

UNIVERSITY OF WALES TRINITY SAINT DAVID

MSc Cybersecurity and Digital Forensics

**Towards Trustworthy AI-Assisted Drone Evidence
Collection for Security and Digital Forensic Investigations**

Master's Dissertation

submitted in partial fulfilment of the requirements for the degree of

Master of Science in Cybersecurity and Digital Forensics

Rishabh Choudhary

Student ID: 2434332

**Supervisor:
Dr. Dipo Dunsin**

**University of Wales Trinity Saint David
London Campus**

September 2026

Abstract

Drones have become popular tools for application in security, emergency, and surveillance operations where they may provide visual, positioning, and telemetry information that potentially has evidential value. However, the existence of drone information does not necessarily mean the presence of evidence. This paper conducts an analysis and evaluation of scientific literature related to such fields as drone digital forensics, UAV cybersecurity, aerial object detection and tracking, environmental robustness, positioning integrity, provenance of evidence, chain of custody, and human oversight. The literature shows considerable technical progress in separate domains but lacks connections between them. Computer vision studies mainly estimate the quality of predictions, cybersecurity studies deal with threats to navigation and communications, while digital forensic studies concentrate on acquisition, preservation, and examination after the incident. As a result, the question of whether the drone evidence generated with the help of AI algorithms is at the same time integrity, traceable, sufficient, timely, environment-proof, and consistent with telemetry information has not been researched empirically. From the review, eight research gaps are defined, including workflow fragmentation, confusing of AI performance and trust, evidential completeness, environmental robustness, timeliness, cyber-forensic verification, provenance and auditability, and multi-dimensional assessment of evidence. Based on this, the review presents a theoretical basis for the Trustworthy Drone Evidence Framework (TDEF), which estimates evidential trust through its independently interpretable dimensions.

Keywords—drone forensics, UAV cybersecurity, digital forensics, provenance of evidence, object detection, multi-object tracking, environmental robustness, telemetric integration, chain of custody, human verification.

Table of Contents

UNIVERSITY OF WALES TRINITY SAINT DAVID	1
<i>Towards Trustworthy AI-Assisted Drone Evidence Collection for Security and Digital Forensic Investigations</i>	1
1. Introduction	6
2. LITERATURE REVIEW	7
2.1 Review Approach and Analytical Lens	7
2.2 Operational Adoption and Evidential Requirements	8
2.3 Drone Digital Forensics and Forensic Readiness	10
2.4 UAV Cybersecurity, Positioning and Telemetry Integrity	11
2.5 AI-Assisted Aerial Object Detection and Evidential Completeness	13
2.6 Temporal Reconstruction and the Limits of Tracking Identity	14
2.7 Activity and Anomaly Analysis: Useful Context but Insufficient Evidence	15
2.8 Environmental Robustness and Evidential Uncertainty	16
2.9 Digital Evidence Integrity, Provenance and Auditability	17
2.10 Trustworthy AI, Explainability and Human Oversight	19
2.11 Critical Synthesis and Comparative Gap Analysis	20
2.12 From Literature Gaps to the Present TDEF Design	25
2.13 Future Research Trajectory	25
2.14 Conclusion	26
3. Methodology	27
3.1 Research Design	27
3.2 TDEF Framework Design	27
3.2.1 Layer 1 — Evidence Acquisition and Provenance	27
3.2.2 Layer 2 — AI-Assisted Evidence Analysis	28
3.2.3 Layer 3 — Trust and Robustness Validation	28
3.2.4 Layer 4 — Multidimensional TDEF Assessment	28
3.2.5 Layer 5 — Investigator Decision Support	28
3.3 Experimental Environment and Implementation	30
3.4 Evidence Acquisition, Integrity and Provenance	30
3.4.1 Evidence Ingestion	30
3.4.2 Video-to-Frame Provenance	31
3.4.3 Audit Logging	31
3.5 AI-Assisted Object Detection	31

3.6 Ground-Truth Evaluation Using VisDrone2019-DET	32
3.6.1 Dataset Selection	32
3.6.2 Annotation Parsing	32
3.6.3 Class Mapping	32
3.6.4 Intersection-over-Union Matching	33
3.6.5 Precision, Recall and F1	34
3.6.6 Ignored-Region Handling	34
3.6.7 Progressive Validation	35
3.7 Temporal Reconstruction and Tracking Reliability	35
3.8 Environmental Robustness Evaluation	36
3.9 Cyber-Forensic Integrity Validation.....	36
3.9.1 Experimental Baseline and Manipulations Design	36
3.9.2 Independent Validation	37
3.10 Multidimensional TDEF Integration.....	37
3.11 Reproducibility and Experimental Traceability.....	38
3.12 Ethical and Safety Considerations	38
4. Results.....	39
4.1 Experimental Overview.....	39
4.2 Evidence Integrity and Provenance Validation.....	40
4.3 Ground-Truth Object-Detection Performance.....	41
4.4 Per-Class Detection Performance	43
4.5 Temporal Reconstruction and Tracker Comparison	44
4.6 Environmental Robustness Results.....	46
4.7 Cyber-Forensic Integrity Validation.....	47
4.8 Integrated TDEF Evidence Assessment.....	49
4.9 Investigator-Facing Prototype	51
5. Discussion.....	52
5.1 Overview.....	52
5.2 Cryptographic integrity and provenance as necessary but insufficient conditions	52
5.3 AI Reliability, Detection Completeness, and Class-Level Variation	53
5.4 Temporal Continuity Is Useful but Should Not Be Confounded with Physical Identity	54
5.5 Environmental Robustness and Evidential Interpretation	54
5.6 Cyber-Forensic Integrity and Complementary Validation	55
5.7 From Fragments to a Trust Profile	55

5.8 Connection between Results and Identified Research Gaps	57
5.9 Explainability and Human Review.....	58
5.10 Practical Implications and Research Contribution	58
6. Limitations and Future Work.....	59
6.1 Limitations.....	59
6.2 Future Work	60
References:.....	62
Appendices:.....	65
Appendix A — Experimental Evidence Screenshots	65
A.1 — Evidence Provenance Record.....	65
A.2 — The important VisDrone raw result	65
A.3 — Tracker evidence	66
A.4 — Environmental robustness	67
A.5 — Cyber-Forensic Validation	68

1. Introduction

Utilization of unmanned aerial vehicles (UAVs) is quickly evolving. Today, law enforcement and emergency services regard drones not only as imaging devices but as critical components of workflows for time-critical operations. For example, in the UK, the National Police Chiefs' Council identified a Drone as First Responder pilot project in Southampton, where a remotely deployed vehicle was able to deliver a live aerial overview of the situation, track suspects fleeing the crime scene, and record video to be used in investigations [1]. As of 2026, the Metropolitan Police reported that three London drone sites were able to cover approximately 200 cases per week [2]. These examples explain why drone-derived material deserves to be considered in digital forensics: it can both inform operational decision making and serve as material in an investigation.

The issue of evidence goes beyond simple identification of whether a drone was able to capture a valid picture. Drone evidence comes from a cyber-physical system of interacting cameras, positioning sensors, flight controllers, communication links, storage media, software, artificial intelligence models, and investigator's actions. A video file may remain cryptographically valid while carrying inaccurate coordinates. Object detector may give highly confident predictions while failing to detect many objects. A tracker may create a stable identifier while having no independent confirmation of correct association with the physical object at all. Environmental factors may affect detection completeness without affecting the integrity of the underlying file. Likewise, a chain of custody may be valid but fail to provide an independent confirmation that the interpretation provided by the AI model is complete and accurate. Therefore, trust in AI-assisted drone evidence cannot be based on a single technical property.

The importance of the problem arises from the fact that each field of the existing research literature is mature independently but not necessarily in combination with others. Drone forensics literature advanced acquisition practices, the concept of forensic readiness, and knowledge about device-specific artifacts [3]-[8]. Cybersecurity research of UAVs explores various attacks related to spoofing, jamming, insecure communications, and navigation integrity [9]-[13]. Computer vision literature contributes progressively advanced object detectors, benchmark datasets, and multi-object trackers [14]-[20]. Anomaly detection literature provides means of detecting unusual spatial or temporal patterns [21], [22]. Environmental robustness literature shows how visual degradation affects model performance [23], [24]. Principles of integrity, preservation, and traceability of digital evidence and chain of custody literature provide a formal framework [7], [25]-[28]. Finally, trustworthy AI guidelines call for the validity, reliability, transparency, explainability, and human accountability [29]. Each research strand addresses a legitimate concern but not the forensic question of when an investigator should trust AI-assisted interpretation of drone-derived evidence.

The main claim of the present review paper is that AI confidence should not be equated to evidential trust. Evidential trustworthiness should be seen as a set of independently interpretable

properties. This review operationalizes the idea in terms of Trustworthy Drone Evidence Framework (TDEF), which consists of such dimensions as cryptographic integrity, evidence provenance, AI detection reliability, detection completeness, temporal reliability, environmental robustness, cyber-forensic integrity, and auditability. The dimensions are further combined in an assessment that does not reduce them to an arbitrary weighted score. The goal of the literature review is not to claim novelty of any of the considered techniques but to assess how their separation creates a methodological gap and build the scholarly foundations for their integration into a provenance-aware and explainable forensic workflow.

The literature review aims at achieving four goals. First, it assesses the state of drone digital forensics and forensic readiness focusing specifically on evidence acquisition, preservation, and reconstruction. Second, it explores the reliability limits of AI-assisted aerial analysis in terms of object detection, tracking, anomaly analysis, and environmental sensitivity. Third, it evaluates cyber-forensic threats to telemetry and positioning data and discusses how cryptographic and semantic consistency checks can complement each other. Fourth, it identifies research gaps arising from these aspects and develops a research agenda that motivates the implemented TDEF methodology and realistic research trajectory built around synchronized multimodal acquisition, model validation, and calibrated decision boundaries.

2. LITERATURE REVIEW

2.1 Review Approach and Analytical Lens

Unlike the purely descriptive survey, the present review employs an integrative critical approach. As the research problem ranges over digital forensics, cybersecurity of UAVs, computer vision, multi-object tracking, AI robustness, evidence management and privacy, restriction of the review to a single subject discipline database or a single tradition of evaluation risks fragmenting a topic that needs to be integrated. Sources have been found using IEEE Xplore, Google Scholar, ScienceDirect, SpringerLink, ACM-focused indexing, official standards repositories, and official technical and policing authorities. Search terms included UAV/drone, digital forensics, forensic readiness, object detection, tracking, aerial anomaly detection, GPS spoofing, telemetry integrity, environmental degradation, chain of custody, evidence provenance, trustworthy AI, and human review.

The priority has been given to peer-reviewed journal articles, major conferences papers, benchmark papers, standards, and technical guidance from official bodies. Public-sector operational sources have been used selectively to provide the contemporary context rather than to replace the scholarly evidence base. Thus, the review identifies three types of sources: empirical and technical sources are used to evaluate methods and performance; standards and guidance are used to define established principles for handling evidence; and operational sources

are used to exemplify the emerging context in which drone-derived evidence becomes meaningful for the purposes of investigation.

The studies have not been evaluated according to their reported accuracy, mean average precision, inference speed, or other specific metrics. Instead, the analysis considers a set of factors including the research problem being solved, data sources, assumptions about authenticity of evidence, relationship between the data from the source and derived from it, evaluation metrics, environmental realism, temporal reasoning, attack models, reproducibility, auditability, and relevance to investigation decisions. This kind of analysis is necessary since benchmark success and forensic adequacy cannot be considered identical. A detector can perform well in general but fail on small or corrupted objects relevant to a specific case. A spoofing classifier can perform well in the training setting but be prone to novel and slow-changing attacks. A cryptographic hash verifies the byte-level integrity of the data but tells nothing whether the sensor measurement taken was correct.

Moreover, this review distinguishes between three related but different concepts: integrity, reliability, and trustworthiness. Integrity refers to consistency with the recorded state of an artefact. Reliability refers to dependable functioning of an analytical or sensing process in the conditions of use. Trustworthiness is a broader concept that addresses the issue of having enough and explainable basis for professional reliance on the evidence and its interpretation. The definition of trustworthy AI as an attribute set rather than a single scalar property resonates well with the NIST AI Risk Management Framework [29].

A methodological principle followed by the review states that limitations observed in adjacent domains can be considered forensic design requirements only when a defensible connection has been made. Image corruption research does not establish a forensic standard but proves that common degradations may influence the model behavior [24]. This establishes the need to examine the environmental robustness when AI outputs serve as evidential elements. Likewise, the multi-object tracking research literature does not prove the identity but illustrates the mechanisms of temporal association formation and fragmentation [19], [20]. Therefore, this review refrains from transforming the engineering metrics into legal conclusions directly. Instead, it evaluates the legitimacy of each technical finding in an evidence assessment process.

2.2 Operational Adoption and Evidential Requirements

As more unmanned aerial systems become part of police and emergency service operation, it becomes clear that taking a forensic approach to the matter should be taken before a technological deployment is done routinely. The example of the Southampton Drone as First Responder Project, as presented by the National Police Chiefs' Council, demonstrates how appealing it can be to deploy an aircraft remotely for rapid access to situations, live situational awareness, suspect tracking and even capturing the situation on camera to be used as evidence

later [1]. Likewise, there is an expanding London deployment model of the Metropolitan Police where drones attend incidents immediately to get live intelligence before other personnel arrives [2]. Although the above systems can help overcome the information gap in terms of the response to situations, it should be noted that the value of operational systems should not be confused with the evidential status of the materials they generate.

The operational intelligence can be advantageous in situations of uncertainty. A control room can act upon the imperfect live feed since the purpose is resource allocation. Still, the subsequent forensic interpretation will face a different evidential burden. For instance, investigators will have to verify what is recorded, when it is recorded, if the file and metadata were changed, how the analysis was conducted, if some of the scene content was overlooked, and if the telemetry is internally consistent. In other words, while the same image can be useful for the operations, it cannot serve the same purpose in the forensic investigation.

It follows the architecture implication. An evidence-producing drone system should not treat camera stream, telemetry, derived frames, artificial intelligence detection, and investigator decision as interchangeable information objects. The acquired evidence is represented by the source artefact; derived frame stands for a transformation of evidence; artificial intelligence detection is an algorithmic observation; track indicates temporal association between observations; and the investigator decision represents professional judgment. Blurring these layers can mean losing track of where the uncertainty was introduced. The idea behind Targeted Digital Forensic Evaluation (TDEF) is quite the opposite. Any subsequent interpretation must be traceable to the source material and reflect information on how it was done.

There are additional concerns regarding the use of drones as operational tools. According to the Information Commissioner's Office, the use of drone can record the personal data about people who are not intended to be recorded and stresses the importance of conducting data-protection impact assessment, being transparent about the process, having the retention control measures in place and making sure that the system meets data protection by design standards [30]. Likewise, the Civil Aviation Authority states that the use of a camera-equipped drone is subject to respecting privacy and protecting the recorded material [31]. It means that the requirement is concerned with how the governance integrity can be preserved and not with the technical reliability of AI detection. In other words, a system that produces highly accurate analysis but does not keep track of access, purpose and retention of the materials is still not good enough from the governance standpoint.

Hence, the operational literature sets a boundary condition for TDEF in the current research. TDEF is not aimed at automating police decisions or inferring criminal intent. It is more about helping the investigator assess the evidential condition of the AI-assisted drone materials.

2.3 Drone Digital Forensics and Forensic Readiness

The field of drone digital forensics arose due to the recognition of the heterogeneity and distribution of drone incident evidence. Mantas and Patsakis describe the operational, procedural, manufacturer and legal barriers to standard investigation across drone platforms [3]. A contemporary drone environment may include onboard storage, removable media, remote controller, connected mobile application, cloud service, flight logs, network traces, and sensor data. Therefore, the problem goes beyond imaging one single device. The investigator must identify potential evidence, preserve it, and analyze the relationship between distributed artifacts.

The goal of the research on forensic readiness is moving the process from reactionary recovery of evidence towards anticipation of availability and organization. Alotaibi et al. propose the Drone Forensics Readiness Framework comprised of proactive and reactive steps and argue that the occurrence and evidence should be anticipated before the incident happens [4]. This is especially true for evidence producing UAVs because the reaction after the incident may fail to recover the volatile, cloud-hosted or externally produced evidence. Forensic readiness lays the groundwork for systematic logging; nevertheless, its focus is primarily on evidence availability and process organization rather than the accuracy of machine-assisted interpretations.

Real examples of device-focused cases help to see both advantages and limitations of the existing forensic knowledge base. Azhar et al. recover flight information and mobile-device artifacts from DJI Phantom 3 Professional and Parrot AR. Drone 2.0 platforms using open-source tools [5]. Taylor's analysis of the DJI Mini 3 Pro and DJI RC drone identifies recorded media, flight logs, and evidence related to usage and operator reconstruction; nevertheless, the researcher shows how evidence can be absent, removed or hidden [6]. These cases are useful in demonstrating recoverable evidence from a real platform; yet, they face the challenge of reproducibility: the hardware, firmware, applications and storage architecture of the drones change rapidly, which means that the procedure that worked on one platform may not work on another one.

Professional guidance highlights the importance of proper preservation. SWGDE's Best Practices for Drone Forensics is a useful starting point in conducting forensic examination of drones and connected systems that stresses on proper handling, documentation and validation [7]. NIST's CFReDS drone collection includes forensic images and associated devices obtained during controlled acquisitions of multiple drone models [8]. Such resources help in testing of the tools and training of investigators; nevertheless, they are mostly concerned with the acquisition and examination of drones rather than end-to-end validation of machine-assisted evidence interpretation. It is the key difference from the present research. Traditional drone-forensics workflow answers the questions about the existence of the artefact, its location, method of acquisition, and preservation of its integrity. Machine-assisted evidence collection raises a new set of questions is the algorithmic interpretation accurate enough, is it sufficient and robust enough to drive the attention of the investigator. Unchanged frames can still be unreliable; a

complete flight log can contain a false location that was initially accepted by the navigation system; a reproducible detection can still miss a significant part of the relevant objects. Therefore, the notion of forensic readiness needs to be expanded from the readiness to collect evidence to the readiness to assess the trustworthiness of the machine-assisted interpretation.

The approach proposed in the Defense-in-Depth for Evidence and Digital forensics (TDEF) framework helps to achieve this goal. Provenance record links video to extracted frames; cryptographic digests ensure byte-level integrity; audit record documents the events of processing; AI output is linked with the configuration of the model and its evaluation context. The novelty is not in individual hashing or logging of each of these elements; it lies in treating the existing forensic controls as dimensions of a more comprehensive assessment including detection completeness, temporal reliability, environmental robustness, and cross-source consistency.

One of the limitations of the literature on the topic is the lack of public multimodal forensic datasets that provide synchronized aerial video, telemetry, known integrity violations, environmental conditions, ground truth scene annotations, and chain of custody events. CFReDS is useful for the acquisition research, whereas VisDrone and UAVDT are useful for computer vision; yet none of them was created with a consideration of the entire evidential workflow of the present study. This separation of the datasets leads to the necessity to conduct a controlled experiment in the present research as well as creation of the multimodal forensic acquisition testbeds in the future.

2.4 UAV Cybersecurity, Positioning and Telemetry Integrity

Unmanned aerial vehicles (UAVs) are cyber-physical systems in which evidential integrity might be compromised without any visible alteration to the recording itself. Sihag et al. review cybersecurity and forensics challenges for next-generation drones, covering threats to communication, control, software, sensing, and data [9]. From the forensic perspective, the significance of these threats is not limited to potential disruptions of flight; they might also alter the semantics of the recording. The video recording may be authentic at the file level, but the metadata, coordinates, or timestamps may be falsified.

Positioning integrity provides a prominent example. GPS and other GNSS information might be used as an evidence component by demonstrating where and when the event took place. Shafique et al. apply machine learning for discrimination of genuine and spoofed UAV signals [10]. Altaweel et al. in a systematic review of GPS spoofing attacks in flying ad hoc networks identify a wide array of detection, prevention, and mitigation methods, noting that many of them are validated in simulations instead of proof-of-concepts under mobility conditions [11]. Together, these studies reveal a high level of technical research activities, as well as the difficulty of generalizing one single defense against multiple platforms, scenarios, and threats.

A second strand of work applies temporal analysis of behavior, rather than just analyzing the signal characteristics at one point in time. Shin et al. propose deep learning approach for UAV location integrity checking based on the idea that an apparently valid position may become suspicious when viewed in the context of time sequence [12]. Pardhasaradhi and Cenkeramaddi use distributed radar tracking and fusion to compare reported GPS positions of UAVs with independently estimated ones [13]. Independent sensing provides a solid ground for corroboration, but it presumes existence of infrastructure that might not be available during the standard forensic investigation.

The interpretation of the above methods in a forensic context should be performed carefully. Discontinuity, abnormal trajectory, or disagreement of different data sources do not imply the existence of malicious GPS spoofing attacks. Similar results could be produced by sensor failures, logging mistakes, transmission latency, data corruption, or malicious actions. Imputation of the the attack implies additional evidence besides these findings. Thus, a credible framework would report only observable integrity and consistency indicators and leave attribution up to the expert.

The above considerations have shaped the cyber-forensics part of TDEF. Specifically, the implemented study avoids the claim of operational GPS spoofing attack detection. The goal is to determine whether artificial changes in telemetry, timestamps, derived frames, location continuity, and cross-source consistency produce the relevant forensic indications. The cryptographic verification and semantic consistency checking techniques are purposefully constructed as complementary to each other. A hash function verifies the change of an artifact relative to the recorded state; semantic check could expose non-monotonic timestamp sequence, abnormal trajectories, or inconsistent timestamps of different evidence sources. None of the two techniques replaces another.

The literature identifies a rationale behind cross-source validation. Check specific to a particular sensor or a particular model could be bypassed when the corrupted value is locally consistent. A trajectory sequence could look plausible but could disagree with visual timestamps; a telemetry file could keep a realistic coordinate, but fail the recorded digest check; a derived frame could look visually the same to a human observer, but differ at byte-level. Hence, the forensic architecture should preserve the opportunity for cross-validation rather than treat any single sensor stream as an authority.

The above consideration reveals a more generic research gap. Literature on UAV security is primarily focused on attack detection, efficiency of defense mechanisms, or navigation robustness; meanwhile, literature on digital forensics mostly evaluates acquisition and preservation techniques. Relatively small amount of research deals with operationalization of cyber-physical inconsistencies as a forensic dimension along with the evidence trust based on AI-assisted visual analysis, provenance, and auditability. This research gap is addressed by the

conceptualization of cyber-forensic integrity in TDEF as an explainable combination of hash, time, location, trajectory, and cross-source checks.

2.5 AI-Assisted Aerial Object Detection and Evidential Completeness

Computer vision is the analytic technology which makes the evidence collected by drones scalable, while, at the same time, brings up the trust issue at the center of this paper. Air imagery has systematic distinctions from a lot of other cases of computer-vision. Objects can take only a few pixels on the image, differ in scale, form clusters, be often occluded, and be observed from the moving camera. VisDrone is a benchmark for drone detection and tracking, as it is designed to facilitate this kind of research in different aerial scenarios and became the leading one in UAV vision [14]. The UAVDT contains about 80,000 images with weather, altitude, camera viewpoints, and occlusion annotations and shows the difficulties of object detection and tracking under uncontrolled UAV scenario [15]. Furthermore, Wu et al. survey highlights the challenges of small objects, scale variability, motion, and computational limitations in deep learning approaches to detection and tracking for UAV vision [16].

The response to those challenges includes ever-more specialized architectures. For instance, An et al. propose the DCE-YOLOv8 which is the light detector aimed at finding the balance between accuracy and efficiency for drone vision [17]. On the other hand, Li et al. tackle the problem of occlusion in the form of occlusion-guided multi-task network [18]. Such research is important because it determines the ability of the models to operate efficiently in realistic UAV and edge-computing environment.

However, the goal of traditional computer vision differs from the goal of forensic vision. Traditional benchmarks evaluate the models comparing them based on metrics such as mean average precision, class-specific average precision, inference time and alike. Those metrics are adequate for the evaluation of the detectors, but they do not guarantee that output of the certain piece of evidence is enough reliable for investigation purposes. The implication of the false positive and false negative are different in this case. The former may create an unnecessary investigation lead, while the latter will leave the object out of the automated scene reconstruction. Therefore, the evidential significance of the non-detection is particularly sensitive to the recall and scene completeness.

It motivates the introduction of two concepts in the context of TDEF: reliability and completeness of the detections. Reliability refers to the quality of predictions made by the model, whereas completeness to the extent of the ground-truth objects detected. The system can make comparatively precise detections while leaving a significant part of objects undetected. To interpret a relatively high precision value or high-confidence individual bounding box as the indicator of the scene reliability means to confuse the local quality measure of the prediction and the global quality measure of the evidential content.

This distinction is especially important regarding object classes. The detector can perform well detecting large vehicles, but be unable to detect smaller bicycles, motorcycles or people. The aggregate measures can hide this differentiation despite the possible different investigative significance of the different object classes. Thus, the forensic framework should keep, if possible, the per-class performance and should avoid interpreting the lack of detection as the proof of non-existence unless the model proved its sensitivity on this class in these conditions.

The current study uses the pretrained lightweight detector as a baseline, not as the innovative detection architecture. It is the reflection of the intentionality of the methodology. The research question does not concern whether a new model could outperform the state-of-the-art, but whether the traditional AI performance can be translated into the reliable evidential assessment. The ground truth evaluation on the full VisDrone2019-DET validation set allows having transparent statistics of true positives, false positives, and false negatives, precision, recall, and F1-score. They are interpreted as different trust dimensions, not collapsed into one model score.

This approach facilitates the reproducibility of the results. Instead of relying only on detector confidence, which is produced in inference time, the study evaluates the model against the external labeled benchmark. The detector confidence is the useful output of the algorithm, but the ground truth is a more defensible measure of the adequacy of the automated interpretation to be considered as sufficiently complete to rely on in investigation. In this respect, the contribution of the forensic science lies in the interpretation of the traditional detection metrics, not in the innovation of the object detector.

The literature, however, suggests several future directions. One model and one dataset cannot prove the general reliability of the technology. Evaluation of multiple detectors, sensor resolutions, altitudes and domains will show whether the distinction between reliability and completeness of TDEF is stable when the underlying AI is changed. Therefore, the future research should consider model replacement as the test of the framework generalizability, not as the quest for the best universal detector.

2.6 Temporal Reconstruction and the Limits of Tracking Identity

However, still imagery is an incomplete record of events. The temporality of drone video footage makes the objects appear, disappear, move in space, get occluded, and reappear later. Multi-object tracking may be applied in this case to build trajectories out of frame-level detections. However, the problem with tracking is that it introduces an additional inference stage whose validity cannot be overstated.

ByteTrack is based on the idea that skipping detections with low confidence results in lost objects and fragmented tracks. An association strategy of the tracker is based on matching low-confidence detections to existing tracklets rather than skipping them [19]. BoT-SORT is based on the idea that tracking-by-detection may be further improved by adding motion and appearance

cues and introducing camera-motion compensation along with a modified Kalman-filter representation [20]. These trackers show excellent benchmark results; however, their aim is computational: maintain consistent object identities across the video.

A forensic application calls for a more limited interpretation of results. A tracking identifier is an algorithmic label associated with a sequence of observations. This label is not an independent proof that a real-world object is recognized unambiguously. Occlusions, missed detections, similar appearance, camera-motion, and errors in associations may cause fragmentation and identity switching. Therefore, the strongest claim a tracker can make about a reconstructed object is that its algorithm consistently identifies the object under test conditions.

This conclusion suggests that temporal consistency of the tracking should be estimated separately from frame-level detection performance. A tracker may miss many objects but have high continuity of trajectories of the objects that were detected. On the contrary, a good-performing detector may produce fragmented trajectories in the tracking phase. Combining both measures into one AI score will prevent us from identifying the weak point of the process.

In the present research, the performance of ByteTrack and BoT-SORT will be compared on the same brief video sequence using an observation-based continuity measure to determine how long tracks are consistently represented in time. The aim of the experiment is not to define the best tracker in general; it is to determine whether temporal consistency can be presented as an independently interpretable trust signal in TDEF. The approach is in line with the existing literature because ByteTrack and BoT-SORT are built on the assumption of the quality of association; however, it is presented in the context of evidential reconstruction and not leaderboard performance.

The limitation is also important to note. A single video cannot prove the reliability of tracking algorithms in terms of crowded scenes, various altitudes, significant occlusions, and complex camera motion. UAVDT shows that these factors significantly impact the tracking difficulty [15]. Therefore, future experiments should use longer and more diversified video sequences to test for identity switches, re-identification, and sensitivity to camera motion. Moreover, in a mature forensic implementation, temporal consistency could be tested against the telemetry and scene geometry, allowing for corroboration of the trajectory within the video.

2.7 Activity and Anomaly Analysis: Useful Context but Insufficient Evidence

Anomaly detection provides another option for artificial intelligence support. Instead of trying to recognize object categories, anomaly models seek to detect observations or patterns which do not conform with the learned patterns. In their UAV-AdNet project, Bozcan and Kayacan use visual arrangements together with GPS labels to develop an unsupervised aerial surveillance anomaly detector [21]. Jin et al. employ the transformer models for the purpose of anomaly detection in aerial videos and consider the temporal aspect which can prove useful in segmenting and

prioritizing anomalous fragments in long videos [22]. They become especially attractive in cases where reviewing lengthy footage manually becomes unfeasible.

Nevertheless, there is an intrinsic limitation to its forensic meaning. Statistical anomaly does not mean suspiciousness, criminality or maliciousness. Emergency vehicles, a maintenance worker, unusual crowd movements or change of route may be statistical anomalies which are nevertheless completely lawful. On the other hand, criminal or malicious actions might appear statistically normal. Therefore, it makes anomaly detection helpful as a triaging mechanism, but insufficient as an autonomous determination.

It was the initial idea in the conception of this research for there to be a more explicit activity analysis layer to it. As the framework developed, this element did not stay in the framework as a separate trust dimension since the focus on the properties that could be experimentally verified against evidence integrity, ground truth, temporal consistency, controlled corruption and validation became the research goal. However, anomaly-detection literature is still important since it emphasizes the necessity of bounded claims and human judgment, but TDEF does not try to classify criminal or malicious actions.

It makes the framework better, not worse since for it to be publishable it is necessary to be able to make distinction between possible and experimentally verifiable features. The key lesson from anomaly detection literature is the methodological one: AI may prioritize observations, but human judgment should remain in the hands of an investigator, and the framework should preserve the evidence and rules behind the alert.

2.8 Environmental Robustness and Evidential Uncertainty

The main deficiency of the AI-powered interpretation of evidence in question is the hidden premise that sensor view is clear enough. There can be motion blur, haze, low illumination, compression artifacts, vibration, small-object scale, and varying contrast in the aerial imagery. UAVDT takes weather, altitude, and occlusion into account, thereby emphasizing their importance for the performance [15]. According to Fang et al., the problem of foggy aerial detection can be approached via multi-task framework combining dehazing and object detection since poor weather is a serious hindrance to the model performance [23].

Another example gives the broader field of robustness research. Hendrycks and Dietterich demonstrate that commonly used corruptions and perturbations can bring out substantial weaknesses which do not appear evident when working on clean datasets [24]. Although the benchmark presented by the authors is not UAV-related, its fundamental idea is fully applicable – the performance tested under one distribution does not guarantee reliable behavior under degraded conditions. It is especially true in the case of drone-based evidence since the acquisition environment is not so controllable as a benchmark dataset.

Enhancement is also not a solution since the dehazed, sharpened, or denoised picture may become better-looking or help with model performance but alter pixel values and create a new artefact. In terms of forensic requirements, the original image should be kept, and the process of enhancement should be registered in terms of parameters, version of software used, and a digest. While enhancing the image may assist the interpretation, it should not rewrite the evidential material.

Environment robustness also exemplifies why model precision cannot be considered reliably. Under degradation, the detector may be more cautious about its decisions and predict less objects. The fraction of correct predictions among those which are made may become higher while the recall and the overall coverage decrease. Such a framework, reporting only the precision, may create an illusion of increased reliability exactly where there are more missed relevant objects. This scenario confirms the necessity of separating AI detector reliability, completeness, and environmental robustness within the TDEF.

The conducted experiment introduces the concept of environmental robustness in practice through consistent transformations of the same VisDrone dataset annotated beforehand. Controlled transformations like blur, reduction of brightness and contrast, and adding Gaussian noise create repeatable degradation conditions instead of mimicking every real-life environment. The main benefit of the approach is the pair comparison – keeping scene content, ground truth, detector, and thresholds unchanged, performance variations may be attributed to the induced degradation.

However, simulated corruption still is the drawback of the approach. Artificial blur or noise will not duplicate the influence of optics, atmosphere, compression, and other sensor processes experienced during real operations. To make environmental robustness more relevant to the operation conditions, the framework needs to be extended so that it can include evidence degraded by nature, i.e., gathered under low illumination, real motion blur, bad weather, altitude, and vibration.

Moreover, a promising direction to explore is the automatic characterization of image quality. Instead of conducting experiments in predefined conditions, a developed system may extract blur, illumination, contrast, noise, and compression attributes directly from the image and compare them with model performance profiles obtained during validation. This method will not validate the evidence itself but will give an explainable indicator of decreased completeness of detections in the current regime.

2.9 Digital Evidence Integrity, Provenance and Auditability

The forensic procedure relies on the demonstrability of evidence handling. NIST FIPS 180-4 defines secure hash algorithms, which include SHA-256, that can be used to detect changes to the data made after the moment of generation of the corresponding message digest [25]. ISO/IEC

27037 gives guidelines for digital-evidence identification, collection, acquisition, and preservation [26]. The SWGDE drone-forensics guidance similarly applies these principles to the UAV systems, focusing on the proper preservation, documentation, and validation of examinations [7].

The hashing process is indispensable but often overinterpreted. The coincidence of stored and newly generated digests is proof of byte-level consistency with a specific state. It proves neither the accuracy of the content at the time of capture, the security of the sensor, the integrity of the timestamp nor the completeness of AI interpretation. In contrast, the semantic difference does not prove that the file was altered post-acquisition. Therefore, TDEF considers cryptographic integrity as a dimension of trust but not as an absolute verification tool.

The provenance concept extends the integrity requirement by providing the documentation of the relations between the source and the derivative artefacts. The AI-supported work process can produce frames, annotated images, detection reports, tracking records, degraded artefacts, telemetry analysis, and integrated analysis. Each transformation adds new analytical value, but each of them may cause detachment of the derivative from the source. Therefore, a provenance record should contain the identifiers, hashes, frame numbers, timestamps, and other processing parameters, as well as relations between the artefacts.

It becomes particularly important in the AI case, because the same source can give different outcomes depending on the model version, the threshold settings, the libraries or pre-processing pipelines applied. Thus, the evidentiary object consists not only of the original artefact but of the whole analytical context. For this reason, the reproducibility requires recording of enough information on how the output was obtained. It does not mean that the source code must always be treated as the evidence, but enough information about the model and conditions of processing must be recorded.

A chain-of-custody analysis examines how such evidence is handled along its lifecycle. Nath et al. classify the modern digital-evidence chain-of-custody approaches and stress the importance of integrity, authenticity, and chronological traceability [27]. The blockchain-based designs, like B-CoC, try to provide tamper-proof evidence of ownership and handling of evidence [28]. These approaches show that technological solutions may help to increase auditability, but immutability must not be confused with truthfulness. Immutable record of incorrect metadata will still be incorrect, and the blockchain will not prove that the AI model did not detect the object.

For the current case, the auditable chain-of-custody is considered better than the blockchain infrastructure. Some evidentiary-processing events can be registered using timestamps, identifiers, status, and metadata specific to the experiment, while Git will independently track the development history of the software. This differentiation needs to be stressed, because the version control system is responsible for the development history of the code, and the evidence

audit log should document the evolution of the evidentiary or analytical artefacts. Such confusion would leave a gap in the reconstruction of the chain of processes.

Auditable practice includes human interaction with the evidence too. If the framework suggests performing additional examination due to some limitations or degradation of the artefact in some dimension, the action of the investigator becomes a part of the traceable process. For this reason, TDEF treats reviewer actions as a separate, traceable event, instead of silently substituting automated examination by the human action.

2.10 Trustworthy AI, Explainability and Human Oversight

The use of artificial intelligence (AI) in forensics falls into the wider discussion of the trustworthy nature of automated processes. The NIST AI Risk Management Framework suggests several characteristics that make automated processes trustworthy including validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy enhancement, and fairness. Importantly, the framework does not suggest a universal way of measuring trustworthiness through one numeric measure since trustworthiness depends on many factors.

The multidimensional character of the definition of trustworthiness mentioned above is fully applicable to the problem being discussed. The cryptographic integrity will not help when the recall rate is low. High temporal continuity will not solve the problem of inconsistencies in timestamps. The provenance will not prove that the GPS data point recorded was true when the record was created. Robustness in environmental conditions cannot prove chain of custody. While these properties are related and contribute to the overall reliability of the data, they cannot be compared and balanced in such a way that some of them may compensate others mathematically.

This is the reason why the last version of the TDEF framework does not include the concept of Evidence Trust Score proposed before. The calculation of the score based on weighted aggregated measures may give a very precise score but hide the problems with the evidence base. For example, several verified integrity measures can balance the negative influence of poor detection completeness and generate the score that is moderate, but the investigator will need information that there is a possibility of a significant part of the scene being missing from the interpretation. A more reasonable way of handling the situation is keeping these dimensions separated and informing which of them require additional investigation.

Explainability in the framework discussed goes beyond the explanation of the prediction made by the neural network. In the current situation, the investigator will need an explanation of why the framework generated such disposition. The trust profile may show that the source is cryptographically intact, and its provenance is checked but the AI-based interpretation of it is

limited because of its incompleteness or lack of robustness. Such an explanation is immediately actionable for the investigator who can decide to investigate the source manually.

Human supervision in this case goes beyond the ceremonial approval of the findings. The investigator should have the opportunity to confirm, reject the findings or declare that the case is inconclusive. These actions should be tracked separately from the automated status. It preserves the difference between the decision support and decision-making system and follows the governance concerns regarding the use of surveillance technology.

It means that the framework should not make any legal claims. TDEF can report on the status of evidence integrity, limitations of the analysis performed, and consistencies. It should not evaluate admissibility, guilt or intent. These decisions are made on the grounds of different legal standards and professional expertise. The literature provides a support for the architectural approach according to which transparency and traceability are the parts of technical quality of the process.

2.11 Critical Synthesis and Comparative Gap Analysis

The literature makes significant progress; however, its findings are fragmented by various research traditions. Investigations of drone-forensics highlight what artefacts are available and how readiness can be built [3]-[8]. Studies on UAV security reveal attacks and integrity threats against navigation and communication [9]-[13]. Contributions in computer vision improve object detection and tracking [14]-[20]. Research on anomaly detection detects anomalous visual or multimodal behavior [21], [22]. Findings on robustness explore the impact of adverse visual conditions on performance [23], [24]. Scholarship on digital evidence defines principles for preservation and chain of custody [25]-[28], while trustworthy AI recommendations emphasize transparent, accountable and context-based evaluations [29].

The question is not about immaturity of any of these domains but rather fragmented architecture. Detector-centric study assumes that an original image is inherently reliable; GPS spoofing study assumes that visual data stream is not important; forensic investigation assumes that an artifact is saved but not validated on the ground of reliability of AI-based inference; chain of custody approach assumes recording handling without completeness assessment. An investigator needs to consider all these aspects at once.

Eight gaps can be identified based on the synthesis:

G1 — Fragmentation of forensic workflow. The existing research usually considers acquisition, visual analytics, temporal tracking, telemetry security, robustness and digital evidence

preservation separately. There is little empirical work explaining how such outputs can be combined without losing their evidential relevance.

G2 — Conflation of AI performance with trust. Computer vision research is correct when optimizing performance metrics. However, high performance or high confidence usually becomes the sole criterion of the system quality. To make an AI-based solution trusted, additional information should be considered: provenance, integrity, completeness, robustness and context-based assessment.

G3 — Lack of assessment of evidential completeness. False negatives are particularly important in aerial imagery since missing objects disappear from the automated scene summary. Average mean average precision (map) or precision do not guarantee that non-detections are reliably made.

G4 — Insufficient consideration of environmental robustness. The literature on adverse conditions shows that the quality of images affects performance of the model, however, robustness is not considered as an evidential factor in the process of digital forensics.

G5 — Insufficient assessment of temporal trust. Research on tracking evaluates performance of association but the forensic process usually fails to evaluate temporal continuity separately and explain the difference between the algorithmic identification of tracks and the identification of physical entities.

G6 — Insufficient cross-validation of cyber-forensics. UAV security research examines spoofing and anomalies, but comparatively little attention is paid to the combination of cryptographic verification with timestamps and cross-checking trajectories.

G7 — Insufficient provenance and audit support. Standards of forensic investigation emphasize the importance of preservation and maintenance of the chain of custody, but the AI-driven investigation requires provenance that includes not only source files but also derived images, model outputs, experimental configurations and reviewer decisions.

G8 — Lack of explainable multidimensional integration. The literature suggests plenty of metrics but there is no evidential framework supporting an investigator in considering these metrics without hiding uncertainties behind an opaque or arbitrarily weighted trust score.

Table 2.1 summarizes representative studies against these gaps and identifies how the present research uses or extends their contribution.

Representative study/domain	Contributions	Limitation for trustworthy drone-forensic evidence	Gap(s)
Mantas and Patsakis [3]	Identifies challenges related to operations, procedures, manufacturers, and legality in drone digital-forensic investigations.	Describes the investigation issue but does not incorporate reliability, provenance, robustness, and cyber-forensic consistency of AI within a single workflow evaluation.	G1, G7
Alotaibi et al. [4]	Proposes proactive and reactive readiness in drone forensics.	Readiness increases the availability of evidence but does not verify the reliability of AI-generated analysis and environmental reliability.	G1, G7
Azhar et al. [5]; Taylor [6]	Shows that various types of artefacts (e.g., flight logs, media, device artefacts, and operator information) can be recovered from actual drone systems.	Acquisition studies on specific platforms and devices; lacks AI-assisted evidence quality and multi-source trust assessment.	G1, G7
SWGDE [7]; NIST CFReDS [8]	Provides professional guidance and reference data for drone evidence acquisition, examination, and tools testing.	Well-established forensic baseline but not an end-to-end benchmark of aerial AI, telemetry, robustness, and investigator review.	G1, G7

Sihag et al. [9]	Synthesises cybersecurity and forensic threats for next-generation UAV systems.	Threats are well-detailed, but no model of evidence trust integration is provided.	G1, G6
Shafique et al. [10]; Altaweel et al. [11]	Develops and reviews GPS-spoofing detection approaches, including machine learning.	Performance of the attack detection is condition-dependent and rarely linked with file integrity, visual evidence or chain of custody.	G6
Shin et al. [12]; Pardhasaradhi and Cenkeramaddi [13]	Uses either trajectory behaviour or radar estimates to evaluate the location integrity.	Increases the confidence in location but does not ensure visual-file integrity, AI completeness, and provenance.	G6, G7
Zhu et al. [14]	Provides VisDrone benchmark for detection and tracking of aerial objects in varied scenes.	Maximizes computer vision performance but ignores evidential provenance, telemetry integrity, and forensic trust aspects.	G2, G3
Du et al. [15]	Provides UAVDT benchmark with varying altitude, weather, view, object density and occlusions.	Probes the real-life vision challenges but remains benchmark-focused without consideration of forensic and provenance aspects.	G2, G4, G5
Wu et al. [16]; An et al. [17]; Li et al. [18]	Surveys and improves UAV detection under efficiency, small-object, and occlusion constraints.	Improvements of models' detection do not define evidential completeness and investigator dependence on non-detections.	G2, G3
Zhang et al. (ByteTrack) [19]	Reduces trajectory fragmentation by associating lower-scored detections with previously identified tracks.	The identity of tracked object remains algorithmically inferred but not proven.	G5

Aharon et al. (BoT-SORT) [20]	Combines motion, appearance, and camera-motion compensation in order to robustly track multiple objects.	Benchmark-level association does not prove the forensic identity and provenance of the tracked object.	G5
Bozcan and Kayacan [21]; Jin et al. [22]	Uses multimodal and temporal learning to detect abnormal events in aerial surveillance.	Abnormal statistical behaviour does not necessarily indicate malicious or criminal activity.	G8
Fang et al. [23]; Hendrycks and Dietterich [24]	Illustrates sensitivity of vision models to weather and other corruptions.	Robustness results are seldom considered as evidential condition in forensics.	G4
NIST FIPS 180-4 [25]; ISO/IEC 27037 [26]	Provides cryptographic and evidence handling foundations.	However, integrity and preservation do not define semantic correctness, AI completeness, and sensor truthfulness.	G7
Nath et al. [27]; Bonomi et al. [28]	Studies chain of custody models and tamper-proofing techniques for evidence handling.	Chain of custody can record the handling process but not ensure the reliability of AI-generated interpretations.	G7, G8
NIST AI RMF [29]	Articulates trustworthy AI as a condition-dependent mixture of validity, reliability, transparency, explainability, security, and accountability.	The general guidance does not propose an evidence assessment workflow in drones forensics.	G8

Table 2.1. Critical comparison of representative literature against the research gaps addressed by TDE

2.12 From Literature Gaps to the Present TDEF Design

This review shows that TDEF is justified as a five-layer framework that utilizes provenance awareness when assessing evidence from AI-assisted drones. Rather than proposing innovative detection, tracking, or cryptographic methods, TDEF incorporates existing methods in view of the evidential constraints revealed by G1–G8.

Firstly, the layer of Evidence Acquisition and Provenance defines integrity of the evidence source, metadata, cryptographic hashes, and source-derived relationships according to established forensic criteria [7], [25], [26]. Secondly, the layer of AI-Assisted Evidence Analysis evaluates object detection and temporal tracking in terms of performance against ground truth and established tracking methods like ByteTrack and BoT-SORT [14], [19], [20]. Thirdly, the layer of Trust and Robustness Validation explores problems of environmental degradation and cyber-forensic inconsistencies with visual and telemetry evidence [23], [24].

Such outputs are summarized by the fourth layer of Multidimensional TDEF Assessment along eight independent dimensions: Cryptographic Integrity, Evidence Provenance, AI Detection Reliability, Detection Completeness, Temporal Reliability, Environmental Robustness, Cyber-Forensic Integrity, and Auditability. Consistent with the multidimensional conception of trustworthy AI [29], the dimensions are not summed into an artificial weighted score. The fifth and final layer, Investigator Decision Support, reveals limiting factors and situations that require professional assessment.

Therefore, TDEF contributes to the literature gap by integrating and interpreting existing methods rather than innovating any algorithms. Its contribution consists in the link between independently measurable forensic and AI-reliability criteria without losing their separate evidential value.

2.13 Future Research Trajectory

Extant literature and current limitations in the experimental scope imply several possibilities for further development of TDEF. Firstly, a unified multimodal UAV-forensic dataset with video, GPS data, telemetry readings, timestamps, and whenever possible inertial measurements needs to be developed. Existing forensic and vision datasets provide separate but useful types of evidence [8], [14], [15], thus limiting end-to-end evaluation.

Future evaluations should also consider multiple object detectors, longer tracking sequences, and a variety of UAVs to examine if reliability, completeness, and time-related features of TDEF are significant under different AI-based configurations. The environment of evaluation should move from the artificial conditions of synthetic degradation towards natural degradation characterized by poor illumination, real motion blur, weather conditions, compression, and camera instability [15], [23], [24].

Moreover, the cyber-forensic module could also be enhanced by means of a controlled UAV testbed that would allow creating realistic inconsistencies in timing, sensors, and navigation. The testbed would help to evaluate differences between post-capture modifications, sensor failures, and intentional inconsistencies not considering all inconsistencies as signs of attack.

Finally, the larger and more varied datasets may help to calibrate thresholds for continuity and spatial consistency. In this way, the goal of future work would be to transform TDEF from a proof-of-concept solution to a publicly validated and reproducible framework of forensic decision-making assistance, retaining human-in-the-loop evaluation.

2.14 Conclusion

As seen from the existing literature, technologies for drone forensics, UAV cybersecurity, computer vision from UAVs, tracking, environmental robustness, and digital evidence preservation have developed enough technical maturity. The major limitation here is that the listed disciplines are usually considered separately. Thus, parameters such as integrity of evidence, AI reliability, completeness of detection, time coherence, environmental conditions, telemetry consistency, etc., are rarely estimated together.

During this review, eight related research gaps are defined, and the theoretical basis for TDEF as a multidimensional framework of forensics evaluation is presented. TDEF does not introduce any novelty in the development of hashing algorithms, object detection and tracking, or telemetry analysis separately. Its innovation is the use of the listed algorithms in the context of an evidence-provenance-aware workflow when each aspect of trust remains independent, and its restrictions are revealed to the investigator and not hidden in one aggregated metric.

Thus, the literature review directly sets the basis for the methodology that follows and evaluates experimentally how those dimensions can be utilized and combined in an AI-assisted workflow for drone-based evidence.

3. Methodology

3.1 Research Design

The study used a design-and-evaluation approach to devise and test the Trustworthy Drone Evidence Framework (TDEF). The underlying methodology framework rested on an essential problem formulated in the literature: the evidential significance of drone analysis assisted by artificial intelligence could not be assessed based on detector confidence and typical model performance measures alone. Digital evidence may maintain cryptographic integrity while its automated interpretation would be partial, environmentally dependant, temporally unreliable, or inconsistent with accompanying telemetry data.

Thus, evidential trust became an object of study in its multidimensional form with cryptographic integrity, evidence provenance, AI detection reliability, detection completeness, temporal reliability, environmental robustness, cyber-forensic integrity, and auditability as its components. Object detection was considered only as one stage of the forensic analysis process rather than the goal of the research.

Development took an incremental form. Evidence ingestion, provenance, object detection, temporal tracking, ground truth, environmental robustness testing, and cyber-forensics were implemented and tested separately. Their results were then aggregated into TDEF to produce a human-interpretable and explainable assessment. Such design reduced the likelihood of implementation mistakes and made it possible to keep the evidential base for every trust dimension separately.

The framework does not calculate any numerically weighted trust score as an arbitrary combination of separate dimensions. It instead reports the limitations of the digital evidence which require the investigator's attention and the basis for such a position among the eight different dimensions of evidential trust.

3.2 TDEF Framework Design

The proposed TDEF architecture consists of five functional layers: evidence acquisition and provenance, AI-assisted evidence analysis, trust and robustness validation, multidimensional evidence assessment, and investigator decision support.

3.2.1 Layer 1 — Evidence Acquisition and Provenance

The first layer represents the forensic starting point. Source artefacts get consumed, metadata get captured, and SHA-256 hashes are generated before any further analysis takes place. Regarding video data that have been converted to derived frames, each individual frame stays linked to the source through source hash, frame number, time stamp, and frame hash. Selection of SHA-256 is consistent with the Secure Hash Standard [25] and forensic practices in terms of demonstrating whether there has been any change to a digital artefact since its recording [32].

3.2.2 Layer 2 — AI-Assisted Evidence Analysis

The second level analyses visually and temporally. The object detection algorithm was developed based on the pretrained YOLOv1n model using the Ultralytics software package [33]. Detection output includes class prediction, probability of the class, and bounding box position coordinates. Temporal reconstruction was done separately through ByteTrack and BoT-SORT [19], [20]. Data generated by artificial intelligence is considered observations supported by data sources, not separate claims of truth.

3.2.3 Layer 3 — Trust and Robustness Validation

The third-tier analyses elements that may lower the reliance on automated interpretations. Robustness of the environment was analysed using blur, decreased brightness, reduced contrast, and Gaussian noise on the annotated images. Integrity of cyber-forensics was tested through controlled alterations in location, timestamp, telemetry, cross-source temporal alignment, derived frames integrity, and flight path continuity. The situations created were meant for validation of forensically consistent methods as opposed to all the UAV attack scenarios encountered in the real world.

3.2.4 Layer 4 — Multidimensional TDEF Assessment

The fourth layer incorporates the experiment outcomes, while retaining the unique evidentiary importance of each. For this reason, TDEF uses categories of states such as VERIFIED, STRONG, LIMITED, LOW, and DEGRADED and not a weightless score. A positive result in one aspect mathematically cannot counterbalance the weaknesses of the other; for example, verified file integrity does not compensate for low detection recall.

3.2.5 Layer 5 — Investigator Decision Support

The last stage involves communicating the outcome of the integrated assessment to the investigator. In case of limiting conditions, the TDEF generates the message **HUMAN REVIEW REQUIRED** and identifies the variables responsible for the assessment. It does not establish admissibility under the law or indicate that the evidence is fraud; it just means that there are one or more conditions that require professional review.

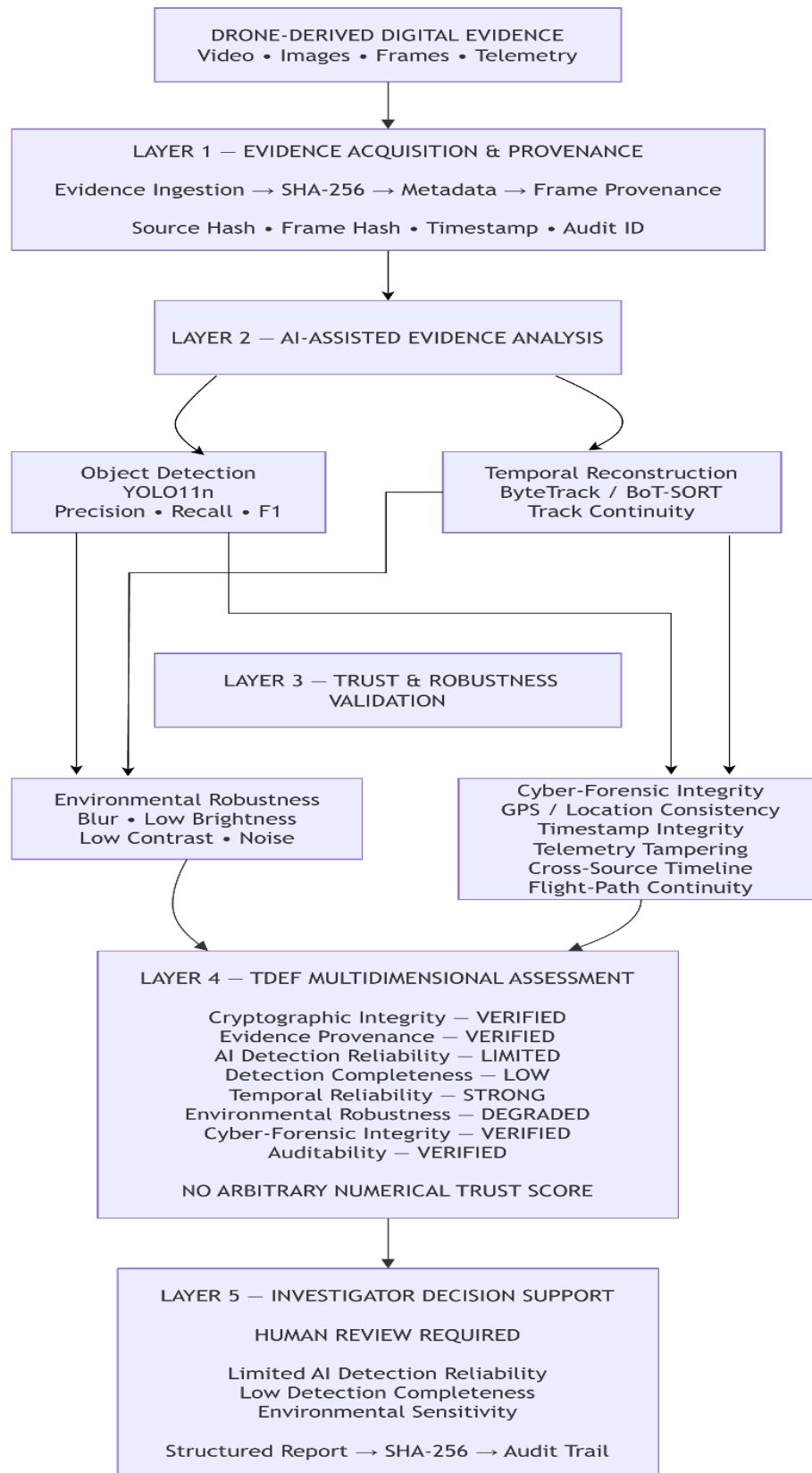


Figure 3.1: Proposed structure of the Trustworthy Drone Evidence Framework (TDEF).

3.3 Experimental Environment and Implementation

The prototype was created on an Apple MacBook Air computer with an M2 chip and 8GB of unified memory running macOS Sonoma 14.5. The development took place in Visual Studio Code and used an isolated Python 3.12 virtual environment. The main software included PyTorch, Ultralytics, and OpenCV, and local hardware acceleration was provided by Apple Metal Performance Shaders where possible.

All the tests were performed in the same local environment to maintain consistent configurations. The source code was kept under Git revision control, while large datasets and results of experiments were kept separate from the version-controlled repositories. The final list of dependencies was kept in the project environment.

3.4 Evidence Acquisition, Integrity and Provenance

3.4.1 Evidence Ingestion

An evidence-ingestion module was specially crafted to provide a repeatable entry point into the digital forensic process.

For every ingested artefact, the framework tracks:

- evidence id;
- filename of the source;
- file type;
- file size;
- SHA-256 digest;
- appropriate media metadata;
- processing timestamp;
- location of the manifest;
- associated audit event.

The cryptographic digest is created before any further downstream artificial intelligence analysis, providing a baseline against which further verification can be accomplished.

Hash of the artefact E is defined as follows:

$$H_E = \text{SHA256}(E),$$

where H_E is saved in the respective evidence manifest. During the next processing phase, the integrity of the artefact is checked through calculating its hash H'_E and evaluating whether

$$H_E = H'_E.$$

Matching of the digests implies that the byte level artefact has not been modified since its creation. This equality, however, cannot prove authenticity or lawful acquisition of the evidence; this difference is preserved in TDEF.

3.4.2 Video-to-Frame Provenance

For temporal analysis, a 10-second-long source video was used, which had 250 frames at 25 FPS and resolution 960 x 540. Ten provenance frames were created from the source video at an interval of one second each. Each artefact contained information about the source video file name, the SHA-256 hash, frame number, timestamp, filename, and frame hash.

The above technique resulted in a provenance relationship of:

Source Video -> Source Hash -> Frame Number -> Timestamp -> Frame Hash

This ensured that the log maintained a direct connection with the generated artefacts and the source evidence.

3.4.3 Audit Logging

These major processing events were described in a JSONL audit log in a well-defined format that included identifiers, Coordinated Universal Time (UTC) timestamps, actions, success or failure of the action performed, and the relevant information pertaining to the experiment. This audit log records evidence processing events, not software development events.

3.5 AI-Assisted Object Detection

For object detection, the YOLOv11n model that had been pretrained was used with the help of the Ultralytics library [33]. The YOLOv11n was selected as a lightweight detector for the local hardware; the paper is not claiming that it is an optimal detector for UAV imagery.

The whole pipeline kept the predicted class, confidence, bounding box coordinates, evidence-source relation, annotation, and the detection report. The confidence threshold of 0.25 was kept for the primary experiments.

The confidence value from the model was not considered as the evidential trust. The confidence value stands for the confidence of the prediction and does not provide any proof of the evidence integrity, completeness of the scene, environment robustness, or lack of undetected objects. These criteria were evaluated in the further experiments separately.

3.6 Ground-Truth Evaluation Using VisDrone2019-DET

3.6.1 Dataset Selection

For ground truth evaluation, VisDrone2019-DET validation dataset was used, comprising 548 annotated aerial images.

VisDrone was selected since it features the annotation for object detection tasks from drone images and serves as a benchmark for visual detection in aerial scenarios [6]. The literature of challenges in object detection on drone images clearly shows that there is much room left for improving aerial detection.

The dataset was saved separately from the source code repository to prevent unmanageable size of the latter and duplication of publicly available datasets.

3.6.2 Annotation Parsing

The VisDrone annotations are expressed using bounding box coordinates, in the form of:

(x, y, w, h) ,

where x, y are the top-left coordinate and w, h is the width and height, respectively, of the bounding box.

The annotations were converted into the format that is more suitable for detectors, as follows:

$(x1, y1, x2, y2)$,

such that:

$$x2 = x1 + w, y2 = y1 + h.$$

Moreover, a custom annotation parser was built that could filter out annotation information as well as remove the hidden Apple Double files created due to macOS external storage use.

3.6.3 Class Mapping

There is no congruence between the taxonomies of VisDrone and COCO/YOLO. Thus, the direct assessment was limited to those classes where there was evidence of a meaningful mapping.

The mapping was done using the following scheme:

VisDrone class → YOLO/COCO class

pedestrian → person

people → person

bicycle → bicycle

car → car

truck → truck

bus → bus

motor → motorcycle

Such categories as van, tricycle, awning-tricycle, and others were not considered because there are no appropriate COCO counterparts.

In such a way, we avoid increased errors caused by incongruences in the taxonomies.

3.6.4 Intersection-over-Union Matching

The performance was evaluated through class-specific one-to-one Intersection-over-Union (IoU) matching.

IoU was defined as follows:

$$\text{IoU} = |\text{Bp} \cap \text{Bg}| / |\text{Bp} \cup \text{Bg}|$$

where Bp represents the predicted bounding box, and Bg represents the ground truth bounding box.

A detection was a true positive if:

the predicted class was equal to the ground truth class;

IoU was greater than or equal to 0.50; and

the ground truth object has not been matched yet.

The predictions were ordered by decreasing confidence and matched greedily one-to-one.

Unmatched predictions were false positives, while unmatched ground truth objects were false negatives.

3.6.5 Precision, Recall and F1

The most important evaluation measures used are as follows:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

Precision is a measure that describes the ratio of the detections made by the algorithm that are correct, based on the chosen matching criteria, while recall is a measure that describes the ratio of the existing objects in the scene that are detected by the model.

It is especially important to understand the difference between precision and recall in terms of the forensic analysis, as a model can have very high precision at the same time when failing to detect many existing objects.

3.6.6 Ignored-Region Handling

VisDrone includes regions classified as being ignored, which are not supposed to affect the evaluation process of the standard object detection.

Therefore, the special evaluator ignores ground truth or prediction boxes in which more than 50% of the candidate box area overlaps with the ignored regions.

This overlap ratio can be formulated as:

$$\text{R_ignore} = \text{Area}(\text{B_c} \cap \text{R_ignored}) / \text{Area}(\text{B_c})$$

where B_c represents the candidate bounding box.

If there are several ignored regions that overlapped one candidate box, then the area of the union of the overlapping rectangles was considered to avoid double counting.

This algorithm intends to avoid any form of unfair consideration of predictions that are mainly affected by the ignored regions. It is important to note that the current version of the software does not intend to mimic the official VisDrone evaluation software.

3.6.7 Progressive Validation

The initial runs of the evaluator were conducted for 25 reproducibly chosen images with random seed 42. Later, the evaluation process was scaled up to include 100 images and then all 548 images in the validation set.

Experiments on 25 and 100 images mostly had a verification character. The main quantitative results of the experiment are obtained from the entire validation set.

This strategy helped to avoid carrying any evaluator mistakes to the full-scale experiment.

3.7 Temporal Reconstruction and Tracking Reliability

Temporal reconstruction used the same 250-frame video as the source data for both ByteTrack and BoT-SORT [4], [5] with a similar YOLO11n detector, confidence threshold, and conditions of processing. In the context of each track, the algorithm preserved its algorithmic identifier, class, start and end of observations, time stamps, number of observation frames, and average confidence.

Continuity of the track is expressed through the Track Observation Ratio (TOR):

$$\text{TOR} = N_{\text{observed}} / N_{\text{possible}}$$

where N_{observed} is the number of observation frames and N_{possible} is the number of frames between the first and the last frame of the track. High continuity was achieved when $\text{TOR} \geq 90\%$, and fragmentation was reached when $\text{TOR} < 50\%$. Those values represent the results of the experiment and do not correspond to any forensic standard.

Tracker comparison includes number of algorithmic tracks, tracking history, mean and median TOR, high-continuity tracks, fragmented tracks, mean detection confidence, and processing time. Algorithmic identifiers must be considered only as algorithmic associations and not the identification of physical objects.

3.8 Environmental Robustness Evaluation

Environmental robustness was tested for the entire dataset of 548 images from the VisDrone validation set under five conditions: the original image, Gaussian blur, low brightness, low contrast, and Gaussian noise. The parameters used for environmental degradations included 15x15 Gaussian blur filter size, brightness factor 0.40, contrast factor 0.40, and Gaussian noise $\sigma=35$. All random operations were set to seed 42.

The detector, confidence threshold, IoU threshold, annotations, class mapping, and ignored region procedure remained the same for all conditions. Thus, the performance of each scene with 548 images was measured under five different conditions, resulting in 2,740 runs of controlled inference.

The performance changes were measured through precision, recall, F1, and the difference between F1 scores in the degraded and the original conditions:

$$\Delta F1_{pp} = (F1_{degraded} - F1_{original}) \times 100$$

Due to the paired design, the performance changes can be attributed only to visual degradation and not to differences in the image content.

3.9 Cyber-Forensic Integrity Validation

3.9.1 Experimental Baseline and Manipulations Design

Cyber-forensics validation evaluated the ability of TDEF to detect controlled violations of integrity and consistency. The experiment did not reproduce actual attacks on UAVs; however, experimental replicas of evidentiary artefacts were manipulated to test whether the framework generated the required forensic evidence.

The clean baseline consisted of the 10-second video and frame provenance log combined with the synthetic 10-second telemetry stream within the same 0-9 second timeframe. The video and telemetry streams were collected separately and only merged to conduct controlled cross-source testing. Baseline validation included video and frame hashes, monotonic timestamps, telemetry hash consistency, and cross-source temporal synchronization.

Six manipulation scenarios were based on the replicas of the baseline and included location manipulation, timestamp manipulation, telemetry value manipulation, 5-second video-telemetry timeline offset, derived frame manipulation, and flight path inconsistency. The original artefacts were not altered.

3.9.2 Independent Validation

In its turn, the validator checked telemetry hashes, timestamp sequence, cross-source timelines, spatial continuity, trajectory anomalies, and frame hashes independently from the scenarios labelling to find out which indicator needs to be produced. Scenarios labels were used only for comparison of the results obtained with the expected ones.

Spatial continuity was estimated based on the Haversine distance between the two consecutive coordinates normalized by the period:

$$v_derived = d_haversine / \Delta t$$

Threshold value 35 m/s was used to detect sudden changes in coordinates. It serves only as a practical boundary of decision and not as a general UAV/forensic standard. Similarly, any location or trajectory anomalies were used just as the forensic indicators and were not identified as the result of a GPS spoofing attack.

3.10 Multidimensional TDEF Integration

After the execution of all individual experiments, their recorded outputs were consolidated via the TDEF integration module. Inputs included cryptographic and provenance validation, VisDrone ground truth analysis, temporal tracking and tracker evaluation, environmental degradation analysis, cyber-forensic validation, and structured audit analysis.

Eight independently interpretable dimensions were obtained as output from the module: Cryptographic Integrity, Evidence Provenance, AI Detection Integrity, Detection Integrity, Temporal Integrity, Environmental Robustness, Cyber-Forensic Integrity, and Auditability.

Dimension states were obtained based on the experimental evidence available for the dimension in question and not some aggregate numeric value. The cryptographic integrity and evidence provenance dimensions were based on hash and provenance validation, AI detection integrity and completeness were based on ground-truth detection, temporal integrity was based on temporal tracking, environmental robustness was based on performance analysis under different environmental degradations, cyber-forensic integrity was based on clean baseline and consistency tests, and auditability was based on existence of processing records.

Dimensions with limited, low, degraded, or any other concerning conditions were kept as explicit triggers for reviews.

3.11 Reproducibility and Experimental Traceability

The consistency of the results was achieved by establishing deterministic experimental conditions, preserving the configuration information, keeping logs and using versioning. A fixed seed (Seed 42) was used when a reproducible sample or noise was needed. The configuration information for models, confidence and IoU threshold values, and degradations were stored inside experimental result files. Structured logs were created by major processing steps, and the source code development was conducted using Git. The principal reports were additionally secured by applying SHA-256 hashing after they were generated.

3.12 Ethical and Safety Considerations

No humans were included as experimental subjects in the research.

The VisDrone dataset was used as a known public benchmark for the research. Human presence in aerial imagery was regarded only as one of the object categories; no identification by name or face recognition was performed on any individual present in the images.

Artificial telemetry was used in conducting the cyber-forensic experiments. No GPS spoofing transmitters or any other live attack vectors against working UAVs' systems were used.

Manipulations were limited to working with copies of experimental data and original artefacts were left intact.

In such a way, potential threats were minimized but the research could be performed.

Therefore, the cyber-forensic experiments can be regarded only as verification of the evidence consistency mechanisms, not as operational penetration testing.

4. Results

4.1 Experimental Overview

The following results show the experimentally evaluated properties of TDEF in terms of evidence integrity and provenance, object detection with assistance from AI, temporal reconstruction, robustness to the environment, cyber forensics consistency, and multi-dimensional integration. The main ground truth assessment was performed using all 548 images from the VisDrone2019-DET validation dataset [34]. In turn, the temporal analysis was conducted using the 250 frames of the source video. The validation in terms of cyber forensics included one baseline and six manipulations.

Experimental module	Data / evidence	Major output
Evidence integrity and authenticity	10 seconds video and 10 frames of provenance	SHA-256 and source-to-derived evidence validation
Object-detection test	548 VisDrone2019-DET images	Precision, recall, F1 score, TP, FP, FN
Per class detection test	Six classes of objects	Class-wise precision, recall, F1
Temporal reconstruction	250 frames of video	Tracking continuity and fragmentation
Trackers comparison	ByteTrack and BoT-SORT	Mean/median continuity, number of tracks, fragmentation
Environment robustness	548 images in five settings	Changes in precision, recall and F1
Cyber-forensic test	1 clean baseline case + 6 controlled manipulation scenarios	Integrity and consistency measures
TDEF implementation	Eight dimensions of evidence	Explained trust profile and investigator's review disposition

Table 4.1. Experimental Evaluation: Summary

4.2 Evidence Integrity and Provenance Validation

The video used was made up of 250 frames shot at 25 fps with 960 x 540 resolution. Ten frames of provenance were taken at one-second intervals and verified using their respective SHA-256 values. The digest for the present source video matched with the provenance data; all the ten hashes for the derived frames were still valid, there were no missing provenance frames, and the frame timestamps had monotonically increased values. Therefore, Cryptographic Integrity and Evidence Provenance were marked as **VERIFIED**.

Validation criterion	Result
Source video frames	250
Frame rate	25 FPS
Resolution	960 × 540
Provenance frames	10
Source-video SHA-256 match	Verified
Derived-frame hashes verified	10/10
Missing provenance frames	0
Frame hash mismatch	0
Frame timestamps monotonic	Yes
Cryptographic integrity	VERIFIED
Evidence provenance	VERIFIED

Table 4.2 provides the results for Evidence Integrity and Provenance.

There was no cryptographic discrepancy in the clean evidentiary baseline.

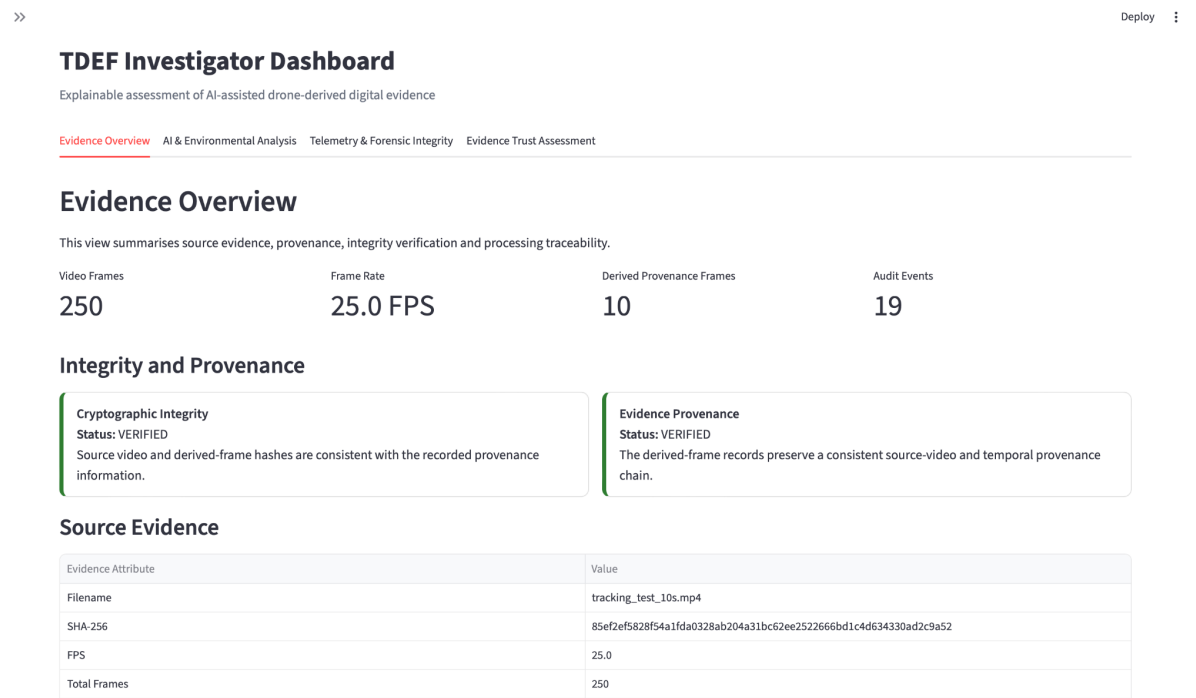


Figure 4.1 — Evidence Provenance Verification

4.3 Ground-Truth Object-Detection Performance

The main object detection experiment involved testing the YOLO11n detector on the full VisDrone2019-DET validation set, including 548 images [34]. The evaluation was performed only on the classes discussed in Section 4 using class-specific one-to-one matching with the IoU criterion of 0.50.

Overall, the total number of comparable ground-truth objects included 35,057. The detector yielded 7,871 predictions, 5,585 of which were marked as true positives. 2,286 predictions were not matched and were considered as false positives. There were also 29,472 false negatives found during evaluation.

Metric	Result
Images evaluated	548

Comparable ground-truth objects	35,057
Predictions	7,871
True positives	5,585
False positives	2,286
False negatives	29,472
Precision	70.96%
Recall	15.93%
F1-score	26.02%
Processing time	18.27 s

Table 4.3. Full VisDrone2019-DET Evaluation Results

The overall statistics show a significant difference between precision and recall values: precision is 70.96% while recall is 15.93%, and the F1-score was 26.02%.

Thus, most of the predictions accepted by the matching procedure are related to real ground-truth objects, however, many more annotated objects were not detected.

This difference is retained in the TDEF as detection reliability and detection completeness are two distinct concepts instead of considering detector precision or confidence as performance indicators.



Figure 4.2 — Overall Detection Performance

4.4 Per-Class Detection Performance

There were huge differences between the performance for each of the six object classes considered.

Class	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)
Bicycle	2	4	1,276	33.33	0.16	0.31
Bus	40	61	207	39.60	16.19	22.99
Car	4,439	1,602	9,541	73.48	31.75	44.34
Motorcycle	60	46	4,822	56.60	1.23	2.41
Person	970	295	12,952	76.68	6.97	12.77
Truck	74	278	674	21.02	9.89	13.45

Table 4.4. Performance by Class on the VisDrone Detection Task

Cars had the best F1 score at 44.34% as well as the best recall at 31.75%. The category person had the best precision rate at 76.68%, although its recall rate was relatively poor at 6.97%.

There were significant performance limitations when it came to the bicycles and motorcycles, whose recall rates were 0.16% and 1.23%, respectively. These numbers suggest that there was significant variability across different classes in relation to the overall aggregate scores.

An illustrative case is that of the category person, whose precision rate of 76.68% existed alongside 12,952 false negatives and a very poor recall rate of 6.97%.

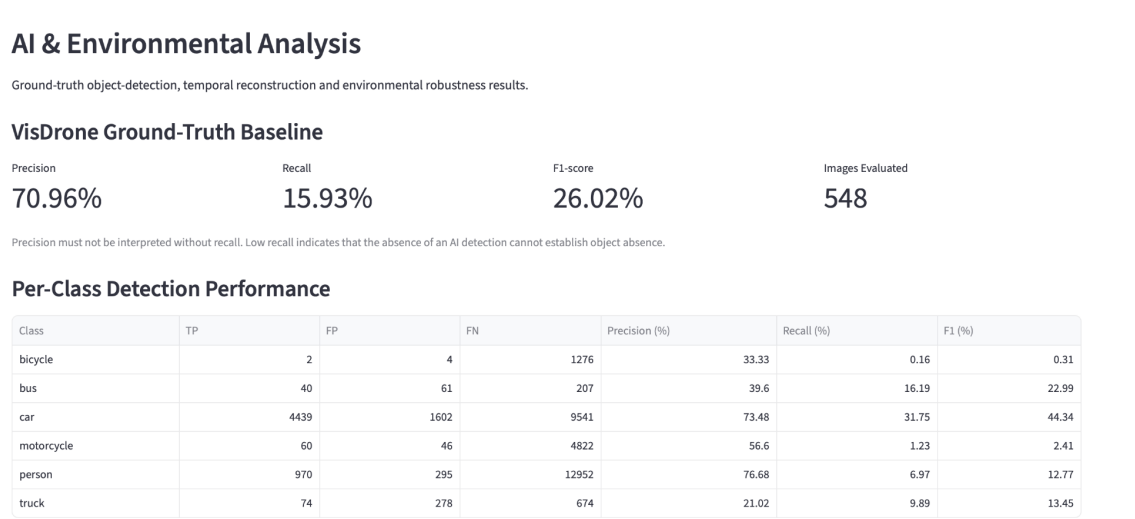


Figure 4.3 — Per-Class Detection (Precision, Recall and F1)

4.5 Temporal Reconstruction and Tracker Comparison

The evaluation of temporal reconstruction was carried out with the help of ByteTrack [19] and BoT-SORT [20] within a certain configuration. Both trackers processed an identical 250-frame video with the use of the same YOLO11n detections with a confidence threshold of 0.25.

Each configuration provided 2,229 frame-level tracking records.

Metric	ByteTrack	BoT-SORT
Video frames	250	250
Tracking records	2,229	2,229
Unique algorithmic tracks	78	75
Mean track continuity	91.78%	92.54%
Median track continuity	100%	100%
High-continuity tracks ($\geq 90\%$)	62	58
Fragmented tracks ($< 50\%$)	1	0
Mean detection confidence	0.559	0.559
Processing time	8.685 s	7.650 s

Table 4.5. The results of ByteTrack and BoT-SORT Temporal Reconstruction

On the other hand, ByteTrack was able to produce 62 tracks classified as having $\geq 90\%$ of high continuity in comparison with 58 by BoT-SORT.

Median continuity of both trackers is 100%.

From the table above, it is seen that both tracking configurations provided high values of temporal continuity on the tested sequence though the distributions of tracks were different.

It is essential to stress the absence of any claims regarding the superiority of one tracker configuration over the other one due to the fact.

Temporal Reconstruction

Tracker	Unique Tracks	Mean Continuity (%)	Median Continuity (%)	High-Continuity Tracks	Fragmented Tracks
ByteTrack	78	91.78	100	62	1
BoT-SORT	75	92.53	100	58	0

Track identifiers represent algorithmic associations and must not be interpreted as verified physical identity.

Figure 4.4 — Tracker Continuity Comparison

4.6 Environmental Robustness Results

The environmental robustness was evaluated using all validation images for five visual conditions: original image, Gaussian blur, decreased brightness, decreased contrast, and Gaussian noise.

As the same set of images, annotation, detection algorithm, and matching rules were used for all visual conditions, the differences in the results obtained are due to the changes caused relative to the original image.

Condition	TP	FP	FN	Precision (%)	Recall (%)	F1 (%)	ΔF1 (pp)
Original	5,585	2,286	29,472	70.96	15.93	26.02	0.00
Blur	4,313	1,738	30,744	71.28	12.30	20.98	−5.04
Low brightness	5,076	1,875	29,981	73.03	14.48	24.17	−1.85
Low contrast	4,526	1,738	30,531	72.25	12.91	21.91	−4.11
Noise	2,900	992	32,157	74.51	8.27	14.89	−11.13

Table 4.6. Detection Performance for Controlled Visual Degradation Conditions

Degradation in all four cases led to lower F1 compared to that in original images. The least decrease was recorded at low brightness (-1.85%), and the greatest decrease was in the case of Gaussian noise, where F1 went down from 26.02% to 14.89% (-11.13%).

An interesting trend could be observed when noise occurred: Precision increased from 70.96% to 74.51%, but recall decreased from 15.93% to 8.27%. Thus, an apparent increase in precision was observed with a significant drop in completeness of detection.

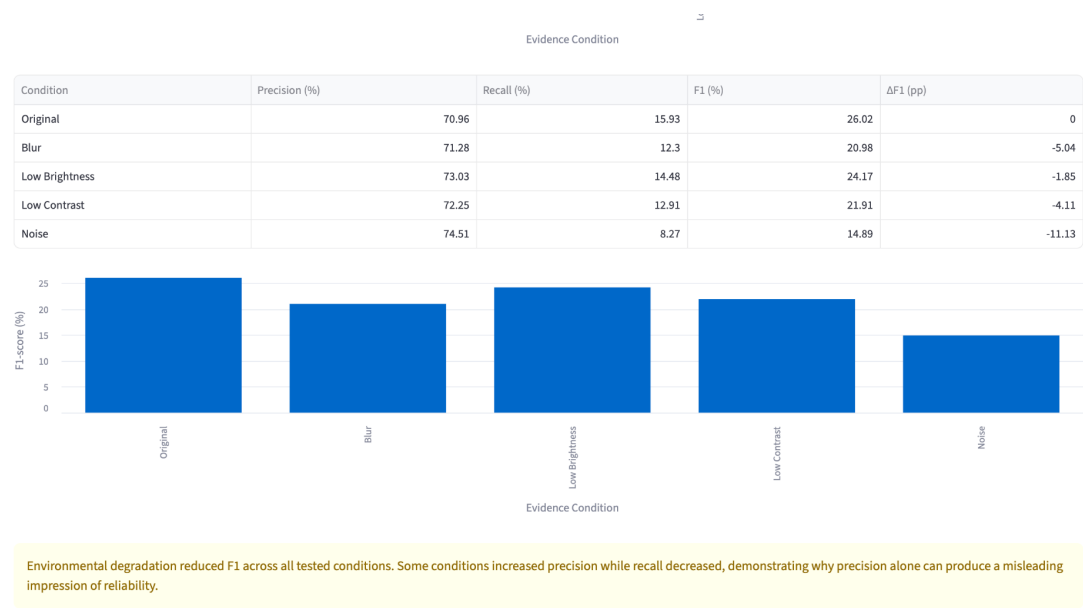


Figure 4.5 — Environmental Robustness

4.7 Cyber-Forensic Integrity Validation

The controlled cyber-forensic experiment started with baseline verification before applying the controlled manipulations on evidence.

The baseline verification revealed that the current video hash was consistent with the recorded value, the frame log relationship was still valid, and both the visual and telemetry timelines covered the range from 0 to 9 seconds. Therefore, the clean case was considered:

VERIFIED

After that, six different controlled manipulations were validated by the independent validator.

Scenario	Detected forensic indicator(s)	Expected target detected
GPS manipulation	Telemetry hash mismatch; location-integrity anomaly	Yes
Timestamp manipulation	Telemetry hash mismatch; timestamp-sequence anomaly	Yes
Telemetry log manipulation	Telemetry hash mismatch	Yes
Time mismatch between video & telemetry data	Telemetry hash mismatch; cross-source timeline mismatch	Yes
Derived frame manipulation	Frame hash mismatch	Yes
Flight path discontinuity	Telemetry hash mismatch; location-integrity anomaly; trajectory discontinuity	Yes

Table 4.7. Results of Controlled Cyber-Forensic Validation

The validator successfully provided the expected forensic indicator in all six controlled cases:

6/6

or 100% of controlled cases.

This result should be treated only as the controlled function validation and should not be considered the real world cyberattack detection rate of 100%.

The clean control was found to be verified, and each of the manipulated cases contained the expected consistency or integrity indicator.

In addition, the results show that multiple manipulations produced two types of the evidence at once. In particular, the GPS manipulation influenced the telemetry digest and caused the location-integrity anomaly. Similarly, in the timestamp manipulation scenario there was both the cryptographic mismatch and the timestamp-sequence anomaly.

Telemetry & Forensic Integrity

Controlled cyber-forensic validation of evidence integrity and consistency.

Baseline Status

VERIFIED

Controlled Scenarios

6

Expected Indicators Detected

6/6

The six scenarios are controlled functional validations. The result must not be interpreted as 100% real-world cyberattack detection accuracy.

Controlled Manipulation Results

Scenario	Artifact Type	Forensic Flags	Human Review
Gps Spoofing Simulation	Telemetry Evidence	Telemetry Hash Mismatch, Location Integrity Anomaly	Yes
Timestamp Manipulation	Telemetry Evidence	Telemetry Hash Mismatch, Timestamp Sequence Anomaly	Yes
Telemetry Log Tampering	Telemetry Evidence	Telemetry Hash Mismatch	Yes
Video Telemetry Time Mismatch	Telemetry Evidence	Telemetry Hash Mismatch, Cross Source Timeline Mismatch	Yes
Evidence Hash Mismatch	Derived Visual Evidence	Frame Hash Mismatch	Yes
Flight Path Discontinuity	Telemetry Evidence	Telemetry Hash Mismatch, Location Integrity Anomaly, Trajectory Discontinuity	Yes

Figure 4.6- Generated forensic indicators for six controlled evidence manipulation scenarios.

4.8 Integrated TDEF Evidence Assessment

After completing all individual tests, the results were consolidated via the TDEF assessment module.

A total of eight distinct dimensions were considered.

TDEF dimension	Final status	Major evidence
Cryptographic Integrity	VERIFIED	Source frame hashes did not change
Evidence Provenance	VERIFIED	Source-to-frame provenance chain and timestamp order maintained
AI Detection Reliability	LIMITED	Precision 70.96%; F1 26.02%

Detection Completeness	LOW	Recall 15.93%; 29,472 false negatives
Temporal Reliability	STRONG	Average tracker persistence >91% across both configurations
Environmental Robustness	DEGRADED	All degradation types decreased F1
Cyber-Forensic Integrity	VERIFIED	Clean baseline verified; 6/6 indicators detected successfully
Auditability	VERIFIED	Required analysis steps supported by successful structured audit events

Table 4.8. Integrated TDEF Trust Profile

An integrated assessment did not provide an aggregate numerical trust score.

Instead, the framework retained the individual evidential states and checked whether there were any dimensions requiring the involvement of investigators.

Three such dimensions were determined: AI Detection Reliability, Detection Completeness, and Environmental Robustness.

Thus, the final system disposition was:

HUMAN REVIEW REQUIRED

In no way was this an indication of the lack of integrity in evidence. Cryptographic integrity, provenance, cyber-forensic integrity, and auditability were still validated. Instead, the determination highlighted the problems associated with the integrity of the AI-based interpretation.

Evidence Trust Assessment

TDEF retains individual forensic trust dimensions rather than collapsing them into an arbitrary numerical score.

Overall TDEF disposition: **HUMAN REVIEW REQUIRED**

Multidimensional Trust Profile

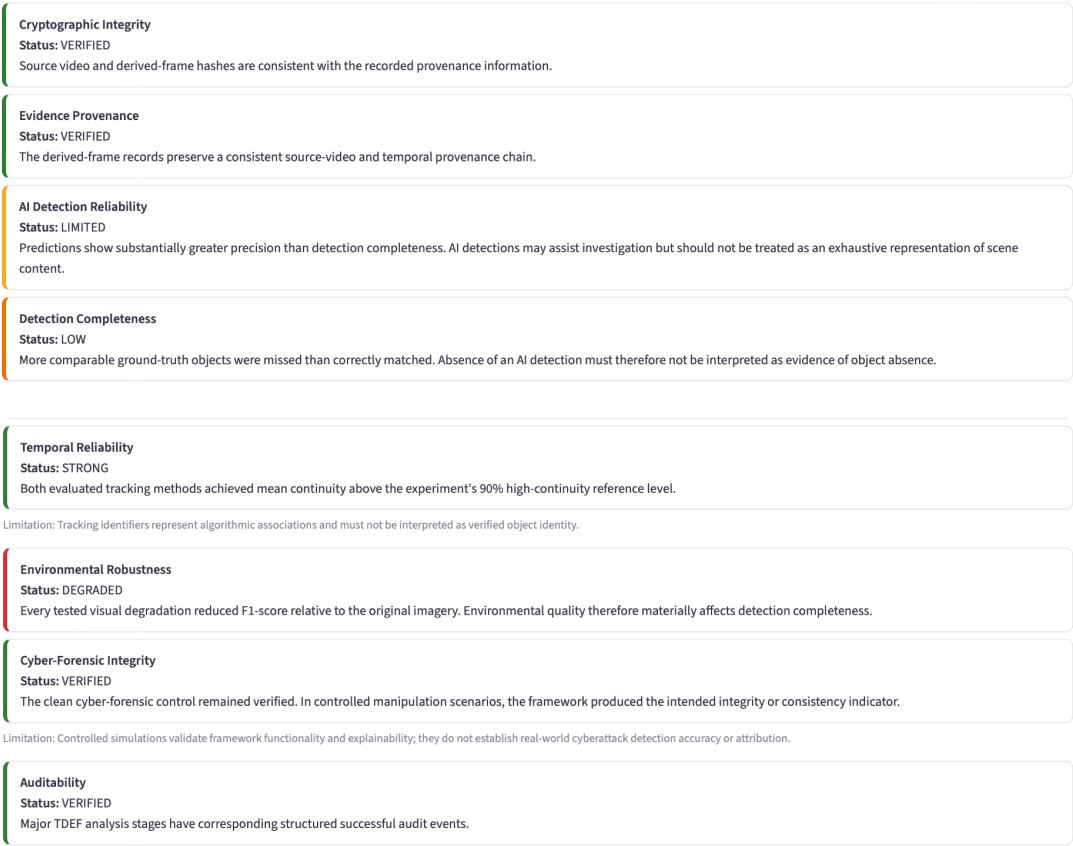


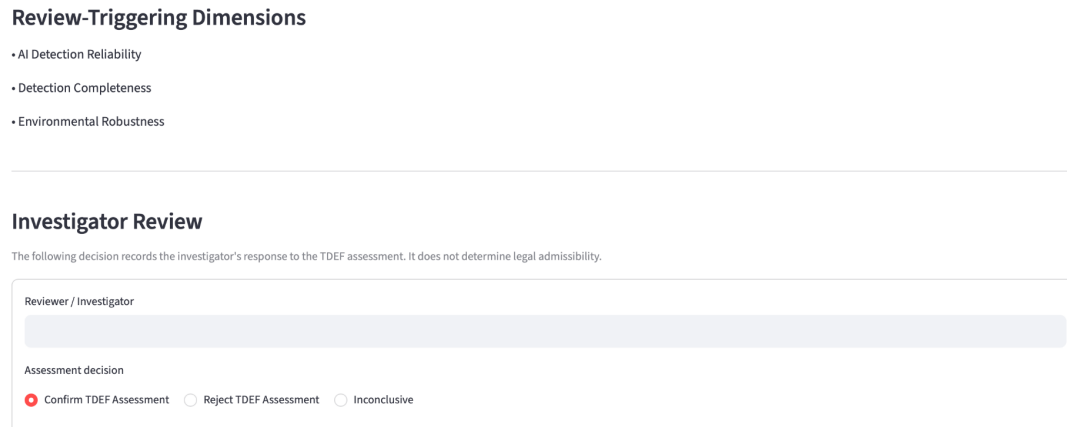
Figure 4.7 — TDEF Multidimensional Trust Profile

4.9 Investigator-Facing Prototype

The integrated assessment approach was realized through an investigator dashboard using Streamlit for the Evidence Overview, AI and Environment Assessment, Telemetry and Forensic Integrity, and Evidence Trust Assessment. The interface provides access to the evidence provenance, output of AI, environmental and cyber-forensic results, and therefore the eight-dimensional trust profile generated.

When human assessment is required, the investigator can choose either Confirm TDEF Assessment, Reject TDEF Assessment, or Inconclusive, along with explanations. The assessment details include the assessment conclusion, reviewer’s decision, time stamp, notes, and the dimensions triggering the review and are stored through the auditing process. As a result, the

interface differentiates automated assessment from human judgment and does not show the investigator's decision as the admissibility decision.



The figure shows a web-based interface for investigator review. It is divided into two main sections. The first section, titled "Review-Triggering Dimensions", lists three bullet points: "AI Detection Reliability", "Detection Completeness", and "Environmental Robustness". The second section, titled "Investigator Review", contains a sub-header "The following decision records the investigator's response to the TDEF assessment. It does not determine legal admissibility." Below this is a form with a label "Reviewer / Investigator" and a large text input field. Underneath the input field is the label "Assessment decision" followed by three radio button options: "Confirm TDEF Assessment" (which is selected), "Reject TDEF Assessment", and "Inconclusive".

Figure 4.8 — Investigator Review Interface

5. Discussion

5.1 Overview

The findings are consistent with the core principle of TDEF: the trustworthiness of AI-assisted drone evidence is not reducible to any one performance measure. There were an intentionally diverse range of findings with respect to the evidential profile. It was easy to establish integrity and provenance, temporal consistency was relatively strong, and controlled cyber-forensic tests led to the expected findings; but object detection coverage was limited, and environmental interference reduced effectiveness.

The distinction carries on the earlier work on drone forensics and preparedness [3], [4] to address AI-assisted interpretation reliability. As such, the contribution of TDEF is not to develop a new detector or tracking algorithm, but to incorporate independent measures of evidence within the forensics decision support process. The findings will be examined below in terms of how they supplement each other, rather than replacing each other.

5.2 Cryptographic integrity and provenance as necessary but insufficient conditions

The hash values of the verified source artefacts and the derived frames support the claim that the considered artefacts were consistent with their provenance. This is consistent with forensic practice, where the use of cryptographic validation and provenance documents provides proof that the artefact has not changed compared to a known state [7].

However, integrity and provenance respond to different questions than reliability. Hash value cannot tell whether an object detector had missed something in the scene, whether the tracker had maintained its correct association, or whether telemetry was accurate when initially recorded. Equally, provenance can prove the provenance of the artefact, but it cannot prove all conclusions drawn from it.

For this reason, TDEF considers cryptographic integrity and provenance as a necessary evidential base rather than a sufficient one with respect to AI interpretability. Such approach reflects the G7 requirement to consider derived artefacts with respect to the verified sources.

5.3 AI Reliability, Detection Completeness, and Class-Level Variation

The main finding regarding visual analysis relates to the discrepancy between precision (70.96%) and recall (15.93%). While most detections turned out to be valid ground-truth objects, 29,472 objects of equal kind were not detected. This is important from a forensic perspective due to the different risks entailed by false positives and false negatives since the former may require further inspection while the latter will simply lack an AI detection.

This result correlates with the well-known difficulties of aerial object detection. The datasets, such as VisDrone and UAVDT, reveal problems that occur frequently within UAV imaging, including small target size, high object density, viewpoint variation, and camera movement [15], [34]. In TDEF, the emphasis is placed on the forensic value of those problems rather than their use to assess detection methods by metrics only. Low recall means that the absence of AI detection should not be considered as evidence of the absence of the object.

Moreover, class-level analysis confirms this finding. For example, the F1-score for cars was 44.34%, and recall was 31.75%, whereas the recall rate for bicycles and motorcycles was 0.16% and 1.23%, respectively. Persons demonstrated quite high precision (76.68%), but very low recall (6.97%) with 12,952 false negatives. Therefore, the general precision score can hide considerable heterogeneity of evidentiary completeness among different object categories.

Accordingly, TDEF evaluates AI Detection Reliability – LIMITED in conjunction with Detection Completeness – LOW. The distinction deals with G2 and G3: while detections look plausible in isolation, they do not provide sufficient evidence that automated analysis gives

adequate reconstruction of the scene. In practice, AI detection can give some grounds for the positive findings, but absence of detection is not proof of the absence of the object.

5.4 Temporal Continuity Is Useful but Should Not Be Confounded with Physical Identity

Reconstruction in time produced much higher levels of continuity compared to the static detection completeness scores. The continuity scores of ByteTrack were 91.78% and that of BoT-SORT – 92.54% with the median of 100% continuity for both setups. This places the test in the category of Temporal Reliability—STRONG within the evaluated sequence (TDEF).

The result is consistent with the objective of the two systems, which are to maintain associations over the course of frames by considering detection, motion, and appearance data in case of BoT-SORT [19], [20]. However, the experiment cannot serve as proof of the overall superiority of either system since they were tested on the same short sequence.

Most importantly, high levels of continuity should not be equated with physically proven identity. A tracking identifier represents an association made by the algorithm and thus is a sign of continuity in terms of the model's reconstruction and not necessarily the identity of the same physical object. For this reason, TDEF considers tracking to be a temporal trust signal while making it clear that the identity is not necessarily proven.

Such a framework provides a solution to G5 by incorporating temporal continuity into the forensic interpretation process while not making any conclusions beyond the scope of the tracking data available.

5.5 Environmental Robustness and Evidential Interpretation

Degradation in the environment gives us many reasons for the multidimensional nature of TDEF. In every case, the F1 of the degradation is lower than the original one, with the Gaussian noise causing the greatest deviation: F1 dropped from 26.02% to 14.89% and recall from 15.93% to 8.27%, while precision increased from 70.96% to 74.51%.

The presented evidence clearly shows why the use of a single performance indicator is enough to get a wrong estimate of the reliability. Due to the presence of noise, the proportion of correct predictions among all predictions grows while the number of recognized relevant objects becomes much smaller. Therefore, the growth of precision goes together with a decrease in completeness of the scene.

This result corresponds to other findings that image corruption can show some problems in the performance which do not appear when we test the model with non-corrupted images [24]. The novelty of our experiment lies in using corruption as the evidential condition of forensic reliance.

Therefore, TDEF includes Environmental Robustness – DEGRADED regardless of Cryptographic Integrity. The image can be authentic and not altered but still become unfit for reliable analysis by a machine learning algorithm. This statement is the answer to G4.

5.6 Cyber-Forensic Integrity and Complementary Validation

In the controlled cyber-forensic experiment, the desired forensic indicator was generated in all six manipulated cases, while the intact baseline was confirmed. The former can be seen as functional validation of the checks performed, rather than as proof of their real-world success in detecting attacks on UAVs.

This point is particularly relevant when talking about location anomalies. It has been reported in previous research that GPS spoofing is one of the possible threats to UAV security [10]. However, while an anomaly in the recorded coordinates transition in the data is an indicator of the attack, it may also be caused by sensor errors or corruptions in data. For this reason, TDEF detects location-integrity or trajectory anomaly, but does not claim any attack.

Furthermore, the experiment showed the complementarity of cryptographic and semantic approaches. While hash check determines whether the artefact is modified since the last recorded baseline, semantic check identifies the nature of the detected anomaly in time, location, trajectory, or cross-source consistency. Cryptographic and semantic checks deliver different types of information because hash check does not reveal which part of the artefact is modified and vice versa – a semantic anomaly does not mean that it was caused by changes in the artefact since its last baseline.

Implementation of both checks in one solution satisfies G6 and G7 by including both cryptographic verification and internal and cross-source consistency checks. Thus, cyber-forensic integrity in TDEF is seen as a set of indicators.

5.7 From Fragments to a Trust Profile

One of the most important results of the study may not be an isolated metric but the culmination of the assessment as a whole:

Dimension	TDEF result
Cryptographic Integrity	VERIFIED
Evidence Provenance	VERIFIED
AI Detection Reliability	LIMITED
Detection Completeness	LOW
Temporal Reliability	STRONG
Environmental Robustness	DEGRADED
Cyber-Forensic Integrity	VERIFIED
Auditability	VERIFIED

This is a case where the evidence is clearly differentiated. Cryptographic integrity cannot substitute for missing objects, as temporal continuity cannot correct for an inconsistent timestamp. The use of a weighted score could make this distinction blurry because it allows for the strength in one dimension to counteract the weakness in another without any forensic basis for doing so.

Thus, trust is modelled by TDEF in its multidimensionality. Hashing provides the dimension of integrity, provenance provides the dimension of traceability, ground truth evaluation provides the dimension of knowledge of detection and completeness, tracking provides the dimension of temporal continuity, degradation testing provides the dimension of robustness, telemetry provides the dimension of cyber forensics, and auditing provides the dimension of process traceability.

5.8 Connection between Results and Identified Research Gaps

The findings have direct correlation with the eight gaps revealed in the literature review and are presented in Table below.

Gap	Results from the current study
G1 — Fragmented forensic workflow	Ingestion, AI-based analysis, monitoring, robustness, telemetry consistency, and forensic evaluation have been integrated into a single TDEF pipeline.
G2 — AI performance—trust conflation	Prediction precision of 70.96% alongside the recall of 15.93% prove that prediction performance alone does not indicate evidentiary trustworthiness.
G3 — Lack of evaluation of evidentiary completeness	29,472 false negatives and low class-specific recalls were specifically maintained via the Detection Completeness dimension.
G4 — Lack of robustness inclusion	Four degradation conditions decreased F1 scores, and noise in particular lowered the score by 11.13 percentage points.
G5 — Lack of temporal trust evaluation	ByteTrack and BoT-SORT algorithms were evaluated regarding continuity and fragmentation, and track identity restrictions were maintained specifically.
G6 — Lack of cyber-forensic validation	Indicators related to hash, time, location, cross-source timeline, and trajectory were evaluated together for six controlled cases.
G7 — Incomplete integration of provenance and audit	Source and derived hashes, provenance records, audit events, and investigator reviews have been maintained all through the process.

G8 — Lack of explainable multidimensional integration

Eight separately interpretable dimensions have been integrated into TDEF trust profile and disposition for human review.

The present research does not suggest that these issues have been solved completely. On the contrary, the paper suggests an approach to the problem that has been practically tested in the established research framework. Therefore, TDEF must be viewed as proof that a multifaceted evaluation of forensic processes can uncover problems which remain unseen when integrity, AI, robustness, and cyber-forensics are evaluated separately.

5.9 Explainability and Human Review

The final disposition of HUMAN REVIEW REQUIRED is based on the three factors, i.e., AI Detection Reliability, Detection Completeness, and Environmental Robustness, but this disposition does not result from any doubts in terms of the underlying evidential facts. It is important to note that the uncertainty in automation-based analysis cannot be automatically treated as an issue of authenticity, admissibility, and veracity of evidence.

These features of the system are manifested by the transparency of the relevant review dimensions and the possibility for the investigator to choose Confirm TDEF Assessment, Reject TDEF Assessment, or Inconclusive with explanations. The response of the investigator is collected separately from the automated one.

In other words, explainability in TDEF includes not only reporting of technical metrics but also providing explanation of why particular disposition is selected. For example, the combination of the following parameters, VERIFIED cryptographic integrity, LOW detection completeness, and DEGRADED environmental robustness, means that the concern pertains to reliability rather than authenticity of evidence. Otherwise, the issue will be concealed in a single numeric evaluation of trust.

5.10 Practical Implications and Research Contribution

Implications of this work are numerous. All outputs produced by AI should maintain connections to the inputs to distinguish the collected artefacts from the interpretation results. The detection outputs should be presented together with the known limitations concerning completeness, while the environmental context should be considered when relying on automated interpretation. Temporal tracking should be viewed as algorithmic association, rather than verified identity, and

telemetry should be analysed in terms of both cryptography and semantics. Human reactions to the automated interpretation should also be traceable in the forensic process.

These implications complement the existing demands for structured drone-forensic investigation and forensic readiness [3], [4] considering an added challenge introduced by AI participation in the evidentiary interpretation.

Three interconnected contributions are made by this research. Technical contribution consists in the integration of cryptographic verification, provenance, object detection, temporal reconstruction, environmental robustness, cyber-forensic consistency, audit logging, and investigation in one workflow. Experimental contribution involves evaluation of these aspects on the whole validation data set of 548 images of VisDrone dataset, a temporal reconstruction experiment and cyber-forensic manipulations in six scenarios. Conceptual contribution shows why the trust in AI-assisted drone evidence should be described in terms of multi-dimensional trust profile instead of model confidence or some other arbitrarily chosen measure.

The individual techniques used by TDEF are not claimed to be new, but are well-known: SHA-256, object detection, ByteTrack, BoT-SORT, and telemetry consistency analysis are among them. Contribution is made in the combination, experimental evaluation and conceptual interpretation of these techniques.

The research contribution can be stated as follows:

TDEF is a provenance-aware framework that integrates heterogeneous indicators of trust in AI-assisted drone evidence with their explanations while maintaining their forensic meanings and preserving the investigator's judgment.

6. Limitations and Future Work

6.1 Limitations

The conclusions should be considered with regards to the limitations of the experiment. Firstly, the main visual experiments involved just one pretrained detector, YOLO11n, and one aerial benchmark, VisDrone2019-DET. Therefore, the reported detection characteristics apply only to this configuration and cannot be automatically applied to any UAV system, sensors, environment, and detection models. Secondly, only VisDrone classes that have meaningful COCO analogy were evaluated, and the specially designed TP/FP/FN assessment at IoU 0.50 should not be regarded as reproduction of VisDrone map benchmark.

The robustness to environmental conditions was studied by means of reproducible manipulation with blur, brightness, contrast, and Gaussian noise. However, this list of manipulations does not cover all atmospheric, optical, compression, or sensor-based factors that can impact UAV evidence during operations. Future research should involve naturally degraded imagery from specific UAV platforms.

The temporal validation was performed using a single short video; thus, the obtained results related to the ByteTrack and BoT-SORT should not be considered as a comparison of trackers' performance. Validation using multiple longer sequences with various object densities and camera movement will improve the reliability of this test.

The cyber-forensics scenarios were also limited. The video and telemetry data were connected to each other to perform experiments, while they were not captured simultaneously by the same UAV. The six manipulations test the framework's behaviour but not its capability to detect attacks in real life. Future research should involve the use of TDEF in case of synchronized multimodal captures and, if possible, ethically and legally, on UAV incident data.

6.2 Future Work

Next steps should improve the TDEF framework from experimental research prototype to a framework that is applicable to evaluation of operationally relevant synchronized UAV evidence. First, it should involve building an acquisition environment for TDEF in which video, GPS data, telemetry data, time stamps, and inertial sensor data where available would be acquired from the same drone flight. This way the current dependence of the TDEF framework on experimentally associated video and telemetry data would be addressed, and provenance, temporal correlation, and cross source correlation could be evaluated under operationally relevant conditions of acquisition.

Another research direction could be assessment of TDEF in multiple object detection and tracking configurations rather than only with YOLO11n and one short sequence of tracking. Experiments could include comparing detectors with different architecture and computational load while applying the same forensic trust dimensions. The idea would be not to find a best model but to assess whether TDEF is able to consistently describe changes in reliability, completeness, temporal consistency, and robustness for different AI systems and datasets.

Evaluation of TDEF performance in the environmental aspect should transition from artificial corruption to naturally degraded drone evidence. The real flight data with naturally occurring issues caused by low illumination, motion blur, adverse weather, compression, height, and unstable camera movements would be much more appropriate for assessment of whether the degradation patterns that were discovered in the current study would transfer to operational

evidence. It could be supplemented with automatic image quality measures to estimate the environmental robustness prior to the interpretation of the AI output.

In terms of the cyber forensic aspect, the experiments could transition to a controlled UAV testbed where telemetry would be corrupted with realistic sensor, timing, and navigation issues instead of only file modifications. Such experiments would allow testing the ability of TDEF to distinguish post-acquisition manipulation, sensor failure, and navigation errors. The key point is that TDEF should keep on reporting the observable forensic indicators without attributing the malicious intention automatically.

Another important research direction is the development of empirical trust thresholds. The current thresholds of the temporal continuity and spatial anomalies were selected specifically for the experiment purpose, and they are not forensic standards. With the help of larger multi-modal datasets, the decision boundaries could be defined based on statistical grounds while maintaining the multidimensional structure of TDEF.

Overall, such development of TDEF could result in investigator-oriented framework that would allow comparison of evidence with respect to different UAVs, different AI models, and different operating conditions without replacing human judgment.

References:

References

- [1] National Police Chiefs' Council. *Drones strengthening city centre policing*. Available: <https://news.npcc.police.uk/releases/drones-strengthening-city-centre-policing>.
- [2] M. Police. *Met to announce ambitious plans to scale drones, LFR and AI in fight against criminals*. Available: <https://news.met.police.uk/news/met-to-announce-ambitious-plans-to-scale-drones-lfr-and-ai-in-fight-against-criminals-510705>.
- [3] E. Mantas and C. Patsakis, "Who watches the new watchmen? The challenges for drone digital forensics investigations," *Array*, vol. 14, pp. 100135, 2022. . DOI: 10.1016/j.array.2022.100135.
- [4] F. M. Alotaibi, A. Al-Dhaqm and Y. Al-Otaibi, "A novel forensic readiness framework applicable to the drone forensics field," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 8002963, 2022. . DOI: 10.1155/2022/8002963.
- [5] M. A. H. Bin Azhar, T. E. A. Barton and T. Islam, "Drone forensic analysis using open source tools," *Journal of Digital Forensics, Security and Law*, vol. 13, (1), pp. 6, 2018. . DOI: 10.15394/jdfsl.2018.1513.
- [6] A. Taylor, "A digital forensics case study of the DJI Mini 3 Pro and DJI RC," *arXiv*, vol. 2309.10487, 2023. Available: <https://arxiv.org/abs/2309.10487>. DOI: 10.48550/arXiv.2309.10487.
- [7] D. E. Scientific Working Group on, "Best practices for drone forensics," 2025 Available: <https://www.swgde.org/documents/published-complete-listing/21-f-002-2/>.
- [8] a. T. National Institute of Standards. *Computer Forensic Reference Data Sets: Drone Data Set*. Available: <https://cfreds-archive.nist.gov/drone-images.html>.
- [9] V. Sihag *et al*, "Cyber4Drone: A systematic review of cyber security and forensics in next-generation drones," *Drones*, vol. 7, (7), pp. 430, 2023. . DOI: 10.3390/drones7070430.
- [10] A. Shafique, A. Mehmood and M. Elhadeif, "Detecting signal spoofing attack in UAVs using machine learning models," *IEEE Access*, vol. 9, pp. 93803–93815, 2021. . DOI: 10.1109/ACCESS.2021.3089847.
- [11] A. Altaweel, H. Mulkath and I. Kamel, "GPS spoofing attacks in FANETs: A systematic literature review," *IEEE Access*, vol. 11, pp. 55233–55280, 2023. . DOI: 10.1109/ACCESS.2023.3281731.
- [12] M. Shin *et al*, "Trajectory-driven deep learning for UAV location integrity checks," *IEEE Access*, vol. 12, pp. 178789–178804, 2024. . DOI: 10.1109/ACCESS.2024.3507637.

- [13] B. Pardhasaradhi and L. R. Cenkeramaddi, "GPS spoofing detection and mitigation for drones using distributed radar tracking and fusion," *IEEE Sensors Journal*, vol. 22, (11), pp. 11122–11134, 2022. . DOI: 10.1109/JSEN.2022.3168940.
- [14] P. Zhu *et al*, "Detection and tracking meet drones challenge," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, (11), pp. 7380–7399, 2022. . DOI: 10.1109/TPAMI.2021.3119563.
- [15] D. Du *et al*, "The unmanned aerial vehicle benchmark: Object detection and tracking," *Computer Vision – ECCV 2018*, vol. LNCS 11214, pp. 375–391, 2018. . DOI: 10.1007/978-3-030-01249-6_23.
- [16] X. Wu *et al*, "Deep learning for UAV-based object detection and tracking: A survey," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, (1), pp. 91–124, 2022. . DOI: 10.1109/MGRS.2021.3115137.
- [17] J. An *et al*, "DCE-YOLOv8: Lightweight and accurate object detection for drone vision," *IEEE Access*, vol. 12, pp. 170898–170912, 2024. . DOI: 10.1109/ACCESS.2024.3481410.
- [18] X. Li *et al*, "OGMN: Occlusion-guided multi-task network for object detection in UAV images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 199, pp. 242–257, 2023. . DOI: 10.1016/j.isprsjprs.2023.04.009.
- [19] Y. Zhang *et al*, "ByteTrack: Multi-object tracking by associating every detection box," *Computer Vision – ECCV 2022*, pp. 1–21, 2022. . DOI: 10.1007/978-3-031-20047-2_1.
- [20] N. Aharon, R. Orfaig and B. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," *arXiv*, vol. 2206.14651, 2022. Available: <https://arxiv.org/abs/2206.14651>. DOI: 10.48550/arXiv.2206.14651.
- [21] I. Bozcan and E. Kayacan, "UAV-AdNet: Unsupervised anomaly detection using deep neural networks for aerial surveillance," *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1158–1164, 2020. . DOI: 10.1109/IROS45743.2020.9341790.
- [22] P. Jin *et al*, "Anomaly detection in aerial videos with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 5628213, 2022. . DOI: 10.1109/TGRS.2022.3198130.
- [23] W. Fang *et al*, "Multi-task learning for UAV aerial object detection in foggy weather condition," *Remote Sensing*, vol. 15, (18), pp. 4617, 2023. . DOI: 10.3390/rs15184617.
- [24] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *International Conference on Learning Representations (ICLR)*, 2019. Available: <https://arxiv.org/abs/1903.12261>.
- [25] a. T. National Institute of Standards, "Secure hash standard (SHS)," 2015 Available: <https://csrc.nist.gov/pubs/fips/180-4/upd1/final>. DOI: 10.6028/NIST.FIPS.180-4.

- [26] f. S. International Organization, "ISO/IEC 27037:2012 — Information technology — Security techniques — Guidelines for identification, collection, acquisition and preservation of digital evidence," 2012. Available: <https://www.iso.org/standard/44381.html>.
- [27] S. Nath *et al*, "Digital evidence chain of custody: Navigating new realities of digital forensics," *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, pp. 11–20, 2024. . DOI: 10.1109/TPS-ISA62245.2024.00012.
- [28] S. Bonomi, M. Casini and C. Ciccotelli, "B-CoC: A blockchain-based chain of custody for evidences management in digital forensics," *arXiv*, vol. 1807.10359, 2018. Available: <https://arxiv.org/abs/1807.10359>.
- [29] E. Tabassi, "Artificial intelligence risk management framework (AI RMF 1.0)," National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.AI.100-1.
- [30] Information Commissioner's Office. *Unmanned Aerial Systems (UAS) / Drones*. Available: <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/cctv-and-video-surveillance/guidance-on-video-surveillance-including-cctv/additional-considerations-for-technologies-other-than-cctv/unmanned-aerial-systems-uas-drones/>.
- [31] A. A. UK Civil. *Privacy rules when flying drones*. Available: <https://www.caa.co.uk/drones/open-category/moving-on-to-more-advanced-flying/privacy-rules-when-flying-drones/>.
- [32] D. E. Scientific Working Group on, "Best practices for computer forensic acquisitions," 2025 Available: <https://www.swgde.org/documents/published-complete-listing/17-f-002-2-1/>.
- [33] G. Jocher and J. Qiu. *Ultralytics YOLO11*. Available: <https://github.com/ultralytics/ultralytics>.
- [34] D. Du *et al*, "VisDrone-DET2019: The Vision Meets Drone Object Detection in Image Challenge Results," *2019 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 213–226, 2019. . DOI: 10.1109/ICCVW.2019.00030.

Appendices:

Appendix A — Experimental Evidence Screenshots

A.1 — Evidence Provenance Record

```
data > processed > frame_log.csv

You, last month | 1 author (You)

1 video_name,video_sha256,frame_number,timestamp_seconds,saved_frame,frame_sha256
2 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,0,0.000,frame_000000.jpg,f5c
3 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,25,1.000,frame_000025.jpg,3f
4 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,50,2.000,frame_000050.jpg,b9
5 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,75,3.000,frame_000075.jpg,30
6 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,100,4.000,frame_000100.jpg,8
7 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,125,5.000,frame_000125.jpg,5
8 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,150,6.000,frame_000150.jpg,0
9 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,175,7.000,frame_000175.jpg,6
10 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,200,8.000,frame_000200.jpg,f
11 tracking_test_10s.mp4,85ef2ef5828f54a1fda0328ab204a31bc62ee2522666bd1c4d634330ad2c9a52,225,9.000,frame_000225.jpg,9
```

Figure A.1. Frame-level evidence provenance record

A.2 — The important VisDrone raw result

```
outputs > visdrone_evaluation > {} visdrone_full_548_results.json > ...

1 {
2   "experiment": "VisDrone YOLO11n subset evaluation",
3   "subset_size": 548,
4   "random_seed": 42,
5   "model": "yolo11n.pt",
6   "confidence_threshold": 0.25,
7   "iou_threshold": 0.5,
8   "device": "mps",
9   "total_ground_truth_objects": 35057,
10  "total_predictions": 7871,
11  "true_positives": 5585,
12  "false_positives": 2286,
13  "false_negatives": 29472,
14  "precision": 0.709567,
15  "recall": 0.159312,
16  "f1_score": 0.260203,
17  "processing_seconds": 18.27,
18  "per_class_metrics": {
19    "bicycle": {
20      "true_positives": 2,
21      "false_positives": 4,
22      "false_negatives": 1276,
23      "precision": 0.333333,
```

Figure A.2. Full VisDrone2019-DET evaluation results

A.3 — Tracker evidence

outputs > tracker_comparison > {} tracker_comparison_results.json > ...

```
18 "results": [  
19   {  
20     "tracker": "bytetrack.yaml",  
21     "model": "yolo11n.pt",  
22     "confidence_threshold": 0.25,  
23     "video_frames": 250,  
24     "processed_frames": 250,  
25     "fps": 25.0,  
26     "unique_tracks": 78,  
27     "tracking_records": 2229,  
28     "mean_track_continuity": 0.917753,  
29     "median_track_continuity": 1.0,  
30     "high_continuity_tracks": 62,  
31     "fragmented_tracks": 1,  
32     "mean_detection_confidence": 0.559275,  
33     "processing_seconds": 8.685  
34   },  
35   {  
36     "tracker": "botsort.yaml",  
37     "model": "yolo11n.pt",  
38     "confidence_threshold": 0.25,  
39     "video_frames": 250,  
40     "processed_frames": 250,  
41     "fps": 25.0,  
42     "unique_tracks": 75,  
43     "tracking_records": 2229,  
44     "mean_track_continuity": 0.92535,  
45     "median_track_continuity": 1.0,  
46     "high_continuity_tracks": 58,  
47     "fragmented_tracks": 0,  
48     "mean_detection_confidence": 0.558644,  
49     "processing_seconds": 7.65  
50   }  
]
```

Figure A.3. ByteTrack and BoT-SORT comparison results

A.4 — Environmental robustness

outputs > visdrone_degradation > {} visdrone_degradation_full_548_images.json > ...

```
567     "results": {
568       "original": {
569         "ground_truth": 35057,
570         "predictions": 7871,
571         "true_positives": 5585,
572         "false_positives": 2286,
573         "false_negatives": 29472,
574         "processing_seconds": 15.927,
575         "precision": 0.709567,
576         "recall": 0.159312,
577         "f1_score": 0.260203,
578         "precision_change_pp": 0.0,
579         "recall_change_pp": 0.0,
580         "f1_change_pp": 0.0
581       },
582       "blur": {
583         "ground_truth": 35057,
584         "predictions": 6051,
585         "true_positives": 4313,
586         "false_positives": 1738,
587         "false_negatives": 30744,
588         "processing_seconds": 10.194,
589         "precision": 0.712775,
590         "recall": 0.123028,
591         "f1_score": 0.209838,
592         "precision_change_pp": 0.321,
593         "recall_change_pp": -3.628,
594         "f1_change_pp": -5.037
595       },
596       "low_brightness": {
597         "ground_truth": 35057,
598         "predictions": 6951,
599         "true_positives": 5076,
600         "false_positives": 1875,
601         "false_negatives": 29981,
602         "processing_seconds": 9.442,
603         "precision": 0.730255,
604         "recall": 0.144793,
605         "f1_score": 0.241668,
606         "precision_change_pp": 2.069,
607         "recall_change_pp": -1.452,
608         "f1_change_pp": -1.854
609       },
610       "low_contrast": {
611         "ground_truth": 35057,
612         "predictions": 6264,
613         "true_positives": 4526,
614         "false_positives": 1738,
615         "false_negatives": 30531,
616         "processing_seconds": 9.491,
617         "precision": 0.722542,
618         "recall": 0.129104,
619         "f1_score": 0.219065,
620         "precision_change_pp": 1.298,
621         "recall_change_pp": -3.021,
622         "f1_change_pp": -4.114
623       },
624       "noise": {
625         "ground_truth": 35057,
626         "predictions": 3892,
627         "true_positives": 2900,
628         "false_positives": 992,
629         "false_negatives": 32157,
630         "processing_seconds": 9.553,
631         "precision": 0.745118,
632         "recall": 0.082722,
633         "f1_score": 0.148913,
634         "precision_change_pp": 3.555,
635         "recall_change_pp": -7.659,
636         "f1_change_pp": -11.129
637       }
638     }
```

Figure A.4. Environmental robustness evaluation results

A.5 — Cyber-Forensic Validation

```
outputs > cyber_forensic_integrity > {} day12_final_validation_results.json > ...
268     "evaluation": {
269         "scenario_count": 6,
270         "correct_target_detections": 6,
271         "target_detection_rate": 1.0,
272         "scenario_evaluations": [
273             {
274                 "scenario": "gps_spoofing_simulation",
275                 "expected_flag": "location_integrity_anomaly",
276                 "observed_flags": [
277                     "telemetry_hash_mismatch",
278                     "location_integrity_anomaly"
279                 ],
280                 "target_detected": true
281             },
282             {
283                 "scenario": "timestamp_manipulation",
284                 "expected_flag": "timestamp_sequence_anomaly",
285                 "observed_flags": [
286                     "telemetry_hash_mismatch",
287                     "timestamp_sequence_anomaly"
288                 ],
289                 "target_detected": true
290             },
291             {
292                 "scenario": "telemetry_log_tampering",
293                 "expected_flag": "telemetry_hash_mismatch",
294                 "observed_flags": [
295                     "telemetry_hash_mismatch"
296                 ],
297                 "target_detected": true
298             },
299             {
300                 "scenario": "video_telemetry_time_mismatch",
301                 "expected_flag": "cross_source_timeline_mismatch",
302                 "observed_flags": [
303                     "telemetry_hash_mismatch",
304                     "cross_source_timeline_mismatch"
305                 ],
306                 "target_detected": true
307             },
308             {
309                 "scenario": "evidence_hash_mismatch",
310                 "expected_flag": "frame_hash_mismatch",
311                 "observed_flags": [
312                     "frame_hash_mismatch"
313                 ],
314                 "target_detected": true
315             },
316             {
317                 "scenario": "flight_path_discontinuity",
318                 "expected_flag": "trajectory_discontinuity",
319                 "observed_flags": [
320                     "telemetry_hash_mismatch",
321                     "location_integrity_anomaly",
322                     "trajectory_discontinuity"
323                 ],
324                 "target_detected": true
325             }
326         ]
327     }
```

Figure A.5. Cyber-forensic integrity validation results