Banshee: Target Switch Attacks on Gimbal-Stabilized Visual Tracking Systems via Acoustic Injection

Jiarui Li
University of Michigan
jiaruili@umich.edu

Joseph Brewington
University of Michigan
brewing@umich.edu

Qingzhao Zhang*
The University of Arizona
qzzhang@arizona.edu

Z. Morley Mao*
University of Michigan
zmao@umich.edu

***Abstract*—Gimbal-stabilized visual tracking is critical for modern autonomous systems such as Unmanned Aerial Vehicles (UAVs). While prior work shows acoustic signals can disturb gimbal internals, the impact of such attacks on real-world applications like UAV tracking and following remains underexplored. Existing demonstrations largely overlook practical challenges for real-world attacks, such as object-motion uncertainty and runtime latency. To bridge this gap, we present Banshee[1], the first physically realizable attack that induces target switching in UAV visual tracking systems by exploiting acoustic vulnerabilities in gimbal-camera systems. Banshee generates carefully crafted acoustic waveforms that induce optimized adversarial gimbal oscillations, causing directionally biased camera-view drifts that break inter-frame target associations. Consequently, the onboard tracker is driven to switch from the original target to an attacker-selected object with high probability, with occasional target loss. Banshee achieves a 93.6% success rate in simulation across two commercial gimbal systems and five trackers. Real-world benchtop and in-flight black-box attacks against a commercial drone across varied scenarios show an overall 95.5% attack success rate. Our results reveal a practical cross-domain vulnerability between acoustics and vision, highlighting the need for robust designs of gimbal systems and applications.**

1. Introduction

Gimbal-stabilized visual tracking is a core capability of modern camera systems, enabling persistent, high-precision object following in dynamic scenes. As a prominent example, commercial Unmanned Aerial Vehicles (UAVs) widely deploy target following, which typically pairs a multi-axis gimbal with an onboard camera, combined with object tracking algorithms running in software, to enable active tracking and following on a selected mobile target [1], [2], [3], [4], [5], [6]. Gimbal-stabilized visual tracking enables applications such as autonomous filming, surveillance, and infrastructure inspection, but also creates a single point of failure: compromising this pipeline can lead to severe consequences, including flight hazards, loss of vehicle control, and tracking of false targets [7], [8], [9], [10], [11].

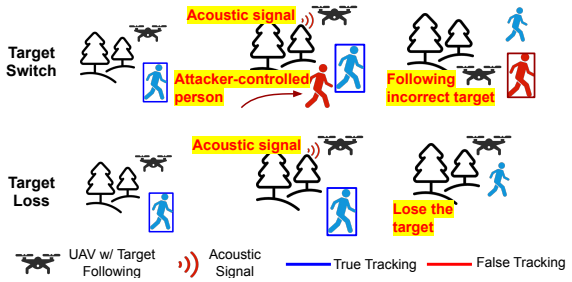

Figure 1: Illustration of Banshee in UAV target following. Crafted acoustic signals induce the UAV's visual tracker to switch to an incorrect target or lose track.

Gimbal systems commonly rely on real-time inertial measurement unit (IMU) feedback to mechanically stabilize onboard cameras during rapid motion [12], [13]. Prior research has shown that carefully crafted acoustic patterns, delivered via speakers, ultrasonic transducers, or even laser systems, can manipulate IMU readings and, in turn, disrupt gimbal stabilization [14], [15], [16], [17], [18]. While these studies establish the feasibility of influencing gimbal motion at the sensor level, they remain disconnected from real-world applications such as UAV target tracking and following, leaving their practical impact uncertain.

To address this gap, we propose Banshee, the first physical target switch attack against UAV visual tracking systems via acoustic injection, linking a hardware acoustic vulnerability with application-level tracking weaknesses to enable end-to-end system compromise. By injecting acoustic signals to the UAV, the attack induces gimbal oscillations with directional bias that accumulates drift along a vulnerable axis. This abnormal motion disrupts motion smoothness and corrupts the tracker's association with the true target, increasing the likelihood of a incorrect tracking output. Figure 1 illustrates two scenarios. (1) In the *target switch* attack, a UAV performing target following switches from its original target to an attacker-selected object, enabling the attacker to potentially steal the UAV from its owner or transfer tracking onto an unintended target. (2) In the *target loss* attack, the UAV loses its tracking target, allowing a

*. These authors are corresponding authors.

1. Our attack is named after the banshee, a mythical spirit whose scream causes or signals harm, reflecting the attack's acoustic nature.

suspect under surveillance to escape.

The attack has two stages. In *offline gimbal profiling*, the attacker uses an identical gimbal to learn a black-box mapping from acoustic signals to induced gimbal motion. In the *online attack*, the attacker runs two loops simultaneously. A *surrogate tracking* loop that runs a surrogate of the UAV onboard tracking mimicking the actual UAV tracking behavior using black-box knowledge. A *planning-execution* loop then optimizes a sequence of acoustic signals under physical and algorithmic constraints, leveraging both the gimbal acoustic response model and the tracking surrogate. These signals are then injected through a speaker or piezoelectric transducers to induce the desired gimbal motion, probabilistically biasing the tracker away from the true target toward incorrect associations.

Designing this application-aware acoustic attack raises several key challenges. First, the attack must achieve an empirically sufficient alignment between physical acoustic signals and their induced camera motion, in order to produce desired adversarial motion that disrupts tracking. Second, the attack must operate at runtime without prior knowledge of the UAV behavior, which motivates an adaptive online strategy that updates the surrogate tracker and signal injection plan as the scene evolves. Third, the attack must remain effective under real-world uncertainties, including unknown future object motion, which we address with optimization algorithms that tolerate uncertainties. `Banshee` overcomes these challenges and achieves high-probability empirical success in inducing target switch or target loss.

Extensive experiments prove the practicality of the proposed attack. First, offline profiling on built-in gimbals of two commercial UAV models shows that consistent runtime directional bias in gimbal motion is feasible. Second, we run large-scale simulation in Gazebo simulator, which deploys PX4-Autopilot flight stack, uses the profiled gimbal parameters, and tests the attack on five representative trackers and diverse scenarios. The results show that `Banshee` overall corrupts the tracking in 93.6% of trials (including 75.0% target switch and 18.6% target loss), proving attack effectiveness and robustness. Finally, real-world experiments further validate successful black-box target switch on commercial drones, including a realistic exploit on the built-in object tracking during flight.[2] We summarize our contributions:

- We design `Banshee`, the first end-to-end attack that uses acoustic injection to compromise UAV visual tracking, bridging gimbal-level vulnerabilities to application-level impacts, and adapts to diverse real-world conditions.
- We propose the first systematic method to empirically model gimbal motion under acoustic injection. Using offline profiling and online phase modulation, the attacker can induce consistent, directionally biased gimbal angular offsets in real time.

2. We have completed responsible disclosure to the vendor and, at their request, anonymized the commercial product models. This anonymization does not affect reproducibility; the vulnerability is not specific to the vendor's product and may apply broadly to gimbal mechanisms, and we provide proof-of-concept simulation environments at GitHub.

Acoustic Attack	**Consistent Biasing**	**Affected Module**	**Signal Source**	**Evaluation Platform**
Rocking. [16]	○	Firmware	Speaker	Drone controller
WALNUT [14]	●	Sensor	Speaker	Toy car
Injected [15]	●	Firmware	Speaker	Smart phone
Poltergeist [19]	○	Detection	Speaker	Smart phone
KITE [17]	●	Localization	Speaker&Piezo	Drone controller
Laser [18]	○	Detection	Laser	Drone gimbal
Ours	●	Tracking	Speaker&Piezo	Drone gimbal
Tracking Attack	**Online Optimize**	**Real-world Robustness**	**Attack Vector**	**Impact/Algo.**
Fooling [20]	○	○	Patch	Loss/MOT
AttrackZone [21]	●	○	Projector	Loss/SOT
ControlLoc [22]	○	◒‡	Patch	Loss/MOT
AdvTraj [23]	●	○	Person	Switch/MOT
FlyTrap [24]	○	◒‡	Patch	Shrink/SOT
Ours	●	●	Acoustic	Switch,Loss/SOT,MOT

‡ Robust patch generation by applying image transformations.

TABLE 1: Comparison with prior acoustic/tracking attacks.

- We extensively evaluate `Banshee` in high-fidelity simulation (Gazebo + PX4-Autopilot), with realistic simulation of gimbal vulnerability. We also demonstrate real-world online attacks on a commercial HighEndDrone.

2. Related Works

UAV visual tracking systems. Visual tracking enables core UAV functions such as target following, obstacle avoidance, and aerial monitoring [3], [4], [25], [26], [27]. Tracking algorithms generally fall into single-object tracking (SOT) and multi-object tracking (MOT). SOT methods, widely used for target following, include correlation-filter-based [28], [29] and siamese-network-based approaches [30], [31], both relying on appearance similarity. More advanced methods (e.g., transformer-based [32] and online learning-based [33], [34]) are often too resource-intensive for onboard deployment. MOT methods, used in aerial surveillance and obstacle avoidance [4], [27], [35], maintain associations across multiple dynamic objects and largely rely on motion consistency. Across these systems, a shared assumption is smooth inter-frame object displacement under benign conditions—an assumption our attack exploits.

Acoustic vulnerabilities of gimbal systems. Gimbal-stabilized cameras are standard in UAVs for stable video capture [4], [5], [6], relying on IMU-based feedback control [12], [13]. Prior work (Table 1) has shown that MEMS IMUs are vulnerable to acoustic injection [14], [15], [16], [17], [18], [36]. By exploiting resonant frequencies, attackers can inject false IMU signals, causing the gimbal to compensate for nonexistent motion and shift camera orientation. While these attacks demonstrate destabilization or sensor spoofing, they have not been systematically leveraged to induce target switch on visual tracking systems.

Prior physical attacks on visual tracking. Physical attacks on visual tracking have been explored but face unique challenges in UAV settings (Table 1). Adversarial patch–based attacks [20], [22], [37] can degrade tracking, yet lack online adaptation [38], [39] Real-time attacks using projectors or adversarial trajectories [21], [23] struggle with real-world uncertainty and latency. Acoustic attacks on object detection [19], [40] do not extend to the more complex tracking pipeline. FlyTrap [24] disrupts tracking via physical patches

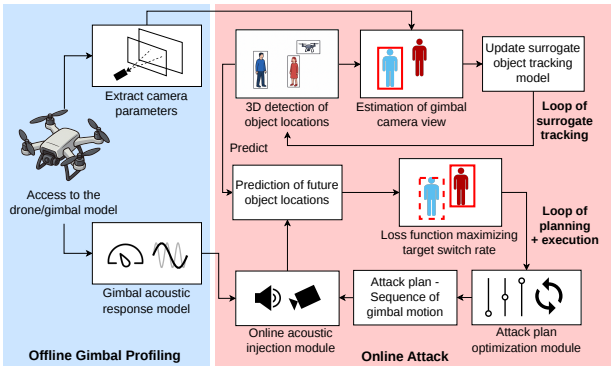

Figure 2: Overview of Banshee attack.

but under a different threat model. To date, no prior work demonstrates a physically realizable acoustic target switch attack on UAV visual tracking.

3. Threat Model

System and environment assumptions. In the attack scenario, a UAV performs target following using a gimbal-stabilized camera, whose video stream is processed by an onboard visual tracker. The flight controller then commands the UAV to follow the tracked target. Besides the legitimate target, the scene contains an attacker-controlled object that can be placed near the target during the attack. For example, if the UAV is following a pedestrian, an attacker-guided pedestrian can casually walk nearby (e.g., within a few meters), creating a normal and inconspicuous scenario. We consider a setting with no other nearby objects, as commonly seen in open-space tracking, and leave attacks in crowded scenes to future work. We denote the originally tracked object as **the true target** and the object after switching as **the false target**.

Attack goal. The adversary aims to launch an online **target switch** attack that achieves persistent target switching from the true target to the false target. The attack succeeds if the tracker consistently associates the false target, within the same semantic category (e.g., pedestrian or vehicle), as the true target. In different application contexts, the attacker can achieve two representative outcomes: (1) *Takeover*: the UAV, following its legitimate target (e.g., its owner), switches to an attacker-controlled object, resulting in effective capture; (2) *Escape*: the UAV, surveilling an attacker-controlled target, switches to another object, allowing the attacker to evade tracking. Both are instances of target switch, and we therefore do not distinguish them in evaluation.

A secondary goal is the **target loss**, where the tracker loses the target entirely, resulting in denial of service. While target loss often arises when target switch fails, it still has significant adversarial impact.

Due to real-world uncertainties and limited attacker knowledge (discussed below), the attack is inherently probabilistic rather than deterministic; nevertheless, it should achieve a high probability of tracking corruption.

Attacker knowledge and capability. First, we assume the attacker knows the target gimbal-camera model and can obtain an identical device, which is feasible for commercially available UAVs. Using this device, the attacker performs offline gimbal profiling (Section 4.1) to characterize its acoustic vulnerability and learn an empirical mapping from acoustic signals to induced gimbal motion, which supports the online attack (Section 4.2). This mapping guides not only signal optimization but also attacker object placement, e.g., aligning with the most vulnerable gimbal axis and maintaining a moderate distance from the true target. The attacker also obtains key system parameters, including gimbal constraints (e.g., rotation limits and degrees of freedom) and the camera intrinsic matrix.

Second, we assume the attacker has only black-box knowledge of the tracking algorithm deployed on the UAV system. The attacker does not know any implementation details including the algorithm and parameters, but the attacker can exploit surrogate models to mount transfer attacks.

Third, the attacker needs a mechanism to deliver acoustic impact during flight. The most reliable option is attaching a **malicious payload** (such as a piezoelectric transducer [17]) to ensure stable acoustic signal injection. Installing a malicious payload requires physical access, for example, an attacker posing as a maintenance technician could covertly attach the device during routine UAV servicing.

Finally, the attacker can observe the UAV system, the true target, and the false target throughout the attack, allowing a sensor system (such as LiDAR, radar, or camera) to perform **real-time 3D object detection** on the UAV system and key objects to be tracked, obtaining location information necessary for runtime attack optimization. The attacker also needs to know the **real-time gimbal orientation** to inject biased gimbal motion (Section 4.1) and achieve an accurate online surrogate of UAV tracking (Section 4.2.3). To this end, the attacker can install an IMU on the gimbal (as part of the malicious payload). Note that we do not assume any access to the UAV system's camera stream or internal software data. Our attack predicts the behavior of the tracking algorithm relying only on the external sensing and the profiled gimbal system parameters.

Relaxed requirements. Beyond the **primary** threat model, we consider a **relaxed** setting that reduces attacker requirements and validate the effectiveness of this threat model through physical experiments (Section 5.3). To avoid physical access, the attacker may perform a **remote attack** via long-range ultrasonic emitters [19], [41] or laser-induced vibrations [18] (see Section 6). The attacker may also estimate gimbal orientation using vision-based methods [42], [43] instead of installing a malicious IMU.

4. Banshee Design

Figure 2 shows the two stages of Banshee: offline gimbal profiling and online attack.

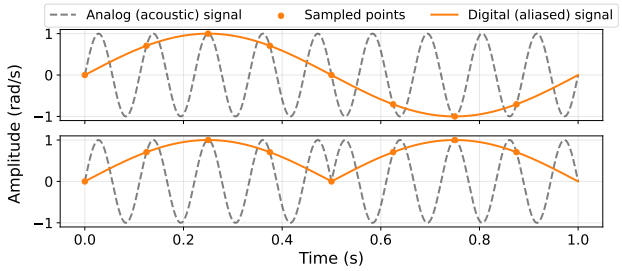

Figure 3: Illustration of phase-based directional bias in gimbal motion. Injected and aliased signals with (top) and without (bottom) phase shifting.

Offline gimbal profiling requires access to a gimbal identical to the target. Through controlled acoustic injection experiments, the attacker infers camera parameters, identifies the most vulnerable gimbal axis, and learns a black-box mapping from acoustic signals to induced motion (Gimbal Acoustic Response Model). This stage guides the selection of attack signals that induce abnormal gimbal motion. Details are in Section 4.1.

Online attack injects crafted acoustic signals into the target gimbal to induce abnormal motion along a vulnerable axis and cause tracking errors such as target switch. It consists of two iterative loops: *surrogate tracking* and *planning-execution*. The surrogate tracking loop maintains a surrogate tracker that mimics the UAV tracker, updated using the attacker's external sensor data and the estimated gimbal view. The planning-execution loop optimizes a sequence of acoustic signals under physical and algorithmic constraints, using the offline profiling results, the surrogate tracker, and the current 3D scene as inputs. These signals are then transmitted through the attack vector (for example, speakers and piezoelectric transducers) to gradually redirect the tracking from the legitimate target to the attacker-selected object. Details are in Section 4.2.

4.1. Adversarial Acoustic Gimbal Biasing

We design a physical attack that injects selected acoustic waveforms *at runtime* to induce biased, abnormal gimbal motion. Specifically, the acoustic signal excites resonance and corrupts gyroscope readings, causing the stabilization controller to apply incorrect compensation. This results in *directionally biased oscillations* that accumulate into abnormal gimbal motion.

Formally, by offline measurements on the the gimbal device, we empirically approximate a mapping from acoustic waveform with frequency f_{in} and amplitude A_{in} to the gimbal motion response as a three-axis angular velocity $\boldsymbol{\omega}[t]$. Let the acoustic drive be $x(t) = A_{\text{in}} \sin(2\pi f_{\text{in}} t)$. We learn a **Gimbal Acoustic Response Model** $\hat{\mathcal{M}}$ satisfying

$$\boldsymbol{\omega}[t] \approx \hat{\mathcal{M}}\big(A_{\text{in}},\ f_{\text{in}},\ t\big) + \boldsymbol{\varepsilon}[t], \tag{1}$$

where $\boldsymbol{\varepsilon}[t]$ is the residual error. This section presents the theoretical background and the offline profiling procedure used to obtain $\hat{\mathcal{M}}$, detailed in Sections 4.1.2-4.1.3. In addition, runtime directional biasing must handle uncertainty in the gyroscope sampling phase, which affects the induced motion direction; we therefore apply a phase-modulation routine that estimates and aligns the sampling phase at runtime, as detailed in Section 4.1.4.

4.1.1. Gimbal Fundamentals: Analog-Digital-Control. MEMS gyroscopes measure angular velocity via the Coriolis force on a vibrating proof mass. Suspended by springs, the mass resonates at its natural frequency f_n. In previous work, it has been established that an external acoustic signal with frequency $f \approx f_n$ can drive large oscillations even without actual motion [14], [15], [16]. We use the definitions standard among works on acoustic injection attacks and model the response of the proof mass as a driven harmonic oscillator:

$$\omega(t) = A \sin(2\pi f t + \phi), \tag{2}$$

where A is the vibration amplitude and ϕ is the phase of the signal.

These oscillations are then sampled by an analog-to-digital converter (ADC). Since the sampling rate f_s is much lower than the resonant frequency, the digitized signal appears as a lower-frequency alias rather than at f (Figure 3). For an ideal, constant-rate ADC, the sampled signal is

$$\omega[t] = A \sin(2\pi f_d \frac{t}{f_s} + \phi),\ \{t \in \mathbb{N}\}. \tag{3}$$

The digitized frequency of the aliased signal is subject to the Nyquist theorem

$$f_d = f - n \cdot f_s\ \ \{n \in \mathbb{N},\ \ f_d \leq \frac{1}{2} f_s\}. \tag{4}$$

The aliased frequency f_d is always less than half of f_s and approaches zero as the frequency of the injected signal nears an integer multiple of this rate.

The digitized acoustic signal propagates into the system as perceived motion, leading to compensatory control algorithm behaviors [44], [45]. Because commercial IMUs operate at modest sampling frequencies typically in the tens to hundreds of hertz and rarely exceeding 1 kHz [46], [47], [48], [49], the resulting aliased oscillations appear at low, easily observable frequencies, enabling our proposed profiling and modeling methodology described below.

4.1.2. The Model of Adversarial Gimbal Biasing. In theory, an adversary can influence and bias gimbal motion through modulated acoustic signals. To approximate this goal, we design a simplified gimbal motion abstraction leveraging structural and algorithmic insights as follows.

Scaling vibration amplitude. Amplitude scaling leverages the linear relationship between the injected signal amplitude and the resulting vibration amplitude. Following the amplitude equation for a driven harmonic oscillator we have

$$A = \frac{1}{m\ Z_m\ 2\pi f} \cdot A_{in}, \tag{5}$$

where the specific IMU model determines mass m and mechanical impedance Z_m, while the attacker controls signal frequency f and the acoustic amplitude A_{in}. We capture this linear relationship by fitting a linear regression for each of the identified resonant frequencies, simplifying the denominator to a single constant a

$$\omega[t] = (a \cdot A_{in}) \sin(2\pi f_d \frac{t}{f_s} + \phi). \tag{6}$$

Correcting direction of gimbal motion. The direction of gimbal motion is subject to real-world uncertainty since the phase of $\omega[t]$ depends on the relative timing between the injected signal and the sampling of the gimbal gyroscope. We extend the phase modulation technique in [14] to address limitations when the timing of samples is subjected to drift. We modulate phase in response to the direction of detected gimbal motion to improve robustness in the presence of sampling rate drifts. Specifically, when the direction of the observed gimbal motion begins to differ from the intended motion, we shift the phase of the injected signal by π so as to invert ω, also shown in Figure 3

$$-\omega[t] = A \sin(2\pi f_d \frac{t}{f_s} + \phi + \pi) \tag{7}$$

Adversarial biasing is independent of the gimbal orientation at the time of signal injection. UAV gimbal systems exhibit significant variability in servo actuation as a response to injected acoustic signals depending on gimbal orientation. The cause of this phenomenon lies in the gimbal stabilization mechanism. A 3-axis gimbal stabilizes a mounted camera using servos fixed in the drone body reference frame, while the camera body houses a gyroscope that is used in a feedback control loop to maintain stability in the world reference frame. To achieve this, gyroscope readings must be rotated by the orientation of the camera body relative to the drone body in order to actuate the correct servos. The closed-form matrix derived in [50] describes the rotation from the drone reference frame to the camera body reference frame that is used to transform gyroscope readings into correct servo actuation. Given angular values of pitch, roll, and yaw θ, the rotation matrix is

$$R = R_p R_r R_y = \begin{bmatrix} C_y C_p - S_y S_r S_p & -C_r S_y & C_y S_p + C_p S_y S_r \\ C_p S_y + C_y S_r S_p & C_y C_r & S_y S_p - C_y C_p S_r \\ -C_r S_p & S_r & C_r C_p \end{bmatrix} \tag{8}$$

where $C_i = \cos\theta_i$, $S_i = \sin\theta_i$, and current gimbal orientation θ. In short, when motion is perceived in the gimbal gyroscope the stabilization system will always actuate the correct servos in order to produce an opposite motion along the same axis in the camera-body frame, resulting in a stable motion response independent of the gimbal orientation. Section 5.1 also shows supporting results.

Overall, our Gimbal Acoustic Response Model $\hat{\mathcal{M}}$ has the following form

$$\omega[t] = a \cdot A_{\text{in}} \begin{cases} \sin(\theta_t), & \text{if } \operatorname{sgn}(\sin\theta_t) = s, \\ \sin(\theta_t + \pi), & \text{if } \operatorname{sgn}(\sin\theta_t) \neq s, \end{cases} \tag{9}$$

where $\theta_t = 2\pi f_d \frac{t}{f_s} + \phi$ and intended direction $s \in \{-1, 1\}$. In the following section, we present the profiling steps to obtain f_d and a. The additional residual error term $\varepsilon[t]$ is modeled as Gaussian noise, parameterized using the real-world experimental traces.

4.1.3. Gimbal Profiling for Obtaining the Model. The profiling procedure consists of two major steps that sweep over a single selected parameter of interest – **frequency** and **amplitude**, which produces f_d and a, respectively. In each sweep we first collect raw profiling data, then proceed to extract key information using spectral analysis and construct the necessary model.

Black-box gyroscope modeling via observation. Without loss of generality, we consider a three-axis gimbal camera system with an embedded three-axis gyroscope. By applying acoustic signals, the attacker induces oscillating motion readings in each axis, $\omega[t] = [\omega_p[t], \omega_r[t], \omega_y[t]]^T$, resulting in observable gimbal motion. The attacker installs an external malicious gyroscope on the camera body and collects the external gimbal motions $\omega_{obs}[t]$.

Frequency sweep. The first profiling step returns the set of resonant frequencies Q of the gyroscope that are most effective in driving proof-mass oscillations. (1) Following [15], sweep a sine wave across a wide range of acoustic signal frequencies (e.g., 1 Hz to 40 kHz) and record the motion response of the gimbal. (2) The acoustic amplitude A_{in} should be set to a fixed maximum value. For each injected test frequency f, extract aliased frequency f_d and amplitude A using spectral analysis. (3) Examining amplitude as a function of injected frequency and identifying local maxima of $A(f)$; these peaks reveal ideal injected signal frequencies for attacks and we collect them into a set Q.

Amplitude sweep. The second profiling step returns the linear relationships $\mathcal{A}$ between the injected acoustic amplitude and the amplitude of the induced gimbal motion. (1) For a selected frequency in Q, sweep the injected acoustic signal amplitude A_{in} from low to high (e.g., 1% to 100% signal power) and record the motion response of the gimbal. (2) For each injected signal, extract the observed motion amplitude A using spectral analysis. (3) Fit a linear regression of A as a function of A_{in} to obtain a constant a such that $A \approx a \cdot A_{in}$. (4) Repeat the procedure for remaining frequencies in Q.

4.1.4. Directional Biasing During Online Attacks. Since the phase of the sampled signal cannot be determined during offline profiling, we bias the direction of gimbal motion with real-time feedback using the theoretical model in Section 4.1.2. For enhanced robustness against real-world uncertainties, we can use a history-based method and only trigger a switch when the past N samples show motion in the incorrect direction. We also incorporate a time restriction, limiting the number of phase switches to a period of once per elapsed time T. The attacker can tune N and T to allow the proper level of feedback sensitivity. As a general rule of thumb, $N < f_{\text{feedback}}$ and $T < \frac{1}{f_d}$, where f_{feedback} is the frequency of the motion direction detection.

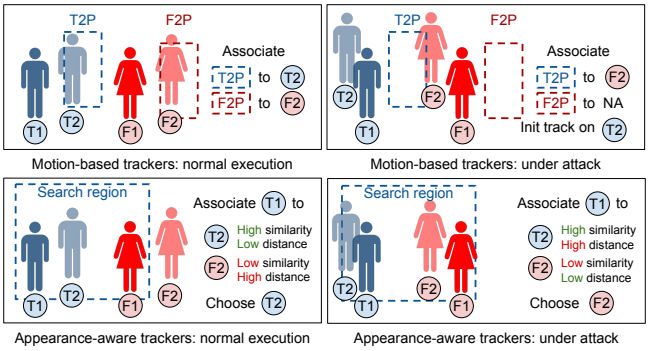

Figure 4: Illustration of how visual tracking is compromised by malicious gimbal motion. The subfigures show the gimbal camera view; T/F stand for the true target and the false target; numbers 1/2 are frame IDs; P is the predicted object location in motion-based trackers. The attack triggers a yaw-axis gimbal rotation and all objects are shifted to the left.

In summary, the attacker can inject consistent and directionally biased gimbal angular offsets in real-time. The accuracy of each component and more detailed settings of our profiling approach are available in Section 5.1. The online attack (Section 4.2) takes advantage of this constructed gimbal biasing pipeline to maximize the attack success rate.

4.2. Online Attack

As summarized above and illustrated in Figure 2, the online attack runs in two iterative loops: *surrogate tracking* and *planning–execution*. In this section, we first explain the underlying visual tracking vulnerabilities (Section 4.2.1) and key attack parameter choices (Section 4.2.2), then detail the surrogate tracking loop and the planning–execution loop (Sections 4.2.3–4.2.4).

4.2.1. Vulnerability of Visual Object Tracking. Visual object tracking processes a 2D image stream to localize a designated object across frames. While tolerant of occasional detection errors, trackers become unstable when gimbal motion is perturbed. Below, we explain the reasons, also illustrated in Figure 4.

Motion-based trackers. Such trackers (e.g., SORT [51]) rely on spatio-temporal consistency, such as bounding-box overlap and motion-model prediction, to associate objects across frames. These algorithms assume that target displacement between consecutive frames is small and predictable: they predict the object's location using a motion model and then match the predicted box to actual detections based on distance or overlap. For instance, SORT uses a Kalman filter for motion prediction and Hungarian matching with IoU cost for object association. Acoustic injection disrupts this assumption by inducing abrupt shifts in the gimbal's field of view, creating large apparent jumps in target position. As a result, the tracker may fail to maintain association with the legitimate target and instead link the track to a nearby object. In Figure 4, the predicted location of the true target (T2P) under attack falls closer to the false target 's detection (F2), causing the tracker to switch its association.

Appearance-aware trackers. Appearance-aware trackers (e.g., Siamese trackers) employ deep similarity networks to match object patches across frames. However, they still depend on spatial proximity constraints to narrow the candidate search space. For instance, KCF [52] considers object candidates only within a search region predicted from the last occurrence of the object; SiamRPN [53] further penalizes candidates that are far from the search region center and whose scale deviates significantly from the previous frame. Abrupt displacements caused by the adversarial gimbal motion could render the true target falling outside or further from the search region, reducing the reliability of matching.

Therefore, in both categories, acoustic-induced gimbal perturbations could be optimized to undermine the fundamental assumptions of inter-frame correlation that tracking algorithms depend upon.

4.2.2. Online Attack Parameters. As introduced in the threat model (Section 3), the attack seeks to mislead the tracker from following **the true target** to **the false target**. The attacker should determine the following parameters.

Geometric relationship of the two objects. Depending on the attack goal – *takeover* or *escape*, the adversary can manipulate either the false target or the true target. By controlling the trajectory of one object, the attacker determines the relative positioning between the two, tailoring it to the vulnerabilities of the target gimbal system. Specifically, in the offline profiling, we access which axis of the target gimbal is the most vulnerable to acoustic injection. For example, if profiling reveals that the gimbal's yaw axis is most susceptible to acoustic injection, the attacker should arrange the true target and the false target horizontally within the camera's field of view. In this case, perturbations along the yaw axis can shift the tracker's focus from the true target to the false target. During the attack window, the two objects are typically placed in close proximity (e.g., 1–3 meters apart), as shorter separation reduces the required attack duration given the limited induced angular velocity.

Cycle times. The two in-loop processes should run as frequently as possible to enable fine-grained online optimization of the attack. However, several constraints govern the choice of cycle times. (1) Both loops must respect computational limits, which depend on the attacker's hardware. In our case, the attack algorithm is lightweight, and its efficiency is evaluated in Section 5.2.5. (2) Stable gimbal motion requires that the planning–execution loop operate with a sufficiently long cycle time to allow acoustic signals to take effect. In practice, we align the planning-execution cycle time to an integer multiple of the attack signal frequency (Section 4.1). (3) The two loops must remain synchronized as the planning-execution takes the 3D object detection and surrogate model status as input. In practice, because the planning-execution loop is usually slower than surrogate tracking, its cycle time is set to be

an integer multiple of the surrogate tracker's cycle time, ensuring coordination.

4.2.3. Loop of Surrogate Tracking. The attacker uses its own sensors to track objects in the scene, estimate the viewpoint of the target UAV gimbal camera, and run a black-box surrogate of the UAV's tracking algorithm that mimics its runtime behavior, preparing for later optimization.

Real-time gimbal view estimation. A key challenge of the black-box attack is that the video stream from the target gimbal is unavailable, limiting direct use of a surrogate tracker. To overcome this, `Banshee` estimates the gimbal's camera view in real time using external observations of the scene. Specifically, the attacker first applies 3D object detection (via an external sensor suite) to obtain the world coordinates of the true target, the false target, and the UAV. The gimbal's orientation is then derived either by (1) a malicious IMU attached to the gimbal (requiring physical access), or (2) a high-resolution external camera with orientation detection algorithms (also see Section 3). Additionally, camera parameters are already known from the offline profiling stage. With the above information, `Banshee` projects the 3D world coordinates of the true target and the false target into the estimated target UAV's 2D camera coordinate system. These projected 2D bounding boxes are then used as inputs to the surrogate tracking model, updated at each cycle of the loop.

Surrogate tracking models. Without access to the UAV system's image stream and internal data, we design surrogate tracking models to enable the optimization.

Motion-based trackers. Modern motion-based trackers typically operate in two steps: object detection followed by ID association. In our surrogate model, object detection is replaced by the real-time gimbal view estimation. As such, at each frame, instead of running an image-based detector (which the attacker cannot access), we estimate the 2D bounding boxes of the true target and the false target and treat them as detection results. These estimated boxes are then the input of remaining surrogate tracking steps, with details depending on the specific surrogate model implementation. If using SORT [51] as the surrogate model, for instance, the two boxes are then matched with Kalman Filter (KF) [54] predictions using the similarity metric of intersection over union (IoU), after which tracks are updated and KF states are refined. To enable efficient optimization, we formulate this pipeline in a differentiable manner and compatible with gradient descent.

Appearance-aware trackers. Appearance-aware trackers localize targets by searching within a region around the last known position, leveraging both appearance cues and spatial consistency. In our black-box setting, the surrogate model cannot replicate the deep learning component that processes raw images. Instead, it directly estimates the deep learning component's outputs and uses them as inputs to the subsequent object association stage. Taking a SiamRPN-based surrogate model [53] as an example, it crops a search region centered on the previous tracking box, typically four times its size. In the original tracker, this region is processed by a neural network to produce appearance-correlated proposals of boxes to be tracked. In our surrogate model, we mock this step by generating box proposals directly around the estimated positions of the true target and the false target in the gimbal camera view. Specifically, we sample multiple box proposals (e.g., 10) around each of the two objects, with slight random perturbations in their sizes and locations, and assign them high confidence scores (e.g., 1.0), reflecting the assumption that the component can reliably recognize foreground objects. These proposals are then passed to the association module, which selects the final tracked box. The entire process can also be implemented in a differentiable manner. Note that while the mocked proposals and confidence scores differ from those produced by the actual tracking with image feeds, they capture the key association behavior of appearance-aware trackers that they favor proposals closer to the search region center, which is sufficient for the optimization to exploit the tracker vulnerability.

Algorithm 1 Planning of angular velocity to inject.

Input: S_i: list of 3D bounding boxes for the true target, the false target, and UAV system at cycle i, θ_i: gimbal orientation , Surr: surrogate tracking algorithm , ΔP_i: position offset to be injected , t^c: planning-execution cycle time.
Output: ω_{i+1}: the planned angular velocity for cycle $i+1$.

function FINDOPTIMALANGULARVELOCITY
  for $iter = 0$ to N **do**
    $T_{i:i+2} \leftarrow$ GENERATETRAJECTORIES(S_i, S_{i-1}, t^c)
    $\Theta_{i:i+1} \leftarrow$ GENERATEROTATIONS(θ_i, ΔP_i, t^c)
    UPDATE(Surr, $T_{i:i+1}$, $\Theta_{i:i+1}$)
    for tr in $T_{i+1:i+2}$ **do**
      $X_{i+1} \leftarrow$ 3DTO2DPROJECT(tr, ω_{i+1})
      COMPUTELOSS(X_{i+1}, Surr)
    end for
    GRADIENTDESCENT(ω_{i+1})
    APPLYCONSTRAINT(ω_{i+1}, Θ_i)
  end for
  return ω_{i+1}
end function

4.2.4. Loop of Planning and Execution. The attacker starts an iterative process with planning and execution run simultaneously in parallel when the surrogate tracking model is initialized and the targets are in position. The iterative process stops after the attack success or timeout.

Planning. In the i-th planning-execution cycle, the algorithm computes the next attack plan ΔP_{i+1}, i.e., the gimbal orientation offset to be executed in cycle $i+1$ (Algorithm 1). The algorithm first loads information from the surrogate tracking loop, including 3D detections of the true target, the false target, the UAV, and the latest surrogate model.

To plan for the future, the algorithm predicts the states of objects, the gimbal, and the tracker. The attacker uses motion/trajectory prediction models to estimate the trajectories of the true target, the false target, and the UAV across cycles i and $i+1$. Future gimbal motion is predicted using the current attack plan ΔP_i and offline profiling results. To increase robustness against prediction errors, we apply

expectation-over-transformation (EoT): each predicted trajectory or motion is augmented with sampled noise, forming sets of trajectories for optimization. For each combination of sampled predicted states, the surrogate tracking model is executed on them and produces an estimated future tracker state, i.e., the location of the tracking box if the scene evolves according to the predicted trajectories. The optimization objective is then defined as the below:

$$\arg \min_{\omega < \omega^{max}} \mathbb{E}_{P^t, P^f, X, G} \left[\sum_{i=1}^{I} \text{dist}(Box_i^{track}, Box_i^{false}) \right], \tag{10}$$

where P^t, P^f, and X denote the trajectory sets of the true target, the false target, and the UAV (from EoT), G is the gimbal orientation distribution, I is the number of frames in a cycle, Box^{track} is the surrogate tracker output tracking box, and Box^{false} is the false target box. In short, the optimization searches for a gimbal angular velocity ω that maximizes the chance of misplacing the tracking box onto the false target by minimizing the distance between the two boxes. The solution is constrained by the maximum achievable velocity ω^{max} from profiling. Because the surrogate tracking model is differentiable (Section 4.2.3), the optimization can be efficiently solved with gradient descent.

The final step of planning is to translate the angular velocity into the acoustic signal to inject. The translation is straightforward by applying the gimbal acoustic response model (Section 4.1).

Execution. The execution step runs in parallel with the planning step. The optimized attack plans (i.e., the acoustic signals to inject) are transmitted to the signal generator immediately after completion. When the i-th planning-execution cycle starts, the attack plan ΔP_i is immediately executed as it is generated in the last cycle (the first cycle has no execution step). This parallel scheduling minimizes the effect of optimization latency on attack performance, with efficiency shown in Section 5.2.5.

5. Evaluation

We extensively evaluate `Banshee` to demonstrate its effectiveness and impact. First, we validate offline gimbal profiling, which underpins our runtime attack, by showing it consistently injects gimbal responses with directional bias (Section 5.1). Next, we evaluate attack effectiveness and robustness in high-fidelity simulation across diverse scenarios, including ablations (Section 5.2). Finally, we demonstrate practicality via physical experiments on a commercial drone, achieving results consistent with simulation (Section 5.3).

5.1. Gimbal Biasing Experiments

In this section, we evaluate the feasibility and consistency of offline gimbal profiling and biasing.

Tested drone and gimbal systems. We profile two state-of-the-art commercial UAV gimbal systems: a high-end industrial drone with a payload-mounted gimbal and a mid-end consumer drone with an integrated gimbal. For HighEndDrone, we perform full profiling and physical experiments, while for MidEndDrone, we conduct frequency sweeps and directly adjust amplitudes in simulation-based evaluations.

Experimental setup. We attach a piezoelectric transducer directly to the gimbal housing and drive it with a function generator (AD9033) and amplifier (TDA8932) to inject crafted acoustic signals into the gimbal through vibration. An IMU (BMI160) is also attached to the gimbal housing to obtain the gimbal orientation.

Frequency sweep. For both HighEndDrone and MidEndDrone, we sweep 5 kHz to 30 kHz. After applying Least-Squares Spectral Analysis and local maxima filtering, we obtain resonant frequency sets $\{7744, 23232\}$ and $\{5526, 9214\}$ for HighEndDrone and MidEndDrone respectively. Figure 5 shows a reconstructed signal using frequency, amplitude, and phase of oscillation in each axis obtained through spectral analysis.

Amplitude regression. Figure 6 shows one regression produced at an injected signal frequency of 23232 Hz. We use a digitally controlled potentiometer to control the signal amplitude in terms of percentage power. We profiled 3 values $a_p = 0.27, a_r = 0.27, a_y = 2.63$ for pitch, roll, and yaw, respectively. These three regressions achieve an average R^2 of 0.997, showcasing the accuracy of our model. Our measurements also show that, at this frequency, the motion response amplitude of the yaw axis is significantly more sensitive to acoustic signal strength than pitch and roll.

Gimbal orientation independence. To validate that acoustic gimbal biasing is independent of gimbal orientation, we test combinations of pitch at $-90°$ to $45°$, roll at $-15°$ to $15°$, and yaw at $-90°$ to $90°$, which covers the full range of possible orientations of HighEndDrone's gimbal. In Figure 9 we plot the average error across all three axes at each orientation, compared to the motion observed at the neutral orientation. Our experiments show that across all orientations, the movement perceived by our external IMU is nearly identical with only slight variation. These minor errors can be most likely attributed to factors such as measurement error and any mechanical differences such as wear in the servos which rotate the different axes.

Biasing results. To evaluate the consistency and accuracy of our online biasing methodology, we conduct 100 trials at a frequency of 23232 Hz over 100% and 50% signal power in clockwise and counterclockwise yaw rotation. In Figure 7 we plot all 100 traces integrated over a 1 second period. We successfully demonstrate that using our online biasing methodology, we can achieve consistent gimbal rotation in the desired direction with only a small degree of error.

Execution latency. A potential concern is latency between acoustic injection and gimbal response. In our experiments, latency was negligible, consistent with expectations: the proof mass responds instantaneously, and gimbals must react in real time to support image stabilization.

Time domain demonstration. To expose intermediate gimbal states and better understand acoustic injection, we simulate the gimbal stabilization pipeline in Gazebo [55]. Our

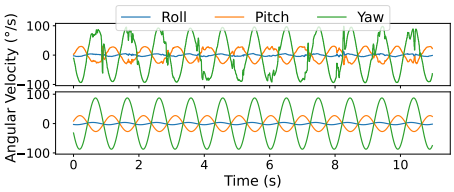

Figure 5: Accuracy of spectral analysis for 7744 Hz. Reconstructed signal (bottom) is close to raw gyroscope readings (top).

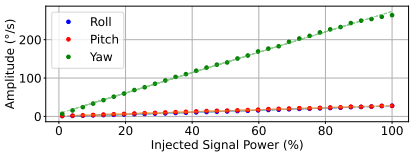

Figure 6: Accuracy of linear amplitude regression at 23232 Hz.

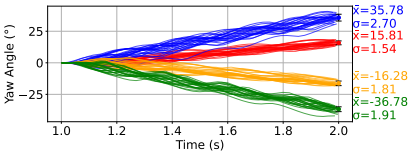

Figure 7: Online gimbal biasing: integrated angle over time for different amplitudes and directions.

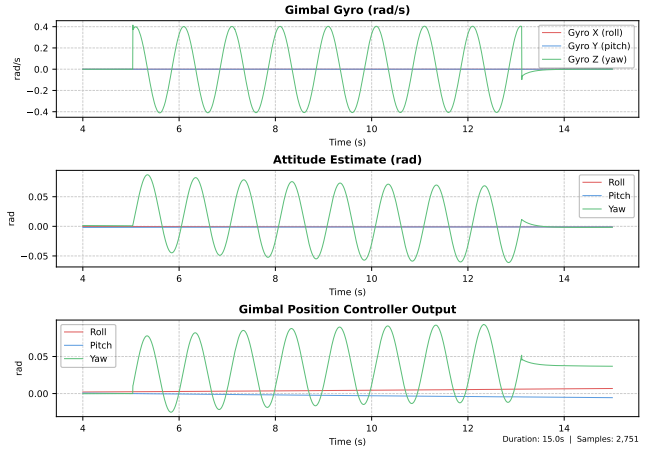

Figure 8: Time domain recording of intermediate gimbal states during acoustic injection.

Algorithm	Motion	Appear.	Robust Design	FPS
KCF [52]	○	●	-	172 / C [52]
SiamRPN [53]	○	●	-	49 / G [56]
DaSiamRPN [57]	○	●	Distractor aware	20.2 / G [58]
SORT [51]	●	○	-	44 / G [59]
UCMCTrack [60]	●	○	Motion compensate	44 / G [59]

TABLE 2: Details of evaluated tracking algorithms. Motion/appearance—tracker types; C/G—CPU or GPU.

	Normal	Noisy	HighEndDrone	MidEndDrone
SORT	11.1%/24%	0%/75%	59.3%/94.4%	55.6%/96.3%
UCMC.	0%/0%	6%/6%	68.5%/92.6%	61.1%/87.0%
Siam.	0%/0%	0%/0%	83.3%/88.9%	90.2%/93.1%
DaSiam.	0%/0%	0%/0%	92.6%/98.1%	97.2%/100%
KCF	0%/0%	17.6%/17.6%	70.3%/94.4%	72.2%/87.5%

TABLE 3: Switch/disable attack success rates in simulation.

implementation follows the STorM32-BGC design: IMU (gyroscope/accelerometer) measurements are fused via a complementary filter for attitude estimation, followed by a cascaded PID controller with an outer attitude loop and a high-gain inner rate loop. Acoustic perturbations are injected directly into gyroscope measurements (prior to Gazebo noise) as sinusoidal signals, with rectified variants for phase modulation along a selected axis. Figure 8 shows gyroscope readings, attitude estimates, and controller outputs under a 1 Hz yaw-axis injection. The resulting sinusoidal motion is consistent with our empirical model despite pipeline nonlinearity.

Overall, these results demonstrate both the feasibility of the profiling and the consistency of gimbal biasing through acoustic injection.

5.2. Simulation Experiments

Our simulation experiments comprehensively evaluate the attack's effectiveness, robustness, and ablation results.

5.2.1. Experiment Setup. We begin by describing how the simulation is configured to approximate real-world attacks.

Tracking algorithms. We select five representative visual tracking algorithms, as shown in Table 2. The algorithms span both major categories: motion-based trackers and appearance-aware trackers (details in Section 4.2.1). In each category of tracker, we include the lightweight method (e.g., KCF and SORT), deep-learning-based methods (e.g., SORT and SiamRPN), as well as the state-of-the-art variants with robustness enhancement (e.g., DaSiamRPN and UCMCTrack). For SORT and UCMCTrack which require an object detection component, we use the pretrained model YOLOv8x [59]. For learning-based methods SiamRPN and DaSiamRPN, we use pretrained models trained on a range of datasets: VID [61], YoutubeBB [62], COCO [63], and ImageNetDet [61]. Our evaluated algorithms are lightweight enough to run on a typical UAV onboard computer. Other algorithms such as transformer-based methods are not suitable for real-time tracking on a mobile system (i.e. STARK [32] runs at only a max 30-40 FPS even using a Tesla V100 GPU).

Evaluation metrics. We evaluate the following metrics:

- *Target switch* is successful if the tracking box contains the center of the false target for at least ten consecutive frames and persists to the end of the simulation.
- *Target loss* happens when the tracking box has no overlap with either the true target or the false target (often a side effect of failed target switch).
- *Disable*: The union of target switch and target loss attacks.

Simulation scenario setup. We simulate the attack with the high-fidelity physical simulator Gazebo [55] and the complete UAV software stack PX4-Autopilot [64]. The simulation experiments are carried out on a desktop with Intel

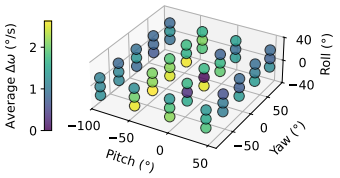

Figure 9: Orientation independence: difference from neutral orientation in amplitude, averaged across pitch/roll/yaw over various orientations.

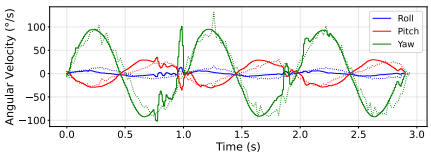

Figure 10: Acoustic gimbal motion injection trace in the physical world (solid line) and simulation (dotted line).

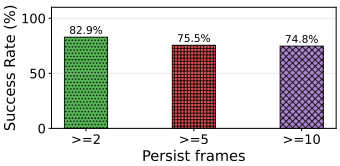

Figure 11: Distribution of number of consecutive frames of target switch, showing that target switch may become target loss in attacks.

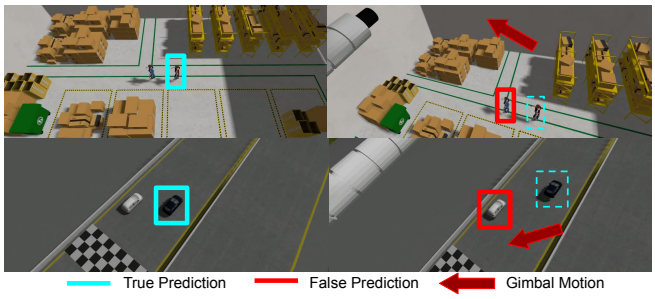

Figure 12: Demonstration of successful simulated target switch attacks.

i9-14900K CPU and RTX 4080 GPU running Ubuntu 24.04. Figure 12 demonstrates some of the attack scenes.

To demonstrate the generality of `Banshee`, the test cases are distributed as follows:

- *Worlds*: We include four distinct environments: factory, urban, field, and raceway. The factory world is an indoor setting, while the others are outdoor scenarios.
- *Object categories, motion, and appearances*: Both the true target and the false target can be either pedestrians or vehicles. Pedestrians move at speeds of 0.5-1.5 m/s in random directions and vehicles move at speeds of 5-15 mph straight. Both pedestrian models (blue, green, red outfits) and vehicle models (silver, blue, and red hatchbacks) have pairwise distinct appearances.

In total, we conduct 108 trials for each evaluation case. 54 are pedestrian cases, evenly distributed across three worlds (factory, urban, field) and six appearance pairs, with randomized initial motions. The remaining 54 trials involve vehicles, conducted in three worlds (raceway, urban, field), also covering six appearance pairs.

Simulation of the system and attacks. For realistic simulation, we follow the threat model (Section 3) and configure the simulation as follows:

- *Adversarial gimbal biasing*: We implement adversarial acoustic gimbal biasing in simulation using the gimbal acoustic response model (Section 4.1). We inject Gaussian noise parameterized by the residual error from physical profiling via spectrum analysis (Section 5.1), modeling realistic motion perturbations (Figure 10). Specifically, we configure the gimbal maximum angular speed for the roll, pitch, and yaw gimbal axes as (0.04, 0.27, 2.35) and (0.76, 2.92, 0.21) rad/s, with realistic motion injection noise of (0.03, 0.04, 0.34) and (0.17, 0.51, 0.05) for HighEndDrone and MidEndDrone respectively. Other simulated gimbal characteristics, including rotation limits, degrees of freedom, and frame rate, are aligned with HighEndDrone or MidEndDrone.
- *UAV altitudes*: UAVs are above the ground by 10 meters and 25 meters for the pedestrian and vehicle scenarios, respectively, which aligns with real-world drone setup.
- *Distance of two objects*: In the pedestrian scenario, the true target and the false target are placed 1–2 meters apart, while in the vehicle scenario they occupy adjacent lanes. Such distances are common in real-world settings and therefore do not appear suspicious as malicious behavior.
- *Uncertain object motion*: To reflect the real-world uncertainty, the true target and the false target move with injected Gaussian noise. Uncertainties in UAV flight are simulated with Gazebo's physical engine.
- *Motion blur*: Abrupt camera shifts naturally introduce motion blur. We model this effect using linear and rotational blur kernels in OpenCV [65]. However, since the gimbal's maximum angular speed is constrained, the blur remains minor, and our attack does not depend on it for success.
- *Baselines*: In Normal, we simulate UAV object tracking without the presence of attack. In Noisy, we simulate UAV object tracking under strong and gusty wind supported by the realistic Gazebo WindEffects plugin.
- The attack uses a surrogate model aligned with the evaluated tracker, with a 4 Hz planning–execution cycle, one gradient descent step per cycle, and three trajectory samples for expectation-over-transformation.

5.2.2. Attack Effectiveness. Table 3 presents the attack results in the simulation setup as described in Section 5.2.1. Our attack demonstrates overall success on different object tracking algorithms with **75.0%/93.6%** averaged success rate for target switch and disable (target switch & target loss) respectively. Target loss arises as a side effect of target switch. Although the attack is optimized for target switch, its probabilistic nature can still disrupt tracking and cause all objects to be lost. Figure 11 shows that 8.1% achieve target switch for at least two frames before transitioning to target loss eventually.

Time domain demonstration. Figure 13 shows a target switch attack example for SiamRPN and SORT in simu-

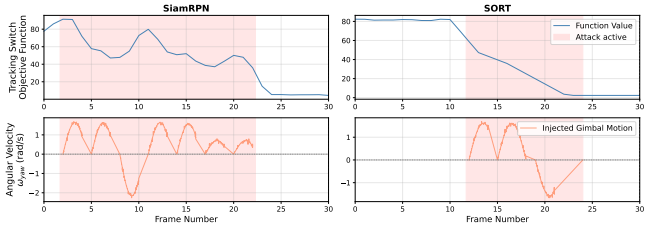

Figure 13: Time domain demonstration of injected gimbal motion induces target switch in simulation.

lation. The target switch objective function (Equation 10) captures key attack success criteria. Minimizing the function value (blue curve) maximizes attack success. Orange curve shows simulated gimbal motion under acoustic injection attack, configured using the acoustic parameters obtained from the physical profiling experiment. It shows that injecting the optimized gimbal motion minimizes the objective function leading to target switch attack success.

Baselines. Switching is negligible under both settings, confirming that target switching does not arise in benign or naturally perturbed conditions. While noise can degrade tracking (e.g., causing target loss for SORT), it does not induce sustained target switching.

Across tracking algorithms. Interestingly, more advanced tracking algorithms, i.e., with embedding of object appearance or compensation to camera instability, do not exhibit stronger resilience to our attacks. Instead, appearance-aware trackers (SiamRPN, DaSiamRPN) achieve up to 29% higher target switch success rates than motion-based trackers (SORT, UCMCTrack). This indicates that the attack perturbation is strong enough for corrupted spatio-temporal consistency to outweigh appearance features and the discriminative appearance awareness assists the attack by identifying the false object instead of losing in the background. The increasing target switch success rate among KCF, SiamRPN, and DaSiamRPN with deeper appearance awareness provides further evidence.

The advanced motion-based tracker UCMCTrack includes mechanisms to compensate for camera instability, yet it still fails under our target switch attack, with a target switch success rate higher than the vanilla SORT. The attack success DaSiamRPN and UCMCTrack highlights that existing robustness designs cannot withstand the malicious perturbations but only assists the attack by reducing target loss. Across three runs, the standard deviations of target switch success (SiamRPN: $\pm 9.1\%$, DaSiamRPN: $\pm 3.5\%$, SORT: $\pm 4.5\%$, UCMCTrack: $\pm 8.2\%$, KCF: $\pm 10.3\%$) show the consistency of this trend. See Finding 1.

Finding 1: Robustness oriented tracking designs, including appearance aware and motion compensated trackers, do not withstand our target switch attack; instead, they often increase target switch success rates by preserving corrupted tracks rather than allowing target loss.

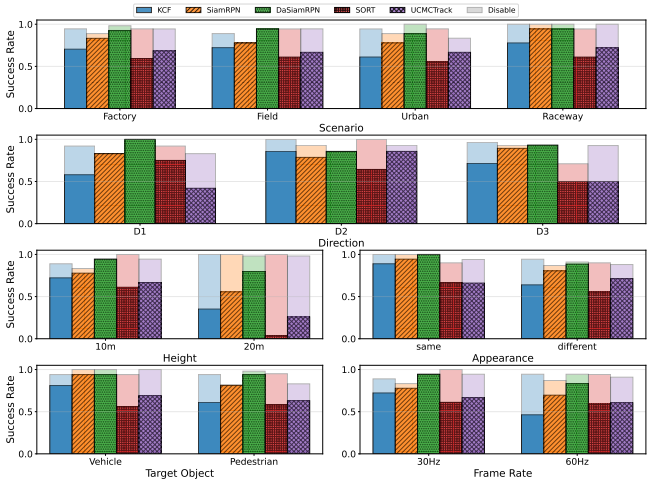

Figure 14: Simulation results analysis. Filled bars: target switch success rate. Translucent bars: disable success rate.

Across different gimbals.. The attack demonstrates stable success across two commercial gimbal camera systems with distinct acoustic vulnerability shown in Section 5.1. This suggests that the proposed online attack automatically adapts to specific gimbal parameters, regardless of the realistic random noise we added to the simulated gimbal motion.

5.2.3. Analysis of Impacting Factors. Beyond the default simulation setup, we vary several factors reflecting real-world conditions to evaluate the robustness of the attack. Results are shown in Figure 14.

Different environmental scenarios. Different environmental scenarios have only limited impact on target switch success rates, showing a minimum of 70.2% in the factory scenario and a maximum of 80.0% in the raceway scenario. This is likely because the raceway vehicles are easier to track than the factory pedestrians because of object sizes.

Different true target directions. This analysis focuses on the pedestrian scenario. We categorize the randomized pedestrian motion into 3 directions in the gimbal camera field of view: D1 (moving away), D2 (moving horizontally), and D3 (moving closer). According to our results, different motion directions have limited impact on target switch attack success, confirming that our optimization algorithm can handle variation in the the true target's motion.

Different UAV heights. The result shows attack success rates at varying UAV heights in the pedestrian scenario. The attack maintains high target switch success at low altitude (10 m), where richer visual cues and reliable detections benefit both category of trackers. However, target switch success drops markedly at higher altitudes (20 m) as smaller object size degrades tracking performance, leading to increased target loss. The vehicle scenario results further support this hypothesis with a 78.9% target switch success rate at 30 m.

Target scenarios. The attack consistently achieves comparable or higher success rates in vehicle scenarios across

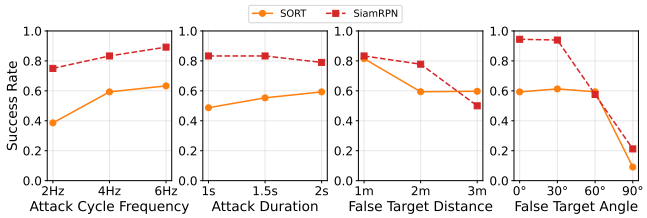

Figure 15: Target switch success rate w/ various conditions.

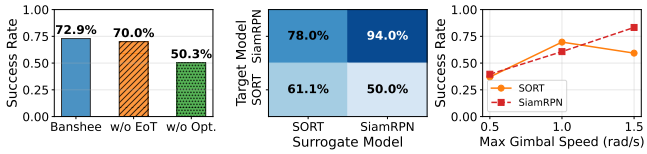

Figure 16: Left: attack components ablation. Middle: surrogate models transferability. Right: gimbal speed ablation.

all tracking algorithms. This indicates that targets which are inherently easier to track, such as vehicles with regular motion patterns and stronger visual features, are also more vulnerable to target switch.

Target appearance. Appearance-aware trackers show markedly higher target switch success when the two objects appear identical, as expected from their reliance on appearance matching. These results indicate that an attacker can enhance success by intentionally making objects visually similar.

Camera frame rate. The 30 Hz setting, used by default in simulation, reflects common real-world configurations (Table 2). Results show that 60 Hz frame rate mitigates the attack, as the same gimbal motion causes smaller inter-frame misalignment. Appearance-based trackers are most affected, suggesting that increasing frame rate can enhance robustness against malicious gimbal motion.

We conclude to Finding 2 summarizing the above:

Finding 2: The attack remains highly effective across real-world variations, and while factors such as higher UAV heights and higher tracking frame rates can reduce success, these challenging cases are limited.

5.2.4. Ablation Study. We perform a set of experiments to understand the effect of different attack parameters and design components. Experiments are performed against SORT and SiamRPN with the same configuration as in Section 5.2.1 with HighEndDrone acoustic parameters.

The usefulness of design components. We perform two ablation studies on the optimization component: removing expectation over transformation (EoT) and removing the entire optimization. Without EoT, optimization samples a single trajectory from the object motion models (4.2.4); without optimization, the attack applies a constant angular speed to shift the camera toward the false target at the maximum controllable rate. As shown in Figure 16, both components substantially improve target switch success.

Distance between two objects. In the pedestrian scenario, we evaluate attack performance as the distance between the true target and the false target varies (either may be the attacker controlled object; see Section 3 and Figure 4). For SiamRPN, target switch success declines sharply with increasing distance, as its fixed search region limits association, making appearance-based trackers harder to compromise when objects are far apart. In contrast, motion-based trackers assign new IDs once geometric consistency breaks, enabling target switch even at larger separations.

Angle between two objects. We vary the relative angle between the true target and the false target. At $0°$ and $90°$, the controlled object aligns with or is orthogonal to the gimbal axis most susceptible to acoustic injection (Section 4.2.2). Deviations from this axis markedly reduce attack effectiveness (Figure 15). The motion-based SORT is more sensitive to angle changes, as its association depends on whether the attacker lies along the injected motion direction, while the appearance-based SiamRPN is less affected due to its broader search region.

These results highlight that both distance and angle of the attacker-controlled object are critical. To maximize success, the attacker should align objects with the gimbal's most vulnerable axis in the camera view—achievable through scene understanding and viewpoint prediction. The distance can then be tuned to balance success and stealth: greater separation slightly lowers target switch probability but benefits inconspicuity. In summary:

Finding 3: Attack success strongly depends on how the attacker places the controlled object, with favorable distance and alignment along the most vulnerable gimbal axis significantly boosting target switch effectiveness.

Planning-execution cycle time. A shorter cycle time yields more frequent optimization updates, making the attack more responsive to subtle scene changes. As shown in Figure 15, shorter cycle time increases success rates, particularly for SORT. Ideally, the attacker would operate at the fastest cycle allowed by available computational power.

Attack duration. The impact of attack duration differs markedly between motion based and appearance aware trackers. For SiamRPN, high target switch success is achieved with short (1 s) perturbations, whereas the motion-based SORT tracker requires longer durations and additional online optimization to reach similar performance. Empirically, excessively long attacks reduce stealth by causing noticeable camera shifts or leading to target loss instead of target switch. Hence, the attacker should determine an appropriate duration through offline testing on the target gimbal system.

Maximum angular speed injected. Our attack remains stealthy, as the attacker-controlled object follows the victim at an unsuspicious distance (1–2 m or in an adjacent lane) and the acoustic signal is invisible. On average, successful

trials induce no more than a 30.2° gimbal movement. Moreover, tracker confidence levels remain consistent before and after the attack, suggesting that the perturbation is difficult to detect. Stealthiness can also be tuned by parameters such as maximum gimbal speed, enabling a trade-off between attack effectiveness and detectability. Figure 16 shows attack success rates with respect to gimbal angular speed, and they are indeed strongly correlated.

Until now, we have identified multiple parameters that influence stealthiness, as summarized in Finding 4.

Finding 4: An attacker can tune attack parameters to trade off stealthiness against effectiveness: placing objects too close makes their motion suspicious, while long attack durations or high gimbal angular speeds cause noticeable camera viewpoint changes.

Surrogate model choices. In Figure 16, we evaluate attacks using different combinations of surrogate and target trackers, such as applying SORT to attack SiamRPN and vice versa. Transferability is limited between motion-based and appearance-aware trackers, consistent with their distinct design principles. This suggests that attackers can perform offline attack before the online attack to select the surrogate model, or infer the type of the UAV tracker. See Finding 5

Finding 5: Beyond gimbal profiling, the attacker can further improve success by inferring whether the UAV uses a motion-based or appearance-aware tracker and identifying effective attack durations or angular speeds.

Attacker-controlled object motion. Beyond noisy pedestrian motion, we evaluate more realistic motion models [66], [67], where the attacker-controlled object follows physically plausible, smooth trajectories. Under the HighEndDrone main settings, the attack achieves 78.2% (SiamRPN) and 67.2% (SORT) target switch success rates, demonstrating robustness to constrained and imprecise object motion.

Gimbal motion pattern variations. The injected motion pattern depends on resonant frequency and amplitude, which may drift due to environmental and hardware factors. To evaluate robustness, we execute an attack plan optimized for 4 Hz under perturbed conditions (3 Hz with profiled amplitudes and noise), resulting in a biased oscillation that differs from the planned pattern. The attack remains effective, achieving 77.8% (SiamRPN) and 55.6% (SORT) target switch success.

5.2.5. Attack Efficiency. Without dedicated optimization, our planning step runs at ∼7.6 Hz on our machine, which is fast enough to tolerate the selected cycle time (4 Hz). The main bottleneck for a real-time attack is the gradient optimization process. The attacker can parallelize the gradient optimization process (e.g., gradient step and EoT) using dedicated GPU devices.

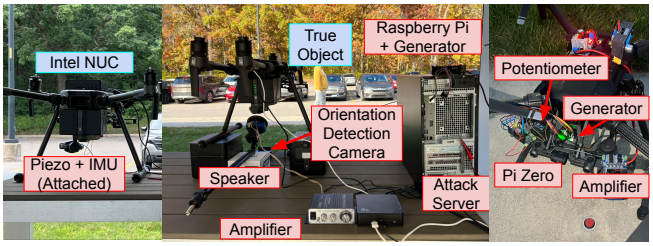

Figure 17: Benchtop direct-contact (left), benchtop contact-free (middle), in-flight direct-contact (right) attack testbeds.

5.3. Physical Experiments

In this section, we demonstrate real-world attack performance through physical experiments on HighEndDrone.

5.3.1. Experiment Setup. As shown in Figure 17, we deploy three types of physical experiments.

- *Benchtop direct-contact*: The drone is placed on an elevated platform about 1.5 m above ground, representing typical consumer tracking scenarios at person height. The gimbal camera is driven by an Intel NUC 13 Pro (Core i5), a common companion computer for onboard tracking algorithms. For acoustic injection, we attach a piezoelectric transducer as described in Section 5.1.
- *Benchtop contact-free*: A Fostex FT17H speaker radiates the crafted acoustic signal through air. An EMEET C960 webcam observes the gimbal from below. The camera feed is processed with an optical-flow algorithm to estimate real-time gimbal orientation. Other hardware and configurations match the direct-contact benchtop setup.
- *In-flight direct-contact*: We build a compact malicious payload with multiple low-power integrated-circuits (which could be reduced to a single embedded unit with further engineering), containing a Raspberry Pi Zero 2W, function generator (AD9033), amplifier (TDA8932), digital potentiometer (DS3502), and IMU (BMI160). The Pi Zero controls the generator and potentiometer to produce the crafted waveform, delivered by a piezoelectric transducer mounted on the gimbal housing. During tests, the drone hovers at about 5 meters.

The online attack algorithm runs on a desktop with Intel i9-14900K CPU and RTX 4080 GPU. An EMEET C960 Webcam is used to obtain the input for the surrogate tracking algorithm. We use an appearance-aware surrogate tracker as it outperforms motion-based alternatives on HighEndDrone, indicating the tested drone may utilize appearance-aware visual tracking.

We evaluate five representative tracking algorithms (Table 2) in benchtop attacks, and the HighEndDrone built-in HighEndDrone's tracking in the in-flight setting. All trackers achieve real-time performance (∼30 FPS or higher) on an Intel NUC 13 Pro (Core i5), except DaSiamRPN. For each tracker, we conduct 10 trials with two pedestrians of distinct appearance and a false-target separation of approximately 3 m. The attack uses a 4 Hz planning–execution cycle, with

	FPS	Direct-contact	Contact-free	In-flight
SORT	25.0	60%/100%	60%/100%	-
UCMCTrack	24.9	80%/100%	100%/100%	-
SiamRPN	23.4	90%/100%	90%/90%	-
DaSiamRPN	7.5	90%/90%	100%/100%	-
KCF	30.0	70%/100%	70%/90%	-
HighEndDrone	-	-	-	60%/80%

TABLE 4: Target switch/disable attack success rate of real-world physical experiments.

one gradient descent step per cycle and three trajectory samples for expectation-over-transformation.

5.3.2. Physical Experiment Results. The attack achieves an average target switch success rate of **79.1%/95.5%** across all benchtop and in-flight trials (Table 4). The overall performance is consistent with the simulation results (Table 3), validating the attack's practicality in the real world.

Across the three experiment settings, as the scenario becomes more challenging in terms of reduced attacker capability and increasing real-world uncertainty and latency, the proposed attack demonstrates stable success. The comparable results between benchtop direct-contact and benchtop contact-free attacks demonstrate that physical access to the victim UAV is not necessary to launch the attack.

More importantly, the attack achieves 60% target switch success rate against the built-in tracking of the industrial-grade HighEndDrone during flight, with complete black-box knowledge (Figure 18). This not only demonstrates that the attack is physically achievable against an operating UAV system but also confirms this security vulnerability in commercial closed-source UAVs.

Failure cases of target switch (target loss and tracking maintained) typically include one or more transient target switch successes across non-consecutive frames. This suggests that the failures arise from acoustic control noise in real-world environments rather than inherent robustness in the target system. The results lead to Finding 6.

Finding 6: Physical experiments confirm that both direct-contact and contact-free attack vectors can corrupt UAV tracking, including successful exploits on a fully closed-source drone, although real-world environmental noise introduces additional challenges for the acoustic attack.

6. Discussion

Potential defenses (details in Appendix A). Several mitigation strategies can be applied against `Banshee`, but each has notable limitations. Hardware approaches aim to eliminate the attack vector via acoustic isolation of the gyroscope [14], [19] or improved signal conditioning and ADC design [14]. However, these solutions are costly, require modifications to off-the-shelf components, and often entail substantial gimbal redesign, making them primarily viable for future UAV systems. Software defenses are easier to deploy, but remain challenging: `Banshee` induces large, rapid motions, while image stabilization demands high-rate processing. Existing detection methods based on simple motion models [68], [69], [70], sensor fusion [71], or specialized hardware [72] also do not align well with typical UAV gimbal architectures.

Figure 18: Before/after target switch HighEndDrone's tracking. First row: daytime. Second row: at dusk.

There are promising defense directions. One option, inspired by GPS spoofing defenses, is to use image-based signals such as visual odometry to detect inconsistencies between visual and inertial measurements [73], [74]. This approach is lightweight and software-based, but can be bypassed by sophisticated adversaries, as visual pipelines themselves are also vulnerable [11], [75], [76].

Attack controllability. Our attack achieves empirical success using a learning-based gimbal response model under a simplified control abstraction. A more principled control theory analysis of modern gimbals with multi-loop control and sensor fusion could further improve the fidelity of this gimbal control approximation. In addition, reliably selecting a specific target in the presence of multiple nearby objects remains an open challenge for practical deployment. We leave these directions to future work.

Remote attack feasibility. While we demonstrate attack success in physical experiments under the benchtop contact-free setup, we observe that motion injection is sensitive to sound pressure level (SPL). In our setup, placing the speaker approximately 4 inches below the gimbal achieves $\sim$110 dB, inducing sufficient motion. Prior work [14], [15], [16], [17], [19], [40] demonstrates acoustic attacks over distances ranging from 10 cm to 7.6 m, while long-range ultrasonic emitters [19], [41] and laser-based approaches [18] can potentially extend attack distances beyond 100 m.

Threats to validity. While simulation enables diverse scenarios, it cannot fully capture real-world complexity. Our physical experiments validate the attack's practicality in physical world, though under limited scenario diversity. Future work will expand real-world evaluations to additional gimbal and system models.

7. Conclusion

This paper introduces `Banshee`, the first physically realizable target switch attack on UAV visual tracking via

acoustic injection on gimbal-camera systems. By empirically profiling gimbal acoustic responses, Banshee generates perturbations that induce directionally biased motion and probabilistically redirect tracking to another object. Our pipeline achieves $> 90\%$ success in simulation and successful black-box attacks on a commercial drone. These results show that acoustic vulnerabilities extend beyond sensor disruption to application-level compromise.

Acknowledgments. We thank the anonymous reviewers and our shepherd for their valuable feedback and constructive suggestions throughout the revision process. This work was partly supported by NSF under the Grant CNS-2321532 and the National AI Institute for Edge Computing Leveraging Next Generation Wireless Networks, Grant # 2112562.

Ethical Considerations

This work reveals an end-to-end vulnerability in UAV visual tracking systems, linking acoustic injection at the sensor level to application-level target-following failures. Given the potential safety risks, we take several measures to ensure responsible conduct.

Responsible disclosure. We have contacted the vendor of the UAV and gimbal systems used and shared our findings prior to publication. Product identifiers are anonymized to reduce misuse risk. We will follow coordinated disclosure practices and release full details only after mitigations are available.

Controlled experiments. All physical experiments were conducted in controlled environments using a single UAV at low altitude within confined areas. Tests were performed under supervision, including benchtop setups and carefully managed outdoor trials, ensuring no risk to bystanders, property, or other aircraft.

Limited artifact release. To support reproducibility, we release simulation code, profiling tools, and non-sensitive datasets, but exclude any software or hardware details that enable direct acoustic exploitation. The artifacts allow validation in simulation and study the implications for defense without providing end-to-end attack capabilities.

Purpose of research. Our goal is to expose a critical class of vulnerabilities in widely deployed UAV systems, not to enable misuse. By demonstrating the feasibility of acoustic-induced target switching, we highlight the need for more robust gimbal design, sensor fusion, and tracking algorithms. We believe responsible disclosure of these findings will support timely mitigation by manufacturers, researchers, and regulators.

In summary, this work follows principles of responsible disclosure, controlled experimentation, limited release, and a focus on improving UAV system safety and robustness.

LLM Usage Considerations

Originality. LLMs were used only for editorial refinement (e.g., clarity, grammar, and style). All scientific content—including problem formulation, methodology, experiments, and analysis—was developed and validated by the authors. The literature review and citation decisions were performed manually. In accordance with conference guidelines: *"LLMs were used for editorial purposes in this manuscript, and all outputs were inspected by the authors to ensure accuracy and originality."*

Transparency. LLMs are not part of our methodology or evaluation. All experiments and analyses are fully reproducible without any LLM service, and no technical contribution or conclusion depends on LLM outputs.

Responsibility. We did not train or fine-tune any LLMs or collect data for that purpose, nor did we provide proprietary or sensitive information to LLM services. As such, the environmental and ethical impact is minimal, and all contributions remain the authors' own.

References

[1] Z. Han, R. Zhang, N. Pan, C. Xu, and F. Gao, "Fast-tracker: A robust aerial system for tracking agile target in cluttered environments," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 328–334.

[2] H. Cheng, L. Lin, Z. Zheng, Y. Guan, and Z. Liu, "An autonomous vision-based target tracking system for rotorcraft unmanned aerial vehicles," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 1732–1738.

[3] A. Maalouf, N. Jadhav, K. M. Jatavallabhula, M. Chahine, D. M. Vogt, R. J. Wood, A. Torralba, and D. Rus, "Follow anything: Open-set detection, tracking, and following in real-time," *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3283–3290, 2024.

[4] DJI, "Best drones that follow you automatically (2024)," https://store.dji.com/content/camera-drone-that-follows-you, 2024, accessed: 2025-08-08.

[5] Skydio, "The best follow me drone in 2022," https://www.skydio.com/blog/10-reasons-skydio-makes-the-best-follow-me-drone, 2022, accessed: 2025-02-25.

[6] Autel, "Autel evo ii drone dynamic track mode full review," https://www.autelpilot.com/blogs/buying-guides/autel-evo-ii-drone-dynamic-tracking-mode, 2020, accessed: 2025-02-25.

[7] H. B. Salameh, M. Alhafnawi, A. Masadeh, and Y. Jararweh, "Federated reinforcement learning approach for detecting uncertain deceptive target using autonomous dual uav system," *Information Processing & Management*, vol. 60, no. 2, p. 103149, 2023.

[8] J. Li, J. Brewington, Q. Zhang, and Z. M. Mao, "Wip: Hijacking attacks on uav follow-me systems in realistic scenarios."

[9] J. Hibberd, "Using drones for peeping, burglaries on rise: "it's gotten dramatically worse"," https://www.hollywoodreporter.com/lifestyle/lifestyle-news/drones-spying-robberies-solutions-hollywood-1236166714/, 2025, accessed: 2025-08-13.

[10] P. Dolan, "Like moths to a false flame: Lethality and protection through deception operations," https://www.army.mil/article/286861/like_moths_to_a_false_flame_lethality_and_protection_through_deception_operations, 2025, accessed: 2025-08-13.

[11] D. Davidson, H. Wu, R. Jellinek, V. Singh, and T. Ristenpart, "Controlling {UAVs} with sensor input spoofing attacks," in *10th USENIX workshop on offensive technologies (WOOT 16)*, 2016.

[12] D. Bereska, K. Daniec, S. Fraś, K. Jedrasiak, M. Malinowski, and A. Nawrat, "System for multi-axial mechanical stabilization of digital camera," in *Vision Based Systems for UAV Applications*. Springer, 2013, pp. 177–189.

[13] A. Altan and R. Hacıoğlu, "Model predictive control of three-axis gimbal system mounted on uav for real-time target tracking under external disturbances," *Mechanical Systems and Signal Processing*, vol. 138, p. 106548, 2020.

[14] T. Trippel, O. Weisse, W. Xu, P. Honeyman, and K. Fu, "Walnut: Waging doubt on the integrity of mems accelerometers with acoustic injection attacks," in *2017 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2017, pp. 3–18.

[15] Y. Tu, Z. Lin, I. Lee, and X. Hei, "Injected and delivered: Fabricating implicit control over actuation systems by spoofing inertial sensors," in *27th USENIX Security Symposium (USENIX Security 18)*. Baltimore, MD: USENIX Association, Aug. 2018, pp. 1545–1562. [Online]. Available: https://www.usenix.org/conference/usenixsecurity18/presentation/tu

[16] Y. Son, H. Shin, D. Kim, Y. Park, J. Noh, K. Choi, J. Choi, and Y. Kim, "Rocking drones with intentional sound noise on gyroscopic sensors," in *24th USENIX security symposium (USENIX Security 15)*, 2015, pp. 881–896.

[17] M. Gao, L. Zhang, L. Shen, X. Zou, J. Han, F. Lin, and K. Ren, "Kite: Exploring the practical threat from acoustic transduction attacks on inertial sensors," in *Proceedings of the 20th ACM conference on embedded networked sensor systems*, 2022, pp. 696–709.

[18] N. Shamsi, K. Chandrasekar, Y. Long, C. Limbach, K. Rebello, and K. Fu, "Wip: Threat modeling laser-induced acoustic interference in computer vision-assisted vehicles."

[19] X. Ji, Y. Cheng, Y. Zhang, K. Wang, C. Yan, W. Xu, and K. Fu, "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision," in *2021 IEEE Symposium on Security and Privacy (SP)*, 2021.

[20] Y. J. Jia, Y. Lu, J. Shen, Q. A. Chen, H. Chen, Z. Zhong, and T. W. Wei, "Fooling detection alone is not enough: Adversarial attack against multiple object tracking," in *International Conference on Learning Representations (ICLR'20)*, 2020.

[21] R. Muller, Y. Man, Z. B. Celik, M. Li, and R. Gerdes, "Physical hijacking attacks against object trackers," in *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2022, pp. 2309–2322.

[22] C. Ma, N. Wang, Z. Zhao, Q. Wang, Q. A. Chen, and C. Shen, "Controlloc: Physical-world hijacking attack on visual perception in autonomous driving," *arXiv preprint arXiv:2406.05810*, 2024.

[23] C. Wang, Y. Man, R. Muller, M. Li, Z. B. Celik, R. Gerdes, and J. Petit, "Physical id-transfer attacks against multi-object tracking via adversarial trajectory."

[24] S. Xie, M. H. Fakih, J. Lu, F. Alshammari, N. Wang, T. Sato, H. Bouzidi, M. A. A. Faruque, and Q. A. Chen, "Flytrap: Physical distance-pulling attack towards camera-based autonomous target tracking systems," 2025. [Online]. Available: https://arxiv.org/abs/2509.20362

[25] J. Ji, N. Pan, C. Xu, and F. Gao, "Elastic tracker: A spatio-temporal trajectory planner for flexible aerial tracking," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 47–53.

[26] Y. Gao, J. Ji, Q. Wang, R. Jin, Y. Lin, Z. Shang, Y. Cao, S. Shen, C. Xu, and F. Gao, "Adaptive tracking and perching for quadrotor in dynamic scenarios," *IEEE Transactions on Robotics*, vol. 40, pp. 499–519, 2023.

[27] S. Liu, X. Li, H. Lu, and Y. He, "Multi-object tracking meets moving uav," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8876–8885.

[28] Y. Li, C. Fu, F. Ding, Z. Huang, and G. Lu, "Autotrack: Towards high-performance visual tracking for uav with automatic spatio-temporal regularization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 923–11 932.

[29] J. Ye, C. Fu, F. Lin, F. Ding, S. An, and G. Lu, "Multi-regularized correlation filter for uav tracking and self-localization," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6004–6014, 2021.

[30] D. Xing, N. Evangeliou, A. Tsoukalas, and A. Tzes, "Siamese transformer pyramid networks for real-time uav tracking," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2022, pp. 2139–2148.

[31] Z. Cao, Z. Huang, L. Pan, S. Zhang, Z. Liu, and C. Fu, "Tctrack: Temporal contexts for aerial tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 14 798–14 808.

[32] B. Yan, H. Peng, J. Fu, D. Wang, and H. Lu, "Learning spatio-temporal transformer for visual tracking," *CoRR*, vol. abs/2103.17154, 2021. [Online]. Available: https://arxiv.org/abs/2103.17154

[33] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg, "ATOM: accurate tracking by overlap maximization," *CoRR*, vol. abs/1811.07628, 2018. [Online]. Available: http://arxiv.org/abs/1811.07628

[34] G. Bhat, M. Danelljan, L. V. Gool, and R. Timofte, "Learning discriminative model prediction for tracking," *CoRR*, vol. abs/1904.07220, 2019. [Online]. Available: http://arxiv.org/abs/1904.07220

[35] Y. Du, J. Wan, Y. Zhao, B. Zhang, Z. Tong, and J. Dong, "Giaotracker: A comprehensive framework for mcmot with global information and optimizing strategies in visdrone 2021," in *Proceedings of the IEEE/CVF International conference on computer vision*, 2021, pp. 2809–2819.

[36] J. Jeong, D. Kim, J.-H. Jang, J. Noh, C. Song, and Y. Kim, "Unrocking drones: Foundations of acoustic injection attacks and recovery thereof." in *NDSS*, 2023.

[37] C. Ma, N. Wang, Q. A. Chen, and C. Shen, "Wip: Towards the practicality of the adversarial attack on object tracking in autonomous driving," in *ISOC Symposium on Vehicle Security and Privacy (VehicleSec)*, 2023.

[38] T. Liu, C. Yang, X. Liu, R. Han, and J. Ma, "Rpau: Fooling the eyes of uavs via physical adversarial patches," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 3, pp. 2586–2598, 2023.

[39] R. R. Wiyatno and A. Xu, "Physical adversarial textures that fool visual object tracking," in *Proceedings of the IEEE/CVF International Conference on computer vision*, 2019, pp. 4822–4831.

[40] W. Zhu, X. Ji, Y. Cheng, S. Zhang, and W. Xu, "TPatch: A triggered physical adversarial patch," in *32nd USENIX Security Symposium (USENIX Security 23)*. Anaheim, CA: USENIX Association, Aug. 2023, pp. 661–678. [Online]. Available: https://www.usenix.org/conference/usenixsecurity23/presentation/zhu

[41] wikipedia, "Long-range acoustic device," https://en.wikipedia.org/wiki/Long-range_acoustic_device, 2025, accessed: 2025-08-24.

[42] X. Yang, J. Yang, J. Yan, Y. Zhang, T. Zhang, Z. Guo, X. Sun, and K. Fu, "Scrdet: Towards more robust detection for small, cluttered and rotated objects," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8232–8241.

[43] C. Xu, J. Ding, J. Wang, W. Yang, H. Yu, L. Yu, and G.-S. Xia, "Dynamic coarse-to-fine learning for oriented tiny object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7318–7328.

[44] OlliW, "Storm32bgc gimbal controller," https://www.olliw.eu/storm32bgc-v1-wiki/Tuning_Guide, 2015, accessed: 2025-11-07.

[45] BaseCam, "Simplebgc gimbal controller," https://www.basecamelectronics.com/, 2025, accessed: 2025-11-07.

[46] InvenSense, "Mpu-6000 and mpu-6050 product specification," https://invensense.tdk.com/wp-content/uploads/2015/02/MPU-6000-Datasheet1.pdf, 2025, accessed: 2025-11-10.

[47] STMicroelectronic, "L3g4200d product specification," http://wikitronica.labc.usb.ve/images/f/fc/L3G4200D.pdf, 2025, accessed: 2025-11-10.

[48] BOSCH, "Bmi160 data sheet," https://www.bosch-sensortec.com/media/boschsensortec/downloads/datasheets/bst-bmi160-ds000.pdf, 2025, accessed: 2025-11-10.

[49] STMicroelectronic, "Lsm6ds3tr-c data sheet," https://www.st.com/resource/en/datasheet/lsm6ds3tr-c.pdf, 2025, accessed: 2025-11-10.

[50] R. J. Rajesh and P. Kavitha, "Camera gimbal stabilization using conventional pid controller and evolutionary algorithms," in *2015 International Conference on Computer, Communication and Control (IC4)*, 2015, pp. 1–6.

[51] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *2016 IEEE international conference on image processing (ICIP)*. Ieee, 2016, pp. 3464–3468.

[52] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "High-speed tracking with kernelized correlation filters," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 3, pp. 583–596, 2014.

[53] B. Li, J. Yan, W. Wu, Z. Zhu, and X. Hu, "High performance visual tracking with siamese region proposal network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 8971–8980.

[54] R. E. Kalman, "A new approach to linear filtering and prediction problems," 1960.

[55] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sendai, Japan, Sep 2004, pp. 2149–2154.

[56] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan, "Siamrpn++: Evolution of siamese visual tracking with very deep networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4282–4291.

[57] Z. Zhu, Q. Wang, B. Li, W. Wu, J. Yan, and W. Hu, "Distractor-aware siamese networks for visual object tracking," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 101–117.

[58] C. Fu, K. Lu, G. Zheng, J. Ye, Z. Cao, B. Li, and G. Lu, "Siamese object tracking for unmanned aerial vehicle: A review and comprehensive analysis," *Artificial Intelligence Review*, vol. 56, no. Suppl 1, pp. 1417–1477, 2023.

[59] Ultralytics, "Quick start guide: Nvidia jetson with ultralytics yolo11," https://docs.ultralytics.com/guides/nvidia-jetson/#nvidia-jetson-agx-orin-developer-kit-64gb, 2025, accessed: 2025-07-04.

[60] K. Yi, K. Luo, X. Luo, J. Huang, H. Wu, R. Hu, and W. Hao, "Ucmctrack: Multi-object tracking with uniform camera motion compensation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, no. 7, 2024, pp. 6702–6710.

[61] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International journal of computer vision*, vol. 115, no. 3, pp. 211–252, 2015.

[62] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke, "Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5296–5305.

[63] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.

[64] P. D. Team, "Px4 autopilot," https://github.com/PX4/PX4-Autopilot, 2025, accessed: 2025-08-26.

[65] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[66] M. Moussaïd, N. Perozo, S. Garnier, D. Helbing, and G. Theraulaz, "The walking behaviour of pedestrian social groups and its impact on crowd dynamics," *PloS one*, vol. 5, no. 4, p. e10047, 2010.

[67] D. Yang, Ü. Özgüner, and K. Redmill, "A social force based pedestrian motion model considering multi-pedestrian interaction with a vehicle," *ACM Transactions on Spatial Algorithms and Systems (TSAS)*, vol. 6, no. 2, pp. 1–27, 2020.

[68] R. Quinonez, J. Giraldo, L. Salazar, E. Bauman, A. Cardenas, and Z. Lin, "Savior: securing autonomous vehicles with robust physical invariants," in *Proceedings of the 29th USENIX Conference on Security Symposium*, ser. SEC'20. USA: USENIX Association, 2020.

[69] Y. Wang, C. Sun, Q. Liu, B. Su, Z. Zhang, M. Norris, G. Tan, and J. Ma, "Vimu: Effective physics-based realtime detection and recovery against stealthy attacks on uavs," 2025. [Online]. Available: https://arxiv.org/abs/2504.20569

[70] H. Meng, S. Luo, Z. Liang, Q. Huang, A. Khazraei, and M. Pajic, "Mars: Defending unmanned aerial vehicles from attacks on inertial sensors with model-based anomaly detection and recovery," 2025. [Online]. Available: https://arxiv.org/abs/2505.00924

[71] S. Lee, "Gyro-mag: Attack-resilient system based on sensor estimation," *Sensors*, vol. 25, no. 7, 2025. [Online]. Available: https://www.mdpi.com/1424-8220/25/7/2208

[72] Y. Wang, Y. Tu, S. Rampazzi, Z. Lin, I. Lee, and X. Hei, *ADC-Bank: Detecting Acoustic Out-of-Band Signal Injection on Inertial Sensors*, 02 2024, pp. 53–72.

[73] N. Gu, F. Xing, and Z. You, "Gnss spoofing detection based on coupled visual/inertial/gnss navigation system," *Sensors*, vol. 21, no. 20, 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/20/6769

[74] M. Varshosaz, A. Afary, B. Mojaradi, M. Saadatseresht, and E. Ghanbari Parmehr, "Spoofing detection of civilian uavs using visual odometry," *ISPRS International Journal of Geo-Information*, vol. 9, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2220-9964/9/1/6

[75] A. Ranjan, J. Janai, A. Geiger, and M. J. Black, "Attacking optical flow," 2019. [Online]. Available: https://arxiv.org/abs/1910.10053

[76] J. Schmalfuss, P. Scholze, and A. Bruhn, "A perturbation-constrained adversarial attack for evaluating the robustness of optical flow," 2022. [Online]. Available: https://arxiv.org/abs/2203.13214

[77] J. Jeong, D. Kim, J. Jang, J. Noh, C. Song, and Y. Kim, "Unrocking drones: Foundations of acoustic injection attacks and recovery thereof," 01 2023.

[78] H. Sathaye, M. Strohmeier, V. Lenders, and A. Ranganathan, "An experimental study of {GPS} spoofing and takeover attacks on {UAVs}," in *31st USENIX security symposium (USENIX security 22)*, 2022, pp. 3503–3520.

[79] D. F. Kune, J. Backes, S. S. Clark, D. Kramer, M. Reynolds, K. Fu, Y. Kim, and W. Xu, "Ghost talk: Mitigating emi signal injection attacks against analog sensors," in *2013 IEEE Symposium on Security and Privacy*, 2013, pp. 145–159.

Appendix A. Detailed Discussion of Defenses

Physical layer. A direct mitigation is to eliminate the attack vector via improved physical housing of the gimbal gyroscope. Acoustic isolation using microfibrous metallic fabric, MEMS acoustic metamaterials, or acoustic foam can attenuate injected signals [14], [19]. However, these approaches increase weight, cost, and design complexity, and may still be bypassed by strong acoustic signals.

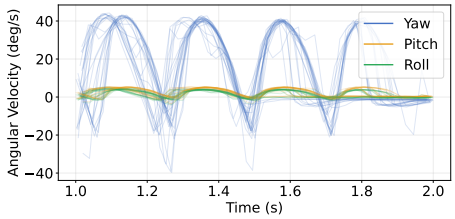

Figure 19: 3-axis angular velocity readings over 25 consecutive injections with directional biasing.

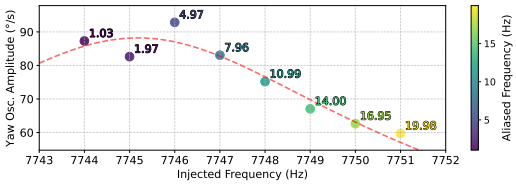

Figure 20: Injected signal frequency vs. yaw oscillation amplitude. Aliased signal frequency shown as changing color. Curve fitted by regression shown as dotted line.

Hardware layer. Secure sensor design can mitigate acoustic injection by strengthening signal conditioning in MEMS IMUs. Gyroscopes and accelerometers infer motion from small proof-mass voltages that are amplified, filtered, and sampled. Acoustic attacks exploit weaknesses in this pipeline, such as amplifier clipping (introducing DC bias) or inadequate low-pass filtering. Secure sampling techniques, e.g., randomized or out-of-phase sampling aligned with sensor resonance, can further reduce vulnerability [14]. However, these defenses require IMU or ADC modifications and are typically infeasible for commodity UAV systems.

Software layer. Software-based defenses include digital image stabilization and learning-based denoising [19], [77]. However, `Banshee` induces much larger motion than typical blur, limiting the effectiveness of stabilization. Learning-based filtering can suppress injected noise but requires large models and extensive training, introducing latency incompatible with real-time gimbal stabilization.

Detection. Detection methods include model-based approaches (e.g., SAVIOR [68], VIMU [69], MARS [70]) that predict IMU readings, and sensor-fusion approaches (e.g., Gyro-Mag [71], ADC-Bank [72]) that detect inconsistencies across sensors. However, model-based methods assume tractable system dynamics (e.g., rigid-body or aerodynamic models), which do not apply to gimbals with nonlinear controllers and friction. Sensor-fusion methods require additional sensors or hardware redundancy, which are often unavailable in gimbal cameras and may themselves be vulnerable to attacks [78], [79].

Camera motion compensation. Incorporating camera motion compensation (CMC) into tracking can reduce attack success rates, especially when trading target switch for target loss. However, the reduction is insufficient to fully mitigate `Banshee`. Moreover, CMC incurs significant computational overhead, limiting feasibility on resource-constrained UAVs, particularly when combined with heavy tracking models [60].

Visual odometry. Visual odometry (VO), which estimates camera egomotion from image sequences, can be used to detect acoustic injection by comparing VO estimates with gyroscope readings. Under normal operation, the two are aligned; under attack, they diverge. VO can run in parallel with stabilization and tracking, requiring only the onboard camera, making it lightweight and deployable via software updates. However, VO is also vulnerable to spoofing, and a sophisticated attacker may evade detection by aligning both modalities [11], [75], [76].

Recovery. Upon detection, the system can enter a recovery mode, e.g., limiting gimbal angular velocity to reduce attack impact, which we observe lowers attack success rates. Alternatively, the UAV can hover or land to preserve safety. These strategies improve robustness but may degrade normal tracking performance.

Appendix B. Gimbal Oscillation with Directional Biasing

Figure 19 shows 25 consecutive trials of injection at 23232 Hz, 50% signal power, and clockwise directional biasing in the yaw axis. Pitch, roll, and yaw axis angular velocities are overlaid into a single plot to show we achieve a consistent induced motion with little variation. While some traces show a slight variation in frequency, the amplitude remains consistent, ensuring that the accumulated angular displacement of the gimbal between different trials remains nearly identical as previously shown in Figure 7.

Appendix C. Effects of Detuning Injected Frequency

As the injected frequency f deviates from the natural resonant frequency f_n, the aliased frequency increases while the oscillation amplitude decreases. Equation (4) shows a linear relationship between f and f_d. The amplitude follows from Equation (5), extended with mechanical impedance:

$$Z_m = \sqrt{(4\pi f_n \zeta)^2 + \frac{1}{(2\pi f)^2}\left((2\pi f_n)^2 - (2\pi f)^2\right)^2}, \quad (11)$$

where ζ is the damping ratio. As $|f - f_n|$ increases, Z_m grows, reducing amplitude.

In our evaluated systems, this reduction is moderate, but it becomes important when signal strength is limited or damping is high. This creates a trade-off: larger amplitudes increase attack impact, while higher frequencies yield more stable responses and faster optimization. We model this trade-off by fitting ζ, F_0, m, and f_n via regression to measured data. The aliased frequency follows a near-linear trend ($R^2 = 0.99$), enabling accurate prediction. Figure 20 confirms that as f moves away from f_n, aliased frequency increases while amplitude decreases, and that the fitted model closely matches observations.

Appendix D. Meta-Review

The following meta-review was prepared by the program committee for the 2026 IEEE Symposium on Security and Privacy (S&P) as part of the review process as detailed in the call for papers.

D.1. Summary

This paper presents `Banshee`, an acoustic injection attack against UAV gimbal-stabilized visual tracking systems. The attack exploits MEMS gyroscope resonance to induce abnormal gimbal motion and camera viewpoint drift, which can cause visual trackers to lose the true target or switch to a nearby object. The paper develops an attack pipeline combining offline gimbal profiling with an online adaptive signal injection strategy. The evaluation includes both simulation and real-world experiments on a commercial drone platform, demonstrating how sensor-level disturbances can propagate through stabilization and affect application-level tracking behavior.

D.2. Scientific Contributions

3) Creates a New Tool to Enable Future Science.
4) Provides a Valuable Step Forward in an Established Field.
5) Establishes a New Research Direction.

D.3. Reasons for Acceptance

1) The paper demonstrates a practically relevant cross-domain attack path in autonomous systems, showing that acoustic injection at the sensing layer can propagate through gimbal stabilization and induce failures in visual tracking pipelines used by UAVs.
2) The work presents a complete attack workflow that combines offline profiling with an online adaptive attack loop, and evaluates the approach across multiple tracking algorithms and experimental settings.
3) The paper provides a useful experimental framework and methodology for studying perception-layer vulnerabilities in UAV systems, which can support follow-on research in robustness, detection, and mitigation.

D.4. Noteworthy Concerns

1) **Interpretation of axis-dependent behavior**. The paper identifies axis-dependent responses (e.g., yaw-dominant behavior) through black-box profiling. While the empirical procedure is clearly described, it remains unclear whether such axis-selective resonance can be consistently identified across different gyroscope designs and implementations. For example, it is unclear whether similar yaw-dominant behavior can always be found for arbitrary commercial gimbal systems, or whether the observed behavior depends on specific hardware characteristics.
2) **Limited characterization of intermediate mechanisms**. The paper demonstrates that acoustic injection induces gimbal motion and leads to tracking degradation. However, the relationship between the induced oscillatory motion and the resulting tracking behavior is not fully characterized. In particular, it remains unclear how different motion patterns (e.g., oscillation versus accumulated drift) contribute to the observed tracking outcomes.
3) **Limitations of the in-flight evaluation**. The in-flight experiments demonstrate that acoustic injection can affect the built-in tracking system of a commercial drone. However, the evaluation is conducted under specific conditions (e.g., limited number of targets, controlled scenarios, and close-range acoustic injection), and does not explore how the attack behaves under more diverse real-world settings. As a result, the results should be interpreted within the scope of the evaluated scenarios.