		Real Data					Noisy Augmented Data										Detec. Time (ms)
		Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Scenario 1					Scenario 2					
							Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	
Audio	LeNet	90.00	97.97	86.07	91.63	89.60	82.50	87.55	84.52	86.01	81.33	63.64	63.64	100.00	77.78	38.89	1.17
	VGG-19	63.64	63.64	100.00	77.78	38.89	63.64	63.64	100.00	77.78	38.89	63.64	63.64	100.00	77.78	38.89	2.13
Video	ResNet-10	96.97	95.45	100.00	97.67	96.66	74.47	71.37	100.00	83.29	64.60	64.02	63.88	100.00	77.96	40.01	2.33
	MobileNet	86.14	94.69	82.86	88.38	85.60	38.56	82.22	4.40	8.36	31.08	37.65	100.00	2.02	3.97	28.90	2.11
RF	ResNet-10	98.41	98.69	98.81	98.75	98.28	81.06	81.58	90.71	85.91	78.52	62.27	65.77	84.88	74.12	52.28	2.42
	MobileNet	92.80	93.47	95.36	94.40	92.16	79.09	78.60	92.26	84.88	75.49	66.52	67.30	92.14	77.79	54.89	1.80

residual connections helped capture complex spatial features, minimizing missed detections. MobileNet, though more efficient (2.11 ms detection time), had lower recall (82.86%), making it more prone to missing drones. However, both models struggled in noisy augmented data. ResNet-10’s accuracy dropped to 74.47% in scenario 1 and 64.02% in scenario 2, while MobileNet’s fell to 38.56% and 37.65%, respectively. The decline highlights the sensitivity of vision-based detection to noise, blur, and occlusions. The confusion matrix for ResNet-10 in scenario 2 (Figure 10b) shows a 100% TPR, detecting all drones, but a FPR of 99.0% and a True Negative Rate (TNR) of only 0.1%, misclassifying nearly all non-drone instances. This pattern mirrors audio-LeNet, indicating a failure to discriminate effectively in highly degraded visual conditions.

**Figure 10: Confusion matrices for the best performing unimodal models in scenario 2.**

RF-based Detection. RF-based detection performed well in real data, with ResNet-10 (98.41% accuracy, 98.75% F1-score) outperforming MobileNet (92.80% accuracy, 94.40% F1-score), as shown in Table 8. These results indicate that RF signals provide strong features for drone detection in a relatively clear spectrum. However, performance declined in noisy augmented data. ResNet-10’s accuracy dropped to 81.06% in scenario 1 and 62.27% in scenario 2, while MobileNet fell to 79.09% and 66.52%, respectively. The confusion matrix for RF-MobileNet (Figure 10c) shows a 92.1% TPR and a 7.9% False Negative Rate (FNR), detecting most drones, but a high 78.3% FPR, frequently misclassifying non-drone signals. This highlights the challenge of distinguishing drones from background RF activity in noisy environments.

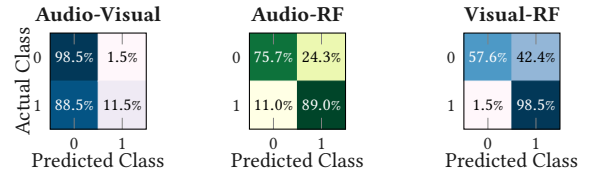
The instability of individual modalities in noisy conditions highlights the need for multi-modal fusion to achieve more reliable and robust drone detection across diverse environments. In the subsequent multi-modal fusion analysis, scenario 2 is considered, as it presents the most challenging conditions for the detection system.

6.3 Dual-Modal Fusion Analysis

In this section, we evaluate the performance of dual-modal fusion, using Late and GMU Fusion techniques to analyze how combining modalities improves detection performance. Table 9 presents

the best-performing dual-modal fusion strategies under both real and noisy augmented data, while the full results for all modality combinations are provided in Appendix C.1.

Best Performing Combinations. In real data, as shown in Table 9, all the listed best combinations exhibit high accuracy, demonstrating the benefit of combining complementary sensor data when environmental factors are benign. Audio-visual (VGG-19 + ResNet-10) with Late Fusion reaches 96.97% accuracy, highlighting the synergy of visual and auditory cues in real data. Audio-RF and visual-RF fusions also achieve strong accuracies of 96.82% and 96.97% respectively, underscoring the effectiveness of dual-modality integration. However, in noisy augmented data, dual-modal fusion experiences a noticeable decline in performance, highlighting its sensitivity to high environmental noise. Visual-RF with Late Fusion shows a noticeable accuracy drop to 72.50%, indicating vulnerability when both visual and RF data streams are degraded. Audio-visual with Late Fusion also sees a performance decrease, reaching 66.82% accuracy in noisy environments. In contrast, audio-RF using GMU Fusion demonstrates superior resilience, maintaining an accuracy of 80.53% under noisy conditions. This highlights audio-RF GMU Fusion as the most robust dual-modal approach in challenging environments.

**Figure 11: Confusion matrices for the best performing dual-modality fusions in noisy augmented data: (blue) Late Fusion; (green) GMU Fusion.**

Confusion Matrices. Analyzing the confusion matrices for augmented noisy conditions (Figure 11) provides further insights. For Late Fusion in Figure 11a, audio-visual (VGG-19 + ResNet-10) shows high drone detection capability (TPR: 98.5%) but also a high false alarm rate (FPR: 88.5%), indicating limited ability to reject non-drone instances (TNR: 11.5%). Visual-RF (MobileNet + ResNet-10) with Late Fusion in Figure 11c, conversely, excels at rejecting non-drones (TNR: 98.5%, FPR: 1.5%) but struggles to detect drones effectively (TPR: 57.6%, FNR: 42.4%). In contrast, Figure 11b, audio-RF GMU fusion (VGG-19 + ResNet-10) achieves a more balanced performance. It provides a good drone detection rate (TPR: 75.7%) while maintaining a significantly lower false alarm rate compared to audio-visual Late Fusion (FPR: 11.0%, TNR: 89.0%).

Table 9: Performance metrics for the best performing two-modal fusions under real and noisy augmented data.

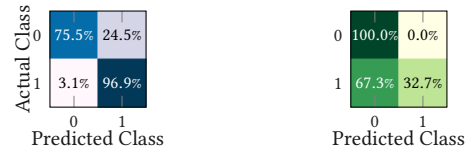
Fusion Type	Modality Combination	Models	Real Data					Noisy Augmented Data					Detec. Time (ms)
			Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	
Late	Audio-Visual	VGG-19 + ResNet-10	96.97	95.45	100.00	97.67	96.66	66.82	66.05	98.45	79.06	49.57	5.22
	Audio-RF	VGG-19 + ResNet-10	96.82	100.00	95.00	97.44	96.62	73.71	85.37	70.83	77.42	72.98	4.30
	Visual-RF	MobileNet + ResNet-10	96.97	95.66	99.76	97.67	96.67	72.50	98.57	57.62	72.73	72.50	3.32
GMU	Audio-Visual	LeNet + MobileNet	99.32	98.94	100.00	99.47	99.26	55.53	60.53	86.55	71.24	36.62	2.40
	Audio-RF	VGG-19 + ResNet-10	97.12	100.00	95.48	97.69	96.94	80.53	92.31	75.71	83.19	80.03	4.25
	Visual-RF	MobileNet + MobileNet	95.38	100.00	92.74	96.23	95.13	62.42	63.23	97.86	76.82	38.81	2.27

Late vs. GMU Fusion. We observe that Late Fusion generally achieves slightly higher accuracy in real data, likely due to its simplicity and effective combination of independently strong unimodal predictions. However, GMU Fusion demonstrates a greater capacity to maintain robustness in augmented noisy environments, particularly evident in the audio-RF combination. This suggests that GMU Fusion’s ability to dynamically gate and integrate features at an intermediate level allows it to better adapt to noisy or partially degraded sensor inputs, leading to more reliable detection in challenging scenarios. Furthermore, it is important to acknowledge the slight increase in detection time when moving from unimodal to dual-modality systems, as seen in Table 9. For example, the detection time for the best-performing GMU Fusion audio-RF (VGG-19 + ResNet10-10) is 4.25 ms, slightly higher than the unimodal VGG (2.13 ms) or ResNet-10 (2.42 ms). Despite these trade-offs, dual-modality approaches still struggle to balance high detection accuracy and low false alarms, motivating the exploration of tri-modality fusion for a more robust drone detection framework.

6.4 Proposed Tri-Modal Fusion Analysis

We evaluate our tri-modal fusion approach, integrating audio, visual, and RF data, focusing on the best-performing combination from Table 10. An overview of all combinations is in Appendix C.2. **Best Performing Combination.** The tri-modal fusion approach achieves consistently high performance, matching the near-perfect accuracy of the best dual-modal setups while providing greater stability across different model configurations in real data. As shown in Table 10, both VGG-19 + MobileNet + ResNet-10 (Late Fusion) and LeNet + ResNet-10 + ResNet-10 (GMU Fusion) maintain accuracy above 96.85%. This highlights the effectiveness of tri-modal fusion in leveraging the complementary strengths of audio, visual, and RF data when environmental factors are not disruptive. However, the true advantage of tri-modal fusion emerges in augmented noisy conditions, where it demonstrates significantly enhanced robustness compared to unimodal and dual-modal systems. As shown in Table 10, the best-performing tri-modal model, VGG-19 + MobileNet + ResNet-10 (Late Fusion), achieves 83.26% accuracy and an 85.16% F1-score, outperforming the best dual-modality configuration (audio-RF GMU Fusion) at 80.53% accuracy (Table 9). This highlights the crucial role of a third modality in improving system resilience, enabling the model to compensate for sensor degradation and maintain reliable detection despite environmental interference. **Confusion Matrices.** The confusion matrices in Figure 12 provide a visual perspective on the performance improvements. For Late Fusion (Figure 12a), VGG-19 + MobileNet + ResNet-10 achieves

high TNR (96.9%), ensuring strong background rejection, and a low FPR (3.1%), maintaining balanced classification. Compared to the best dual-modal fusion (Figure 11b), it reduces false positives by nearly 7.9% (FPR: 3.1% vs. 11.0%), improves background rejection (TNR: 96.9% vs. 89.0%), and maintains recall (TPR: 75.5% vs. 75.7%). These improvements make tri-modal fusion more precise and stable, ensuring fewer misclassifications while preserving strong drone detection, making it more reliable in complex environments. For GMU Fusion, LeNet + ResNet-10 + ResNet-10 (Figure 12b) ensures 100% TPR, detecting all drones with no missed instances (FNR: 0%). However, this comes at the cost of a high false alarm rate (FPR: 67.3%) and weak background rejection (TNR: 32.7%), leading to frequent misclassification of non-drone signals as drones.



(a) VGG-19 + MobileNet + ResNet-10. (b) LeNet + ResNet-10 + ResNet-10. **Figure 12: Confusion matrices for tri-modality fusions in noisy augmented data: (blue) Late Fusion; (green) GMU Fusion.**

Late vs. GMU Fusion. In contrast to dual-modality fusion where GMU Fusion showed robustness advantages in noisy augmented data, in tri-modality, Late Fusion emerges as the superior technique in terms of overall accuracy and balanced performance. Late Fusion’s approach of combining decisions from independently optimized unimodal models appears to be particularly effective when integrating three diverse sensor streams, providing both high accuracy and robustness. Despite its added complexity, tri-modal fusion remains efficient, achieving detection times of 5–6 ms (Table 10), only slightly higher than dual-modal fusion while staying well within real-time constraints for drone detection and response. This efficiency extends to energy consumption, with the best combination in Late Fusion consuming 75.27 mJ and GMU Fusion requiring 80.09 mJ. These fusion approaches were tested on the Jetson Orin Nano using the GPU, demonstrating their feasibility for deployment on resource-constrained devices.

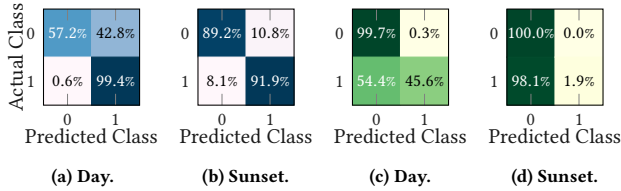
Light and Environmental Conditions Effects. In the following, we evaluate the impact of lighting and environmental conditions on the performance of the proposed tri-modal fusion approach.

① *Light Condition Analysis.* The impact of lighting on the performance of the proposed tri-modal fusion approach was evaluated under noisy augmented data (scenario 2) during both *daytime* and *sunset*. The results in Table 11 and the confusion matrices in Figure

Table 10: Performance metrics for the best performing three-modal fusions under real and noisy augmented data.

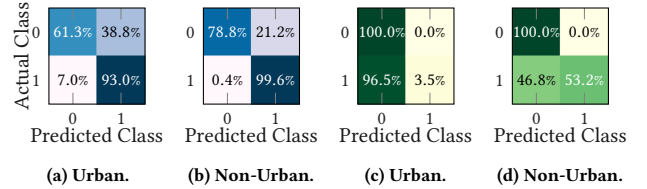
Fusion Type	Modality Combination	Models	Real Data					Noisy Augmented Data					Detec. Time (ms)	Ener. Cons. (mJ)
			Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)		
Late	Audio-Visual-RF	VGG-19 + MobileNet + ResNet-10	96.89	95.35	100.00	97.62	96.58	83.26	97.69	75.48	85.16	82.98	6.09	75.27
GMU	Audio-Visual-RF	LeNet + ResNet-10 + ResNet-10	96.97	95.45	100.00	97.67	96.66	75.53	72.23	100.00	83.87	66.58	6.48	80.09

13 provide deeper insight into how illumination variations influence detection performance. The sunset conditions generally provide more favorable and stable illumination, leading to improved performance across all fusion strategies. Under **Late Fusion**, the VGG-19 + MobileNet + ResNet-10 combination performed best at sunset, achieving 89.84% accuracy, with strong background rejection (TNR: 91.9%) and a low false alarm rate (FPR: 8.1%). The consistent lighting conditions at sunset help reduce visual challenges, such as glare and high-contrast shadows, leading to higher recall (TPR: 89.2%), fewer missed detections (FNR: 10.8%), and a high F1-score (92.94%). During daytime, accuracy dropped to 77.06%, with a significant decline in drone detection (TPR: 57.2%) and increased missed detections (FNR: 42.8%). However, background rejection remained strong (TNR: 99.4%), minimizing false alarms. The fluctuating light intensity and shadows degraded video-based detection, but audio and RF data helped compensate, ensuring more consistent overall performance. This highlights the effectiveness of multi-sensor fusion in mitigating modality-specific weaknesses, making the system more resilient to environmental variations. **GMU Fusion** maintains high recall across lighting conditions but struggles with false positives. During daytime, it achieves 74.26% accuracy, detecting almost all drones (TPR: 99.7%) with minimal missed detections (FNR: 0.3%). However, background rejection is weak (TNR: 45.6%), leading to a high false alarm rate (FPR: 54.4%). At sunset, accuracy improves slightly to 75.47%, with perfect drone detection (TPR: 100%) and no missed detections (FNR: 0%). However, background rejection deteriorates significantly (TNR: 1.9%), resulting in an extremely high false alarm rate (FPR: 98.1%), making it unreliable in environments that require precise classification.

**Figure 13: Confusion matrices showing the impact of lighting conditions across three modalities for Late (blue) and GMU Fusion (green).**

② **Location-Based Analysis.** The results in Table 11 and Figure 14 show how urban and non-urban environments affect the tri-modal fusion approach under noisy augmented data (scenario 2). Non-urban settings offer higher accuracy and fewer false alarms across all fusion strategies, benefiting from cleaner RF conditions, clearer audio signals, and a more uniform visual background. Specifically, under **Late Fusion**, detection is significantly stronger in non-urban environments, achieving a higher accuracy (84.90%), supported by strong background rejection (TNR: 99.6%) and a minimal false alarm rate (FPR: 0.4%). Drones are detected more reliably, with a higher TPR (78.8%) and fewer missed detections (FNR: 21.2%). In contrast, urban environments introduce more noise and interference, leading

to a drop in accuracy (78.89%) and a decrease in drone detection (TPR: 61.3%), with more missed instances (FNR: 38.8%). However, background rejection remains strong (TNR: 93.0%), ensuring that non-drone samples are still well classified despite the increased complexity of the scene. **GMU Fusion**, while maintaining perfect drone detection (TPR: 100%) in both environments, suffers from extremely high false alarms. In urban areas, performance drops drastically (46.39%) due to severe background misclassification (TNR: 3.5%) and an overwhelming false alarm rate (FPR: 96.5%). While every drone is detected (TPR: 100%, FNR: 0%), the model struggles to differentiate drones from background clutter. In non-urban areas, it achieves an accuracy of 86.35%, benefiting from better background separation (TNR: 53.2%) and a lower false alarm rate (FPR: 46.8%) compared to urban environments. These results highlight the trade-off between fusion strategies: Late Fusion offers a more balanced approach, maintaining strong detection performance and minimizing false alarms, making it better suited for urban settings. In contrast, GMU Fusion prioritizes recall, making it more effective in non-urban environments where false positives are less critical.

**Figure 14: Confusion matrices illustrating the impact of locations across three modalities for Late (blue) and GMU Fusion (green).**

7 Limitations and Discussion

Dataset Expansion. Our dataset focuses on capturing synchronized audio, visual, and RF data in diverse real-world environments, providing a strong basis for multi-modal drone detection. It emphasizes small UAVs, such as the DJI Mini series, and reflects common operational conditions. However, including other types of drones, particularly larger models with different acoustic and RF characteristics, along with additional environmental factors such as rain, fog, and dense vegetation, may further enhance generalizability. Despite this, the dataset remains a valuable resource, offering truly synchronized multi-modal data in uncontrolled settings and serving as a strong benchmark for small UAV detection.

Adversarial Robustness Considerations. While our data augmentation enhances TRIDENT’s resilience to natural noise, intentional adversarial attacks by a knowledgeable adversary remain a challenge. An attacker aware of TRIDENT’s tri-modal fusion could strategically manipulate sensor reliability to exploit the system. For example, subtle visual camouflage, targeted audio masking mimicking background sounds, or intermittent RF interference designed to resemble legitimate signals could be employed. These modality-specific manipulations aim to degrade individual sensor reliability,

Table 11: Performance metrics for the top three-modal fusions combos across lighting (day/sunset) and location (urban/non-urban).

Type	Fusion Models Combination	Lighting Condition										Location									
		Day					Sunset					Urban					Non-Urban				
		Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)
Late	VGG-19 + MobileNet + ResNet-10	77.06	99.04	57.22	72.54	76.42	89.84	97.05	89.17	92.94	87.42	78.89	87.50	61.25	72.06	77.55	84.90	99.81	78.82	88.09	83.73
GMU	LeNet + ResNet-10 + ResNet-10	74.26	67.35	99.72	80.40	71.46	75.47	75.35	100.00	85.94	44.81	46.39	45.33	100.00	62.38	34.57	86.35	83.85	100.00	91.21	80.34

causing TRIDENT to misweigh sensor inputs within the fusion process and increasing wrong detections or false alarms. Defending against such intelligent adversarial exploitation, which targets sensor vulnerabilities to perturb the fusion mechanism, is crucial for future work, requiring techniques to enhance sensor-level robustness and develop more resilient fusion strategies.

Counter-UAV Measures. This study establishes a strong foundation for drone detection, a key element of counter-UAV systems. Integrating detection with localization, tracking, and mitigation could significantly improve response time and situational awareness, allowing for more effective countermeasures. While these aspects fall outside the scope of this work, a highly reliable detection system remains a critical first step toward a comprehensive counter-UAV strategy.

8 Related Work

Dataset Diversity. Many existing drone detection datasets consist of synthetic data or are collected in controlled environments, such as laboratories, test sites, or airport runways, where sensor-to-target visibility is high, background noise is minimal, and RF interference is controlled [8, 30, 50]. These conditions fail to reflect real-world challenges, where urban noise, lighting variations, occlusions, and overlapping RF signals significantly impact detection performance [10, 57]. In contrast, TRIDENT introduces a novel dataset collected in real-world, uncontrolled environments, covering both urban and non-urban locations under varying lighting conditions (daylight and sunset). This deliberate focus on real-world data acquisition, with synchronized multi-modal sensor streams, enhances the generalizability of our TRIDENT framework.

Data Augmentation and Testing. Data augmentation is crucial for enhancing the robustness of drone detection models. Existing methods typically apply augmentation to only one sensor at a time, such as adding noise to audio or applying visual transformations independently [4, 29]. However, this approach fails to capture synchronized multi-modal degradations, where real-world factors simultaneously impact multiple sensors, leading to compromised detection accuracy. Additionally, some approaches train models on synthetically augmented data but evaluate them on non-augmented or minimally augmented test sets [3, 56]. This mismatch creates a performance gap, as models trained under these conditions often struggle with severe real-world noise, making them unreliable for practical deployment [14]. Unlike existing approaches, TRIDENT applies synchronized data augmentation across audio, visual, and RF modalities *exclusively during testing*. This approach enables a rigorous evaluation of system performance under varying levels of multi-modal noise (scenario 1 and scenario 2), effectively measuring the lower bound of accuracy when all sensors experience simultaneous degradation. Importantly, TRIDENT models are trained solely on *real, non-augmented data*, ensuring that robustness improvements are not a result of overfitting to specific synthetic noise patterns

but rather reflect the true effectiveness of multi-modal fusion in handling unseen, real-world degradations.

Multi-Modality Strategy. Early UAV detection relied on single-modality systems, which are highly sensitive to environmental noise, lighting conditions, and RF interference [7, 20, 34, 45]. Dual-modality approaches improve resilience by combining two complementary sensors, but they still fail when both modalities are simultaneously degraded [28, 29, 51]. TRIDENT advances the state-of-the-art with a tri-modal (audio-visual-RF) fusion framework, leveraging the strengths of all three modalities to enhance detection robustness in challenging conditions. Unlike existing systems that utilize simpler fusion techniques, TRIDENT incorporates sophisticated architectures that dynamically learn the optimal contribution of each modality during the training process. Approaches relying on basic OR-function logic for detection [18, 46], equally weighted modalities [52], multinomial logistic regression for fixed probability combination [29], or static ensemble stacking [33] inherently lack the adaptability of TRIDENT. These simpler methods are unable to effectively account for the varying reliability of each sensor under different environmental conditions. In contrast, TRIDENT’s learned fusion weights enable intelligent, context-aware sensor integration, significantly enhancing performance in real-world scenarios.

Real-Time and Energy-Efficiency. Real-time drone detection is crucial for timely responses in security and surveillance applications. Many existing systems, while achieving high accuracy, may not fully address the computational constraints of real-time deployment, or lack detailed reporting on detection times [5, 35, 37, 49]. Some works focus on optimizing models for edge devices, but may compromise accuracy for efficiency [48]. TRIDENT prioritizes both accuracy and real-time feasibility, with its best-performing Late Fusion model achieving a detection time of 6.09 ms per segment, ensuring suitability for real-time applications. Energy efficiency was evaluated on Jetson Orin Nano, measuring 75.27 mJ for Late Fusion and 80.09 mJ for GMU Fusion, demonstrating its capability for edge deployment. This detailed assessment of inference time and energy consumption, often overlooked in related work, highlights TRIDENT’s practical deployability for real-world drone detection.

9 Conclusions

This paper shows that robust drone detection in real-world environments requires multi-sensor data fusion. We present TRIDENT, a tri-modal framework for real-time UAV detection that: (i) achieves 96.89% accuracy in real data and 83.26% in noisy augmented data, outperforming unimodal and dual-modal baselines; (ii) remains robust under real-world noise scenarios through targeted data augmentation; (iii) operates in real-time (6.09 ms detection) with low energy consumption (75.27 mJ per detection), making it suitable for resource-constrained devices; and (iv) generalizes well to diverse urban and non-urban environments under varying lighting conditions, ensuring reliable performance under adverse situations.

Acknowledgments

This research was made possible with the support of the Horizon Europe research and innovation programme of the European Union, under grant agreement number 101092912 (project ML-SysOps). The authors also gratefully acknowledge the Intelligent and Autonomous Systems Lab (IASL) at the University of California, Irvine (UCI), particularly Sharon Ladron De Guevara Contreras and Marco Levorato for their support and collaboration during the data collection process.

References

- [1] Airsight. 2018. *Companies and Data Centers Should Consider Drones When Assessing Security Risks*. <https://www.airsight.com/blog/companies-and-data-centers-should-consider-drones-when-assessing-security-risks>
- [2] M Eren Akbiyik. 2023. Data augmentation in training CNNs: injecting noise to images. *arXiv preprint arXiv:2307.06855* (2023).
- [3] Fatih Cagatay Akyon, Ogulcan Eryuksel, Kamil Anil Ozfuttu, and Sinan Onur Altinuc. 2021. Track boosting and synthetic data aided drone detection. In *2021 17th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–5.
- [4] Ildi Alla, Hervé B. Olou, Valeria Loscri, and Marco Levorato. 2024. 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) (*WiSec '24*). 123–133.
- [5] Mhd Saria Allahham, Tamer Khattab, and Amr Mohamed. 2020. Deep learning for RF-based drone detection and identification: A multi-channel 1-D convolutional neural networks approach. In *2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*. IEEE, 112–117.
- [6] Rakesh John Amala Arokia Nathan, Indrajit Kurmi, and Oliver Bimber. 2023. Drone swarm strategy for the detection and tracking of occluded targets in complex environments. *Communications Engineering* 2, 1 (2023), 55.
- [7] Muhammad Zohaib Anwar, Zeeshan Kaleem, and Abbas Jamalipour. 2019. Machine learning inspired sound-based amateur drone detection for public safety applications. *IEEE Transactions on Vehicular Technology* 68, 3 (2019), 2526–2534.
- [8] Sanjoy Basak, Sreeraj Rajendran, Sofie Pollin, and Bart Scheers. 2021. Combined RF-based drone detection and classification. *IEEE Transactions on Cognitive Communications and Networking* 8, 1 (2021), 111–120.
- [9] BBC News. 2023. Drone sighting temporarily shuts runways at New York airport. *BBC News* (15 May 2023). <https://www.bbc.com/news/uk-england-sussex-65591023>
- [10] Abdelouahid Bentamou, Abbass Zein-Eddine, Oussama Messai, and Yann Gavet. 2023. Real-Time Drone Detection and Tracking in Distorted Infrared Images. In *2023 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*. IEEE, 3594–3598.
- [11] Dell Cameron. 2024. Intel Officials Warned Police That US Cities Aren't Ready for Hostile Drones. *WIRED* (17 December 2024). <https://www.wired.com/story/intel-officials-police-us-cities-drones-dhs/>
- [12] Pietro Casabianca and Yu Zhang. 2021. Acoustic-based UAV detection using late fusion of deep neural networks. *Drones* 5, 3 (2021), 54.
- [13] Cybersecurity and Infrastructure Security Agency (CISA). 2024. *Unauthorized Drone Activity Over Sporting Venues*. Technical Report. Cybersecurity and Infrastructure Security Agency (CISA). <https://www.cisa.gov/sites/default/files/2024-03/Unauthorized%20Drone%20Activity%20Over%20Sporting%20Venues.pdf>
- [14] Tamara Regina Dieter, Andreas Weinmann, Stefan Jäger, and Eva Brucherseifer. 2023. Quantifying the simulation–reality gap for deep learning-based drone detection. *Electronics* 12, 10 (2023), 2197.
- [15] Siyi Ding, Xiao Guo, Ti Peng, Xiao Huang, and Xiaoping Hong. 2023. Drone detection and tracking system based on fused acoustical and optical approaches. *Advanced Intelligent Systems* 5, 10 (2023), 2300251.
- [16] DJI. 2023. DJI Mini 2 SE - Specs. <https://www.dji.com/fr/mini-2-se/specs>. Accessed: 2025-02-28.
- [17] DJI. 2023. DJI Mini 3 Pro - Specs. <https://www.dji.com/fr/mini-3-pro/specs>. Accessed: 2025-02-28.
- [18] Qiusi Dong, Yu Liu, and Xiaolin Liu. 2023. Drone sound detection system based on feature result-level fusion using deep learning. *Multimedia Tools and Applications* 82, 1 (2023), 149–171.
- [19] Mohamed Emimi, Mohamed Khaleel, and Abobakr Alkrash. 2023. The current opportunities and challenges in drone technology. *Int. J. Electr. Eng. and Sustain.* (2023), 74–89.
- [20] Martins Ezuma, Fatih Erden, Chethan Kumar Anjinappa, Ozgur Ozdemir, and Ismail Guvenc. 2019. Detection and classification of UAVs using RF fingerprints in the presence of Wi-Fi and Bluetooth interference. *IEEE Open Journal of the Communications Society* 1 (2019), 60–76.
- [21] Alan Frid, Yehuda Ben-Shimol, Erez Manor, and Shlomo Greenberg. 2024. Drones detection using a fusion of RF and acoustic features and deep neural networks. *Sensors* 24, 8 (2024), 2427.
- [22] Vipul Garg, Suman Niranjana, Victor Prybutok, Terrance Pohlen, and David Gligor. 2023. Drones in last-mile delivery: A systematic review on Efficiency, Accessibility, and Sustainability. *Transportation Research Part D: Transport and Environment* 123 (2023), 103831.
- [23] Stefan Glüge, Matthias Nyfeler, Ahmad Aghaebrahimian, Nicola Ramagnano, and Christof Schüpbach. 2024. Robust Low-Cost Drone Detection and Classification Using Convolutional Neural Networks in Low SNR Environments. *IEEE Journal of Radio Frequency Identification* (2024).
- [24] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.
- [25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [26] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. 2017. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv preprint arXiv:1704.04861* (2017).
- [27] Yuzhuo Huang, Jingyi Qu, Haoyu Wang, and Jun Yang. 2024. An All-Time Detection Algorithm for UAV Images in Urban Low Altitude. *Drones* 8, 7 (2024), 332.
- [28] Snezhana Jovanoska, Benjamin Knoedler, D Palur Palanivelu, Luisa Still, T Fiolka, M Oispuu, C Steffes, and W Koch. 2021. Passive sensor processing and data fusion for drone detection. In *Proceedings of the NATO STO Meeting Proceedings: MSG-SET-183 Specialists' Meeting on Drone Detectability: Modelling the Relevant Signature*. 16.
- [29] Hunje Lee, Sujeong Han, Jeong-Il Byeon, Seougyu Han, Rangun Myung, Jongun Joung, and Jihoon Choi. 2023. CNN-based UAV detection and classification using sensor fusion. *IEEE Access* 11 (2023), 68791–68808.
- [30] Tamara R Lenhard, Andreas Weinmann, Kai Franke, and Tobias Koch. 2024. SynDroneVision: A Synthetic Dataset for Image-Based Drone Detection. *arXiv preprint arXiv:2411.05633* (2024).
- [31] Ziyi Liu, Pei An, You Yang, Shaohua Qiu, Qiong Liu, and Xinghua Xu. 2024. Vision-Based Drone Detection in Complex Environments: A Survey. *Drones* 8, 11 (2024), 643.
- [32] Marshall Electronics. 2024. CV505-MB/M Full-HD (3G/HD-SDI) 2.5MP Mini-Broadcast POV Camera. <https://marshall-usa.com/discontinued/cameras/CV505-MB.php>. Accessed: 2025-02-28.
- [33] James McCoy and Danda B Rawat. 2024. Optimized Machine Learning Based Multimodal UAV Detection Using Ensemble Stacking. In *2024 IEEE 6th International Conference on Cognitive Machine Intelligence (CogMI)*. IEEE, 40–49.
- [34] Olusiji O Medaiyese, Martins Ezuma, Adrian P Lauf, and Ismail Guvenc. 2021. Semi-supervised learning framework for UAV detection. In *2021 IEEE 32nd Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE, 1185–1190.
- [35] Olusiji O Medaiyese, Abbas Syed, and Adrian P Lauf. 2021. Machine learning framework for RF-based drone detection and identification system. In *2021 2nd International Conference On Smart Cities, Automation & Intelligent Computing Systems (ICON-SONICS)*. IEEE, 58–64.
- [36] Maham Misbah, Misha Urooj Khan, Zeeshan Kaleem, Ali Muqabail, Muhammad Z Alam, Ran Liu, and Chau Yuen. 2024. MSF-GhostNet: Computationally-Efficient YOLO for Detecting Drones in Low-Light Conditions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* (2024).
- [37] Yongguang Mo, Jianjun Huang, and Gongbin Qian. 2022. Deep learning approach to UAV detection and classification by using compressively sensed RF signal. *Sensors* 22, 8 (2022), 3072.
- [38] Manel Mrabet, Maha Sliti, and Lassaad Ben Ammar. 2024. Machine learning algorithms applied for drone detection and classification: benefits and challenges. *Frontiers in Communications and Networks* 5 (2024), 1440727.
- [39] Trong Thanh Nguyen, Le Cuong Nguyen, et al. 2025. An Effective RF-Based Solution for Drone Detection and Recognition Amid Noise, Bluetooth, and Wi-Fi Interference. (2025).
- [40] Mattes Ohlenbusch, Aike Ahrens, Christian Rollwage, and Jörg Bitzer. 2021. Robust drone detection for acoustic monitoring applications. In *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 6–10.
- [41] Gokulnath M Pillai, Ajith Suresh, Eikansh Gupta, Vinod Ganapathy, and Arpita Patra. 2024. Privadome: Delivery Drones and Citizen Privacy. *Proceedings on Privacy Enhancing Technologies* (2024).
- [42] Jane Saldanha, Shaunak Chakraborty, Shruti Patil, Ketan Kotecha, Satish Kumar, and Anand Nayyar. 2022. Data augmentation using Variational Autoencoders for improvement of respiratory disease classification. *Plos one* 17, 8 (2022), e0266467.
- [43] Samson Technologies. 2024. Meteor Mic - USB Microphone. <https://samsontech.com/products/microphones/usb-microphones/meteormic/>. Accessed: 2025-02-28.

- [44] Guanxiong Shen, Junqing Zhang, Alan Marshall, Mikko Valkama, and Joseph R Cavallaro. 2023. Toward length-versatile and noise-robust radio frequency fingerprint identification. *IEEE Transactions on Information Forensics and Security* 18 (2023), 2355–2367.
- [45] Lin Shi, Ishtiaq Ahmad, YuJing He, and KyungHi Chang. 2018. Hidden Markov model based drone sound recognition using MFCC technique in practical noisy environments. *Journal of Communications and Networks* 20, 5 (2018), 509–518.
- [46] Xiufang Shi, Chaoqun Yang, Weige Xie, Chao Liang, Zhiguo Shi, and Jiming Chen. 2018. Anti-drone system with multiple surveillance technologies: Architecture, implementation, and challenges. *IEEE Communications Magazine* 56, 4 (2018), 68–74.
- [47] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [48] Zhifan Song, Yuan Zhang, and Abd Al Rahman M Abu Ebayyeh. 2025. ED-Net: Edge-Optimized Small Target Detection in UAV Imagery–Faster Context Attention, Better Feature Fusion, and Hardware Acceleration. *arXiv preprint arXiv:2501.05885* (2025).
- [49] K Sricharan and M Venkat. 2023. Real-Time Drone Detection Using Deep Learning. In *Second International Conference on Emerging Trends in Engineering (ICETE 2023)*. Atlantis Press, 905–918.
- [50] Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund. 2021. A dataset for multi-sensor drone detection. *Data in Brief* 39 (2021), 107521.
- [51] Fredrik Svanström, Fernando Alonso-Fernandez, and Cristofer Englund. 2022. Drone detection and tracking in real-time by fusion of different sensing modalities. *Drones* 6, 11 (2022), 317.
- [52] Fredrik Svanström, Cristofer Englund, and Fernando Alonso-Fernandez. 2021. Real-time drone detection and tracking with visible, thermal and acoustic sensors. In *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 7265–7272.
- [53] Clayton Swope. 2024. Why Are There So Many Unexplained Drones Flying Over the United States? *Center for Strategic and International Studies* (16 December 2024). <https://www.csis.org/analysis/why-are-there-so-many-unexplained-drones-flying-over-united-states>
- [54] Pietro Tedeschi, Siva Ganesh Ganti, and Savio Sciancalepore. 2024. Selective authenticated pilot location disclosure for remote ID-enabled drones. *Proceedings on Privacy Enhancing Technologies* (2024).
- [55] The Guardian. 2024. First UK arrest made over filming of women on nights out without consent. *The Guardian* (27 November 2024). <https://www.theguardian.com/uk-news/2024/nov/27/first-uk-arrest-made-over-filming-of-women-on-nights-out-without-consent>
- [56] Mariusz Wisniewski, Zeeshan A Rana, Ivan Petrunin, Alan Holt, and Stephen Harman. 2024. Drone Detection using Deep Neural Networks Trained on Pure Synthetic Data. *arXiv preprint arXiv:2411.09077* (2024).
- [57] Lu Xu, Rui Shi, and Yijia Zhang. 2025. A Radio-Frequency Sensor Based UAV Detection and Identification System Using Improved Vision Transformer Based Model. *IEEE Sensors Journal* (2025).

A Technical Specifications of Drones and Sensing Equipment

This appendix provides detailed technical specifications for the drones (Figure 15) used in our experiments and the sensing equipment employed for data collection. This information is intended to offer a comprehensive understanding of the experimental setup and the technical capabilities of the key components used in this study.



Figure 15: UAVs and their radio controller used during experiments.

① **Drone Specifications.** We utilized two commercially available drones, the DJI Mini 2 and DJI Mini 3 Pro, to evaluate our TRIDENT system. These models were selected due to their subtle yet important differences in characteristics such as size, acoustics, and operational capabilities, which pose varying challenges for detection systems. Table 12 presents a detailed comparison of their key specifications, highlighting the nuances that informed our experimental design.

Table 12: Technical characteristics of drones used in experiments. [16, 17].

Parameter	DJI Mini 2	DJI Mini 3 Pro
Size [mm]	138×81×58	145×90×62
Weight [g]	249	
Material	Plastic and metal components; non-reflective and discernible surfaces	
Sound Propeller [dB/m]	74	72
Diagonal Length [mm]	213	247
Operating Frequency [GHz]	2.400-2.4835, 5.725-5.850	
Max. Transmission Distance [km]	10 (FCC), 6 (CE), 6 (SRRC), 6 (MIC)	12 (FCC), 8 (CE), 8 (SRRC), 8 (MIC)
Max Speed [m/s]	16 (S Mode), 10 (N Mode), 6 (C Mode)	

As shown in Table 12, both drones share a lightweight design at 249g, constructed from plastic and metal components with non-reflective surfaces to minimize visual detectability. Subtle differences exist in their physical dimensions and acoustic profiles, with the DJI Mini 3 Pro being slightly larger and quieter (72 dB/m) than the DJI Mini 2 (74 dB/m). Both operate in the common 2.4 GHz and 5.8 GHz frequency bands and exhibit similar maximum speeds, but the DJI Mini 3 Pro offers an extended maximum transmission distance. These variations are crucial for evaluating the TRIDENT’s ability to discern drones with differing acoustic and signal characteristics.

② **Sensing Equipment Specifications.** Our data collection system is designed to capture synchronized audio, visual, and RF data streams, utilizing a suite of sensors. Figure 16 illustrates the primary sensing equipment setup, while the subsequent tables detail the technical specifications of each sensor component.



Figure 16: Sensing equipment used during the experiments.

• **Audio Sensor.** We employed the Samson Meteor condenser microphone for audio data acquisition, chosen for its high sensitivity and broad frequency response, crucial for capturing the subtle acoustic signatures of small drones. Table 13 outlines the key specifications of this microphone.

Table 13: Specifications of the Samson Meteor microphone [43].

Parameter	Value
Type	Condenser
Polar Pattern	Cardioid
Frequency Response	20Hz - 20kHz
Sample Rate	44.1/48kHz
Bit Rate	16-bit
Sensitivity	-33 dB/Pa

The Samson Meteor is a condenser-type microphone with a cardioid polar pattern, effectively capturing sound sources directly in front while minimizing background noise. Its wide frequency response (20Hz-20kHz) ensures comprehensive capture of the drone’s

acoustic spectrum. It records high-quality audio at sample rates of 44.1 kHz or 48 kHz with a 16-bit bit depth, and features a sensitivity of -33 dB/Pa, enabling it to detect faint drone sounds even in moderately noisy environments.

Table 14: Specifications of the Marshall CV-505 camera [32].

Parameter	Value
Image Sensor	2.5 Megapixel 1/3-inch CMOS
Lens	3.7mm
Effective Pixel	1920(H) x 1080(V) 0.2 Lux (Color), 0.1 Lux (Black/White), 0.005 Lux (DSS on)
Min. Illumination	0.1 Lux (Black/White), 0.005 Lux (DSS on)
Power Supply	12V
Special Features	WDR function, RS-485 remote control
Operation Temperature	-10°C ~ 50°C

• **Video Sensor.** Visual data was captured using the Marshall CV-505 camera, selected for its compact form factor and high-resolution imaging capabilities, suitable for drone detection in diverse lighting conditions. The detailed specifications of the Marshall CV-505 camera are provided in Table 14.

The Marshall CV-505 camera utilizes a 2.5 Megapixel 1/3-inch CMOS sensor and a 3.7mm lens, capturing high-resolution video with 1920x1080 effective pixels. Its minimum illumination sensitivity, as low as 0.005 Lux in DSS mode, ensures effective video capture even in low-light conditions. The camera also features wide dynamic range (WDR) to handle scenes with high contrast lighting and RS-485 remote control for flexible operation.

• **RF Sensor.** We utilized the Universal Software Radio Peripheral (USRP) B210 for RF signal acquisition, a software-defined radio known for its wide frequency range and high sampling rate, essential for capturing drone communication signals. Table 15 details the specifications of the USRP B210.

Table 15: Specifications of the USRP B210 RF receiver.

Parameter	Value
Frequency Range	70 MHz - 6 GHz
Instantaneous Bandwidth	~ 56 MHz
Maximum I/Q sample rate	61.44 MS/s
Number of channels	2
Maximum Receive Gain	76 dB
Maximum Transmit Gain	89.8 dB
ADC resolution	12 bits

The USRP B210 offers a wide frequency range from 70 MHz to 6 GHz, covering typical drone communication bands. It supports an instantaneous bandwidth of approximately 56 MHz and a maximum I/Q sample rate of 61.44 MS/s, enabling high-fidelity capture of RF signals. With 2 channels and significant gain in both receive (76 dB) and transmit (89.8 dB), the USRP B210 is well-suited for detecting weak drone signals amidst background RF noise. The 12-bit ADC resolution ensures accurate digitization of the received RF signals for subsequent processing and analysis.

B Dataset Overview

The dataset collected for this study includes synchronized data from audio, video, and RF sensors, divided into two classes: *drone* and *no drone*. Each data sample consists of ten-second clips, captured within a range of 1 to 50 meters.

Audio Dataset. The audio dataset comprises 277 ten-second clips recorded at a sample rate of 44.1 kHz. The *drone* class captures

the characteristic sounds of drone flight, while the *no drone* class specifically features general outdoor background sounds collected at the acquisition sites. These include noises from vehicles, voices, and other ambient sounds, providing a robust baseline for distinguishing drone-related noises from environmental sounds.

Video Dataset. The video dataset consists of 83,100 frames at a resolution of 640×640 pixels. The *drone* class captures drone activity at distances up to 50 meters, while the *no drone* class includes images of non-drone objects such as clouds, helicopters, birds, buildings, and trees. This variety ensures the system can accurately differentiate drones from other airborne objects and background elements.

RF Dataset. The RF dataset contains I/Q samples recorded at a sampling rate of 55 MS/s in the 2.4 GHz ISM band. These samples were saved in binary files (277 files) and later processed into spectrograms using the STFT, resulting in 11,080 images with a resolution of 640×640 pixels. The *drone* class in the RF dataset captures the characteristic spectral signatures of drone communication signals. The *no drone* class encompasses a diverse range of background RF noise and interference from other signal sources.

C Detailed Analysis of the Fusion Performance

This appendix provides a detailed analysis of the performance metrics for all dual-modality and tri-modality fusion combinations evaluated in this study, extending the discussion presented in the paper. This comprehensive analysis serves to justify the selection of the best-performing models and offers a broader understanding of the comparative effectiveness of different fusion strategies and modality combinations.

C.1 Dual-Modal Fusion

Table 16 presents the performance metrics for all evaluated dual-modality fusion models under both real and noisy augmented data, categorized by fusion type and modality combination.

① **Late Fusion.** In real data, all dual-modality combinations demonstrate high accuracy, generally above 96%, with audio-visual (LeNet + MobileNet) achieving a 100% accuracy. This indicates that Late Fusion effectively leverages complementary information from different modalities when data quality is high. In noisy augmented conditions, however, performance varies significantly.

• **Audio-Visual.** While achieving perfect accuracy in real data with LeNet + MobileNet, performance drops to 65.23% in noisy augmented conditions. The VGG-19 + ResNet-10 combination shows slightly better noisy condition accuracy at 66.82%, suggesting VGG-19 and ResNet-10 models are somewhat more robust to noise in a Late Fusion setup for audio-visual data, although still significantly degraded compared to real data. The other audio-visual combinations show even lower noisy condition performance, indicating that simply combining audio and visual modalities via Late Fusion does not inherently guarantee robustness in noisy environments.

• **Audio-RF.** This modality combination generally exhibits more consistent performance across real and noisy augmented data compared to audio-visual. The VGG-19 + ResNet-10 achieves the highest noisy condition accuracy among all dual-modality combinations, at 73.71%. This suggests that combining audio and RF modalities in a Late Fusion manner provides a more robust approach to noise compared to audio-visual combinations. While LeNet + MobileNet achieves a slightly higher real data accuracy (98.03%) compared

Table 16: Performance metrics for two-modal fusions under real and noisy augmented data.

Fusion Type	Modality Combination	Models	Real Data					Noisy Augmented Data					Detec. Time (ms)
			Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	
Late	Audio-Visual	LeNet + ResNet-10	96.97	95.45	100.00	97.67	96.66	36.36	0.00	0.00	0.00	26.67	3.26
		LeNet + MobileNet	100.00	100.00	100.00	100.00	100.00	65.23	70.68	77.50	73.94	60.86	2.55
		VGG-19 + ResNet-10	96.97	95.45	100.00	97.67	96.66	66.82	66.05	98.45	79.06	49.57	5.22
		VGG-19 + MobileNet	96.97	95.45	100.00	97.67	96.66	56.36	100.00	31.43	47.83	55.16	3.61
	Audio-RF	LeNet + ResNet-10	97.20	99.88	95.71	97.75	97.02	65.53	64.91	99.76	78.65	44.63	3.04
		LeNet + MobileNet	98.03	99.39	97.50	98.44	97.89	63.64	63.64	100.00	77.78	38.89	3.40
		VGG-19 + ResNet-10	96.82	100.00	95.00	97.44	96.62	73.71	85.37	70.83	77.42	72.98	4.30
		VGG-19 + MobileNet	96.67	99.38	95.36	97.33	96.45	63.64	63.64	100.00	77.78	38.89	3.58
	Visual-RF	ResNet-10 + ResNet-10	96.97	95.45	100.00	97.67	96.66	63.18	64.11	95.71	76.79	43.89	4.29
		ResNet-10 + MobileNet	96.97	95.45	100.00	97.67	96.66	63.64	63.64	100.00	77.78	38.89	4.36
		MobileNet + ResNet-10	96.97	95.66	99.76	97.67	96.67	72.50	98.57	57.62	72.73	72.50	3.32
		MobileNet + MobileNet	99.62	100.00	99.40	99.70	99.59	62.42	63.23	97.86	76.82	38.81	2.25
GMU	Audio-Visual	LeNet + ResNet-10	89.77	97.96	85.71	91.43	89.38	63.64	63.64	100.00	77.78	38.89	3.36
		LeNet + MobileNet	99.32	98.94	100.00	99.47	99.26	55.53	60.53	86.55	71.24	36.62	2.40
		VGG-19 + ResNet-10	96.97	95.45	100.00	97.67	96.66	42.35	78.01	13.10	22.43	38.28	4.53
		VGG-19 + MobileNet	96.97	95.45	100.00	97.67	96.66	36.36	0.00	0.00	0.00	26.67	3.63
	Audio-RF	LeNet + ResNet-10	97.05	99.88	95.48	97.63	96.86	59.47	63.70	84.40	72.61	47.37	3.09
		LeNet + MobileNet	97.05	99.51	95.83	97.63	96.85	63.64	63.64	100.00	77.78	38.89	2.37
		VGG-19 + ResNet-10	97.12	100.00	95.48	97.69	96.94	80.53	92.31	75.71	83.19	80.03	4.25
		VGG-19 + MobileNet	94.17	95.36	95.48	95.42	93.70	61.89	63.62	93.69	75.78	43.22	3.43
	Visual-RF	ResNet-10 + ResNet-10	96.29	99.25	94.88	97.02	96.05	59.92	63.68	86.19	73.24	46.73	3.65
		ResNet-10 + MobileNet	96.74	99.38	95.48	97.39	96.53	60.76	62.70	94.64	75.43	39.03	4.56
		MobileNet + ResNet-10	96.97	95.45	100.00	97.67	96.66	47.42	100.00	17.38	29.61	43.83	3.16
		MobileNet + MobileNet	95.38	100.00	92.74	96.23	95.13	62.42	63.23	97.86	76.82	38.81	2.27

to VGG-19 + ResNet-10 (96.82%), its noisy condition performance (63.64%) is considerably lower, further emphasizing the robustness of the VGG-19 + ResNet-10.

- **Visual-RF.** Visual-RF combinations show strong performance in real data, with MobileNet + MobileNet reaching 99.62% accuracy. However, in noisy conditions, their performance is comparable to or slightly lower than audio-visual. The MobileNet + ResNet-10 combination achieves the highest noisy condition accuracy among visual-RF at 72.50%, which is still slightly below the best audio-RF (VGG-19 + ResNet-10 at 73.71%). This indicates that while visual-RF fusion is effective, it does not offer superior robustness in noise compared to audio-RF fusion.

- ② **GMU Fusion.** In real data, GMU Fusion combinations also achieve high accuracy, although generally slightly lower than the best Late Fusion counterparts. However, GMU Fusion demonstrates a different trend in noisy conditions.

- **Audio-Visual.** In noisy augmented conditions, the audio-visual of GMU Fusion performance is generally lower than audio-visual in Late Fusion. The best-performing combination, LeNet + ResNet-10, achieves 63.64% accuracy, significantly lower than the best audio-visual Late Fusion (66.82%). This suggests that GMU Fusion may not be as effective as Late Fusion for audio-visual data in high noisy environments. Notably, the VGG-19 + ResNet-10 combination performs particularly poorly with GMU Fusion in noisy augmented conditions (42.35% accuracy), indicating that complex models combined with GMU Fusion may not generalize well for audio-visual data under noise.

- **Audio-RF.** Audio-RF fusion stands out as the most robust dual-modality approach in noisy conditions. The VGG-19 + ResNet-10 combination achieves the highest noisy condition accuracy among

all dual-modality combinations at 80.53%. This significantly outperforms all Late Fusion dual-modality combinations in noise and demonstrates the effectiveness of GMU Fusion in leveraging audio and RF data for robust drone detection. While LeNet + MobileNet achieves a lower noisy condition accuracy (63.64%), it is still comparable to the most Late Fusion combinations.

- **Visual-RF.** Visual-RF combinations show relatively consistent performance across real and noisy augmented data, but their noisy condition accuracy is generally lower than the best audio-RF GMU Fusion. The MobileNet + MobileNet combination, achieving 62.42% accuracy in noisy conditions, is representative of the typical performance level for visual-RF GMU Fusion. While more robust than audio-visual, visual-RF does not reach the superior noisy condition performance of audio-RF.

C.2 Tri-Modal Fusion

Table 17 presents a detailed breakdown of the performance metrics for all evaluated tri-modality fusion models under both real and noisy augmented data. This table allows for a comprehensive comparison across different model combinations and fusion techniques when integrating audio, visual, and RF data.

- ① **Late Fusion.** In real data, all tri-modality combinations achieve near-perfect accuracy, with several combinations reaching a perfect 100% accuracy. This indicates that Late Fusion is highly effective in leveraging the synergistic strengths of audio, visual, and RF modalities when data quality is optimal. In noisy conditions, performance variation is observed, but overall, Late Fusion demonstrates significant robustness.

- **Audio-Visual-RF.** The VGG-19 + MobileNet + ResNet-10 achieves the highest noisy condition accuracy at 83.26%. This combination

Table 17: Performance metrics for three-modal fusions under real and noisy augmented data.

Fusion Type	Modality Combination	Models	Real Data					Noisy Augmented Data					Detec. Time (ms)
			Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	Acc (%)	Prec (%)	Rec (%)	F1 (%)	F1-M (%)	
Late	Audio-Visual-RF	LeNet + ResNet-10 + ResNet-10	100.00	100.00	100.00	100.00	100.00	67.95	66.67	99.29	79.77	51.36	5.43
		LeNet + ResNet-10 + MobileNet	100.00	100.00	100.00	100.00	100.00	63.18	63.49	99.17	77.42	38.91	7.89
		LeNet + MobileNet + ResNet-10	99.62	100.00	99.40	99.70	99.59	80.98	95.24	73.81	83.17	80.66	5.57
		LeNet + MobileNet + MobileNet	100.00	100.00	100.00	100.00	100.00	65.00	64.65	99.29	78.31	43.86	3.70
		VGG-19 + ResNet-10 + ResNet-10	96.97	95.45	100.00	97.67	96.66	67.27	66.04	100.00	79.55	48.86	6.85
		VGG-19 + ResNet-10 + MobileNet	95.61	93.54	100.00	96.66	95.12	57.35	61.67	87.14	72.22	40.19	8.85
		VGG-19 + MobileNet + ResNet-10	96.89	95.35	100.00	97.62	96.58	83.26	97.69	75.48	85.16	82.98	6.09
GMU	Audio-Visual-RF	VGG-19 + MobileNet + MobileNet	100.00	100.00	100.00	100.00	100.00	63.64	63.64	100.00	77.78	38.89	4.64
		LeNet + ResNet-10 + ResNet-10	96.97	95.45	100.00	97.67	96.66	75.53	72.23	100.00	83.87	66.58	6.48
		LeNet + ResNet-10 + MobileNet	96.97	95.45	100.00	97.67	96.66	74.09	71.07	100.00	83.09	63.87	4.81
		LeNet + MobileNet + ResNet-10	97.20	99.51	96.07	97.76	97.01	68.48	69.78	89.05	78.24	60.55	4.95
		LeNet + MobileNet + MobileNet	99.39	99.06	100.00	99.53	99.34	63.64	63.64	100.00	77.78	38.89	3.80
		VGG-19 + ResNet-10 + ResNet-10	97.05	95.56	100.00	97.73	96.75	63.64	63.64	100.00	77.78	38.89	7.55
		VGG-19 + ResNet-10 + MobileNet	97.27	100.00	95.71	97.81	97.10	53.26	63.58	62.14	62.85	49.92	7.38
		VGG-19 + MobileNet + ResNet-10	96.97	95.45	100.00	97.67	96.66	36.67	100.00	0.48	0.01	27.20	5.67
		VGG-19 + MobileNet + MobileNet	96.97	95.45	100.00	97.67	96.66	68.48	81.45	65.36	72.52	67.79	4.38

leverages the robust feature extraction capabilities of VGG-19 for audio, MobileNet for video, and ResNet-10 for RF, resulting in superior performance even when high noise is present. LeNet + MobileNet + ResNet-10 also performs strongly in noisy conditions with 80.98% accuracy. Both combinations remain significantly above the best dual-modality noisy condition accuracy (80.53% for audio-RF GMU Fusion), highlighting the added robustness of tri-modality Late Fusion. Notably, combinations including MobileNet sometimes tend to have lower detection times due to MobileNet’s efficiency, with LeNet + MobileNet + MobileNet achieving the fastest detection time at 3.70 ms while still maintaining 65.00% accuracy in noisy conditions.

② **GMU Fusion.** In real data, tri-modality GMU Fusion combinations also exhibit high accuracy, comparable to Late Fusion. However, in noisy conditions, GMU Fusion shows a different performance profile.

• **Audio-Visual-RF.** Among GMU Fusion combinations, LeNet + ResNet-10 + ResNet-10 achieves the highest noisy condition accuracy at 75.53%. LeNet + ResNet-10 + MobileNet also performs well at 74.09% accuracy in noisy conditions. While these accuracies are lower than the best Late Fusion tri-modality combination (83.26%), they also remain below the best dual-modality performance under noisy conditions (80.53%). This indicates that GMU Fusion, when applied to tri-modal data, is a weaker alternative to Late Fusion, exhibiting reduced peak accuracy in noisy environments.