Multi-Partner Project: Enhancing Resilience, Efficiency, and Trustworthiness of Edge AI in Safety-Critical Systems (GuardAI)

Antonis Savva*, Mehmet Demirel*, Yeshwanth Kumar Adimoolam*, Rafaella Elia*,
Alexandros Gkillas‖, Christos Anagnostopoulos‖, Erion-Vasilis Pikoulis‖, Amalia Damianou§,
Charmaine Barker¶, Daniel Bethell¶, Ahmed Salah Tawfik Ibrahim‡, Filippo Cugini‡,
Francesco Paolucci‡, Kyriakos Vlachos‡, Simos Gerasimou¶, Antonios Lalas§,
Konstantinos Votis§, Aris Lalos‖, Christos Kyrkou*, Theocharis Theocharides†,
*KIOS Research and Innovation Center of Excellence, University of Cyprus,
†Electrical and Computer Engineering Department, University of Cyprus,
‡National Inter-University Consortium for Telecommunications (CNIT), Italy
§Centre for Research and Technology, Hellas (CERTH), Greece
¶Department of Computer Science, University of York, UK
‖Industrial Systems Institute, ATHENA R.C., Greece

***Abstract*—AI at the network edge promises real-time perception and decision-making in safety-critical domains such as aerial robotics, autonomous vehicles, and 5G-enabled infrastructures. Yet, operating under resource constraints, dynamic, and adversarial conditions exposes edge AI systems to fragility, inefficiency, and security risks that threaten their safe operation. GuardAI, a Horizon Europe project, introduces a framework for resilient and trustworthy edge AI that unites three pillars: adversarial robustness, context-enhanced inference, and security-by-design. Initial project results include a diffusion-based adversarial purification framework optimized for real-time operation, lightweight deep unrolling architectures for LiDAR super-resolution with built-in outlier removal, and robust uncertainty quantification modules to improve confidence calibration. It further develops a context-enhanced inference engine that integrates visual, spatial, and operational context across multi-agent systems, and a risk-aware defense recommender that autonomously selects mitigation strategies based on evolving threat landscapes. Through representative Use Cases, covering monitoring with Unmanned Aerial Vehicle, decentralized 5G network analytics, and secure perception in connected autonomous vehicles, GuardAI demonstrates how robust and adaptive AI can be achieved within stringent edge constraints. Together, these technologies lay the groundwork for a new generation of secure, context-aware, and certifiable AI systems that can be trusted to operate autonomously in the physical world.**

***Index Terms*—adversarial attacks, adversarial resilience, Trustworthy AI, Robust AI, Edge AI, UAVs, connected autonomous vehicles, network edge infrastructure**

I. Introduction

Deploying AI models on edge devices enables real-time, intelligent data processing by minimizing reliance on cloud systems and bringing computation closer to data sources. This supports immediate decisions in high-stakes areas such as Unmanned Aerial Vehicle (UAV)-enabled monitoring [1], connected vehicles [2], and network-edge infrastructures [3]. Compared to cloud computing, edge AI cuts latency and bandwidth use by reducing large data transfers.

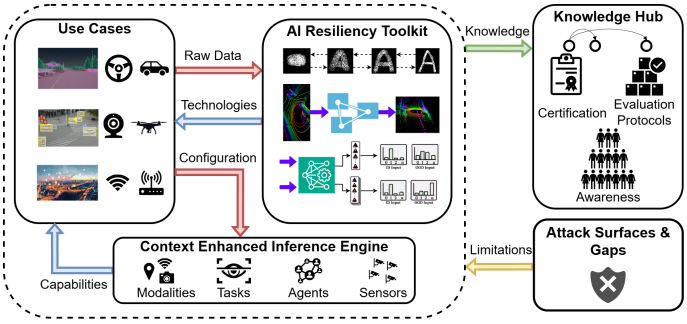

Fig. 1. The multi-faceted GuardAI concept for the new generation of secure, context-aware, and certifiable AI systems operating autonomously in the physical world.

However, substantial challenges exist. Edge AI operates under uncertainty, limited computational resources, and often sensitive data processing. Risks include sensor data manipulation and environmental impacts that may directly compromise human safety [4]. Solutions usually address narrow aspects: adversarial defenses reduce clean-input accuracy [5]; perception systems fail in poor weather conditions [6]; anomaly detection struggles with high-dimensional, imbalanced data and requires extensive labels [7], while robustness methods add computational costs incompatible with edge limits.

These fragmented landscape illustrates the need for holistic solutions that jointly address robustness, computational efficiency, and trustworthiness in real-world deployments. In particular, resilience at the edge requires robust and efficient algorithms, context-awareness (to adapt to dynamic conditions), security-by-design (to anticipate evolving threats), and clear evaluation protocols that can support future certification.

The **GuardAI**[1] Horizon Europe project is focused exactly on tackling the aforementioned issues by developing resilient,

[1]https://www.kios.ucy.ac.cy/guardai/ (3-year project, Oct 2024–Sept 2027)

978-3-9826741-1-7/DATE26/© 2026 EDAA

efficient, and trustworthy AI for safety-critical edge applications. It achieves this through integration of adversarial robustness, context-enhanced inference, automated defense recommendation, and standardized evaluation. This paper provides an overview of the **GuardAI** project while identifying challenges and presenting initial solutions established through:

- Articulation of the GuardAI vision and high-level architecture.
- Identification of key challenges in adversarial robustness, context integration, and security-by-design for edge AI.
- Presentation of methodological pillars, including an adversarially resilient toolkit, a context-aware inference engine, and a defense recommender system, together with initial results.
- Description of representative Use Cases in UAV-based monitoring, Connected Automated Vehicles (CAVs), and 5G-enabled infrastructure.

II. GuardAI Approach

GuardAI (Fig. 1) initially analyzes attack surfaces and identifies vulnerabilities in existing AI deployments. This informs novel defense mechanisms, including robustification mechanisms, unsupervised outlier removal, and robust uncertainty quantification. A particular focus is the development mechanisms that enable AI systems to incorporate spatial, sensory, and operational indicators directly into their inference pipelines. These mechanisms adapt across various applications, such as UAV monitoring versus CAV perception, while remaining compatible with edge resource limitations.

Given the complexity of tailoring defenses to diverse edge scenarios, GuardAI also develops a defense recommender system that selects appropriate security mechanisms for a given threat profile and deployment context, thereby embedding the security-by-design principle in practice.

All knowledge generated is continuously integrated into a collaborative knowledge hub[2], which consolidates resilience insights and formalizes standardized evaluation metrics. Beyond functioning as a repository, the hub provides a medium to support dialogue among researchers, industry, and regulators. This foundation will be crucial for advancing certification and compliance in safety-critical edge AI systems.

III. Challenges of Robust Edge AI Systems

The deployment of Edge AI in safety-critical domains introduces interconnected challenges that must be addressed to ensure resilience, efficiency, and trustworthiness. Systems that depend on autonomous perception and decision-making are vulnerable to unauthorized access, data breaches, and manipulation of sensor data or control flows, creating weaknesses with potentially severe consequences. We use *robustness* to denote resistance to adversarial perturbations, *reliability* for consistent behavior, *resilience* for recovery capability, and *trustworthiness* for the combination of efficiency, robustness, and regulatory compliance.

[2]https://www.kios.ucy.ac.cy/guardai/K4AI/

Among these threats, adversarial attacks pose a particularly critical risk: carefully crafted perturbations or sensor spoofing can deceive AI models, leading to misclassifications and unsafe actions. For example, by changing sensor data, an attacker could cause a drone to misidentify objects or misinterpret its environment, leading to potentially dangerous situations. Adversarial attacks often use gradient-based optimization techniques to find the most effective perturbations. These methods exploit the model's vulnerability to input changes by analyzing variations in the loss function (e.g., Fast Gradient Sign Method, Projected Gradient Descent) [8].

Edge platforms face equally tough challenges due to strict computational and energy constraints. Robustness strategies often add significant overheads; for instance, adversarial training is a highly effective defense method [9]. The main challenge is that, with this approach, training robust models on large datasets dramatically increases training time compared to standard methods [5]. As a result, research focused on improving the efficiency of the basic adversarial training algorithm [10], [11]. Achieving a balance among robustness and efficiency under these constraints remains an ongoing challenge. Specifically, deploying these models on edge devices requires techniques such as pruning and quantization to compress the models [12]. However, these methods often reduce robustness; for example, adversarial training effectiveness drops significantly when parameters are quantized from 32-bit floats to 8-bit integers [13], [14].

Another aspect of complexity arises from the need to ensure reliability in dynamic, unpredictable environments. Edge AI deployments must withstand physical and digital tampering, communication interruptions, hardware failures, and software bugs, while maintaining safety-critical functions. These systems must also adapt to rapidly changing conditions (e.g., lighting, weather, etc.), while providing dependable performance despite uncertainty. Besides data captured in the visible spectrum, other sensing methods such as infrared, hyperspectral imaging, radar, and Light Detection and Ranging (LiDAR) are increasingly used in various applications with UAVs and CAVs. However, their vulnerability to adversarial attacks, both digital and physical, remains underexplored. In this context, it is essential to determine whether attack strategies designed for the visible spectrum can be adapted or extended to other sensing modalities in different operating conditions. This is especially important in settings where multimodal sensing is used, as robustness across multiple data streams is crucial. [15].

Finally, ethical, legal, and regulatory considerations present equally critical challenges. These systems frequently process sensitive personal data, operate in real time, and make autonomous interventions in physical environments, raising concerns about privacy, accountability, and compliance with human rights. Ensuring trustworthiness requires robust safeguards, including strong data protection mechanisms, transparency and explainability of AI decision-making, and appropriate human oversight. Moreover, bias mitigation, compliance with evolving regulations (e.g., General Data Protection Directive (GDPR),

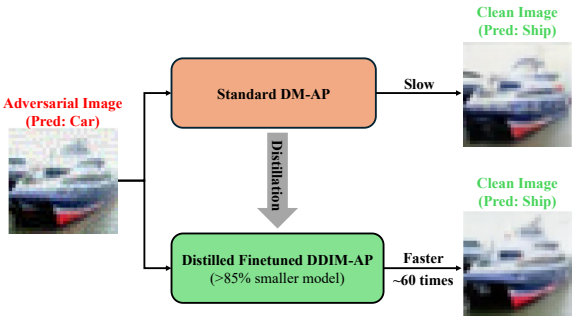

Fig. 2. Efficient adversarial purification framework. Our Distilled Finetuned DDIM-AP provides up to 60x speedup and over 85% model size reduction while successfully removing adversarial perturbations.

Artificial Intelligence Act[3], NIS2 Directive[4], and the Cybersecurity Act[5]), and certification pathways are essential to guarantee that Edge AI technologies can be deployed responsibly in high-stakes applications.

IV. GuardAI Technologies

A. Adversarially Resilient Toolkit

1) Diffusion-based Adversarial Purification: State-of-the-art defenses such as Diffusion Model-based Adversarial Purification (DM-AP) have emerged as a powerful paradigm for protecting AI models against adversarial inputs. DM-AP takes manipulated input, adds noise, then uses the diffusion model's reverse process to restore it to clean data, removing adversarial perturbations before passing to a downstream model. It's modular, doesn't require retraining, and handles unseen threats well. However, it relies on an iterative denoising process that often requires hundreds of steps, each costly and slow, making real-time on-device use impractical [16].

To enable practical edge purification, GuardAI introduces an efficient purification framework (Fig. 2). Instead of standard diffusion models (DDPM) [17], we use Denoising Diffusion Implicit Models (DDIM) [18] that introduce a more flexible, non-Markovian reverse process. This allows larger "jumps" between denoising steps, significantly reducing the number of sampling iterations without retraining. To address edge hardware constraints, we also employ knowledge distillation, followed by a fine-tuning step to maintain generative quality and robustness.

The DDIM-based approach greatly enhances inference speed, delivering up to $60\times$ faster performance compared to standard DM-AP on GPUs and edge devices (NVIDIA Jetson Xavier NX). The distilled, fine-tuned model reduces model size and parameter count by over 85%, lowering memory and computational demands. This efficiency also improves robustness, resulting in higher robust accuracy against strong adaptive attacks (Backward Pass Differentiable Approximation + Expectation over Transformation [19]) from 55.7% (standard DM-AP) to 74.3%.

[3]Regulation (EU) 2024/1689
[4]Directive (EU) 2022/2555
[5]Regulation (EU) 2019/881

2) Deep Unrolling for LiDAR Super-Resolution with Outlier Removal: Low-resolution LiDAR sensors are widely deployed in autonomous vehicles and robotics due to their affordability and energy efficiency [20]. However, their limited vertical resolution severely degrades Simultaneous Localization and Mapping (SLAM) performance, leading to drift, poor scan alignment, and increased vulnerability to noise or adversarial interference. Conventional super-resolution (SR) methods attempt to recover dense point clouds. Still, they often introduce artifacts, rely on heavy neural networks, or require costly post-processing to remove spurious data, rendering them unsuitable for real-time, safety-critical systems [21].

To address these challenges, GuardAI incorporates a model-based deep unrolling architecture for LiDAR SR that is both lightweight and resilient. The approach reformulates SR as an optimization problem in the range-image domain. It unrolls its iterative solution into a deep network, preserving the geometric structure of LiDAR data while enabling fast inference. Crucially, an outlier removal module is embedded directly into the optimization process. This integrated design allows the network to jointly enhance resolution while suppressing spurious measurements arising from environmental noise or adversarial manipulation, without relying on external filters or additional latency-inducing steps.

The resulting SR module runs in real time (400 FPS) and is fully compatible with LiDAR SLAM pipelines, thereby significantly improving trajectory estimation and robustness under challenging sensing conditions. Experimental results (Ouster LiDAR dataset; 15-minute drive through San Francisco) demonstrate up to 58% reduction in Absolute Pose Error and over 99% lower model complexity (0.2M vs. 50M parameters) compared to transformer-based SR methods, while maintaining state-of-the-art accuracy for high-resolution reconstruction.

3) Robust Uncertainty Quantification: Modern deep learning models have achieved remarkable success across multiple decision-making tasks, yet unreliable confidence estimates hinder their deployment in safety-critical domains [22]. Conventional models often yield overconfident predictions under distributional- and covariate-shift, adversarial perturbations, or data noise, leading to unsafe or misleading predictions [23]. Robust uncertainty quantification employs models aware of their predictive reliability.

Uncertainty arises from two sources: epistemic uncertainty reflects the model's poor training or limited knowledge, and aleatoric uncertainty stems from inherent data noise [24]. Methods that prompt the model's own parameters can capture epistemic uncertainty, but can also incur high inference/training costs [25]. Other methods attempt to model the noise inherent to the data using generative or heteroscedastic approaches, thus encoding aleatoric uncertainty [26], [27]. Hybrid methods that attempt to capture both uncertainty sources and reason about their combined effects include Conformal Prediction [28] and Evidential Deep Learning [29].

To enhance robustness in uncertainty quantification, we have been developing the following comprehensive approaches.

First, Monte Carlo Conformal Prediction (MC-CP) [30] combines a lightweight stochastic inference mechanism with conformal prediction to construct prediction sets or intervals rather than singletons or point estimates, providing strong coverage guarantees. Due to the hybrid approach, the head and tail of the predictive distribution are heavily regularised, resulting in compact prediction sets that meaningfully reflect the model's uncertainty. MC-CP, albeit simple, outperforms state-of-the-art CP-based and MC-based methods, e.g., traditional MC dropout, RAPS [31] and CQR [32], both in classification (e.g., test error – MC-CP: 3.99 $\pm$ 0.41, RAPS: 4.57 $\pm$ 0.09 in Tiny ImageNet) and regression benchmarks (e.g., empirical coverage – MC-CP: 96.06 $\pm$ 0.73, CQR: 94.79 $\pm$ 0.01 in Protein structure).

Second, the Gradual Uncertainty Refinement via Noise-Driven Curriculum (GUIDE) meta-model [33] is a post-hoc method that models uncertainty through an evidential meta-model that explicitly learns when and how to be uncertain. By identifying the most salient layers of a pretrained, frozen model, GUIDE can construct an evidential meta-model that uses these features and trains on a noise-driven curriculum to teach the meta-model to express uncertainty proportionally. This strategy does not inherit the base model's overconfidence, enables uncertainty estimation, and allows the base model to be used in downstream tasks, where typical uncertainty estimation techniques operate at an intrusive level. Our empirical evaluation using both intrusive and post-hoc evidential-based uncertainty quantification methods [29], [34] using both near-out-of-distribution (OOD), far-OOD, and adversarially attacked data on CIFAR100, SVHN, and DeepWeeds, shows GUIDE's outperformance in AUROC (above 94.85%), ID ($\approx 87\%$), OOD ($\leq 8\%$), and adversarial ($\leq 5\%$) data coverage.

Third, Conflict-aware Evidential Deep Learning (C-EDL) [35] builds upon the evidential deep learning framework via conflict estimation in prediction estimates. C-EDL generates multiple label-preserving transformations for each input data and quantifies their representational disagreement through intra- and inter-class conflict measures. This resulting conflict score adjusts the pretrained evidential model's uncertainty estimates to reflect higher uncertainty when conflict is high. Our experimental evaluation shows that C-EDL significantly outperforms state-of-the-art EDL variants [29] and competitive baselines, achieving substantial reductions in coverage for OOD data (up to $\approx$55%) and adversarial data (up to $\approx$90%), in several datasets, attack types, and uncertainty metrics.

B. Context-Enhanced Inference Engine

1) Multi-modal Visual Sensing: Multi-modal systems, such as RGB-Infrared sensors, offer a defense by exploiting complementary sensor failure modes: RGB cameras work well in good light, while infrared (IR) sensors function in low-light and resist adversarial attacks [36]. This enables fallback mechanisms, enhancing trust and security. We propose an adversarial framework for testing multimodal detection with Gaussian noise, mimicking sensor degradation. As shown in Figure 3, our defense uses IoU-based fusion, where infrared predictions override RGB if they overlap by over 50%, offering a fallback mechanism.

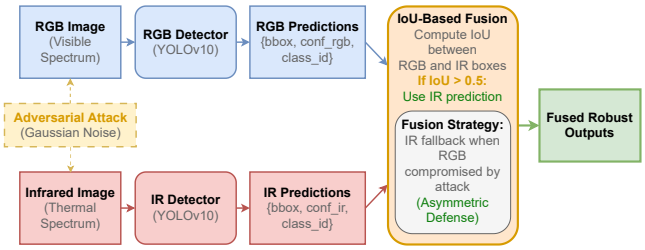

Fig. 3. Multimodal Object Detection Pipeline: IoU based fallback mechanism allows maintaining performance under adversarial conditions.

Evaluation using YOLOv10 on the DroneVehicle dataset [37] shows that multimodal sensing improves the performance (mAP) from 39.87% to 53.31% in unperturbed scenarios. In the presence of Gaussian noise, mAP drops to 16.23% in the case of RGB-only detection, whereas RGB+IR sensing achieves a score of 20.89%. While both sensing techniques degrade in performance due to perturbations, multimodal sensing still outperforms RGB-only sensing.

2) Cross-View Geo-Localization: Autonomous systems face risks from degraded GNSS signals caused by interference, urban occlusion, or environmental issues, jeopardizing localization reliability [38]. To address this, a vision-based fallback mechanism using Cross-View Geo-Localization (CVGL) aligns camera views with satellite imagery to refine or replace GPS positions [39]. This approach supports ground vehicles and aerial platforms, providing a scalable, infrastructure-free solution to GPS spoofing and drift.

The core CVGL pipeline has two stages. First, a coarse pose estimate from GNSS or map data fetches a satellite image patch centered on the vehicle or drone's approximate location, retrieved online or onboard. Second, the system extracts dense feature maps from both the satellite view and onboard camera, then performs feature-based alignment to refine the pose, producing a more robust 3DoF estimate (x, y, yaw) that is less susceptible to spoofing and environmental changes.

In the vehicle-to-satellite setting [40], the vehicle captures a query image with a front-facing monocular camera. After getting an initial position from GNSS, SLAM, or another odometry source, the system selects a satellite tile centered on that estimate. Both the satellite tile and vehicle image pass through feature extractors, typically convolutional or transformer-based encoders, which produce descriptors that capture both structural and semantic cues. These descriptors are then processed by a geometric alignment backend, which may use optimization or learned pose regression. The refined transformation corrects drift and provides a globally accurate pose, eliminating reliance on GNSS alone. Evaluation on the FORD dataset showed the CVGL pipeline consistently improved localization accuracy with different initial errors. When the initial GPS error was $15\ m$, it reduced the final error by about 20%; for 10 m, the improvement was 50%; and for 8 m, over 66%.

In the drone-to-satellite scenario [41], the system processes aerial images captured by UAV-mounted cameras flying at varying altitudes and orientations. While the general CVGL pipeline remains similar to that of ground vehicles, the nature of drone imagery introduces distinct characteristics. Specifically,

drone views often align more closely with a nadir perspective, meaning they fit more naturally with the satellite's top-down geometry. This reduces the severity of perspective distortion, allowing for more efficient geometric transformations between image frames. As a result, the mapping from drone image pixels to satellite image coordinates becomes more constrained and computationally tractable, facilitating faster and more stable convergence during pose refinement. Nonetheless, challenges persist due to variations in illumination, scale shifts, and sensor-specific differences. To address these, the pipeline includes a visual normalization stage that applies color transfer to reduce domain gaps between UAV and satellite imagery. Initial UAV data tests showed the method greatly enhances localization accuracy, reducing yaw error by about 64% and total translation error by 26% from an initial 10m.

3) Secure and Collaborative Context Awareness: GuardAI extends perceptual hashing to CVGL, using fuzzy hashing to maintain visual consistency and verify onboard and reference images to ensure data integrity. Compact hashes detect tampering and spoofing, enhancing efficiency and robustness against perception attacks. Hash-based similarity accelerates secure image registration via feature matching. This approach broadens from tampering detection to trusted inference, improving localization and situational awareness. In multi-agent systems such as CAVs, UAVs, and edge nodes, Federated Swarm Learning shares data without revealing raw information, thereby maintaining privacy. A Dynamic Situational Awareness Scoring System (SASS) evaluates autonomous perception, decision-making, and coordination through machine indicators such as responsiveness and reliability. Each agent creates an internal score, which is aggregated across the swarm to assess overall situational awareness, resilience, and consensus.

C. Defense Recommender System

To implement security-by-design in dynamic edge environments, GuardAI develops a *defense recommender system*. The system uses a continuously updated knowledge base of attacks and corresponding defense mechanisms. Each entry is defined by data modality (e.g., visual, temporal), application constraints (e.g., latency, energy budget), and operational context (e.g., mobility, connectivity). The recommender will use utilize techniques to simulate attack scenarios, assess defensive responses, and rank their effectiveness under different edge constraints. The results will yield a curated dataset for training hybrid recommendation models that combine collaborative filtering, content-based reasoning, and reinforcement learning. During deployment, the system will analyze the current threat profile and system context to recommend the most appropriate defense strategy, balancing robustness, efficiency, and resource use.

V. Use Cases

A. Robust Monitoring with AI-Enabled UAVs

This use case deploys AI-enabled UAVs for real-time monitoring of critical infrastructures [42]. While drones can provide rapid situational awareness and threat detection, their perception pipelines are highly vulnerable to adversarial manipulation—through both physical artifacts (e.g., adversarial elements in the scene) and digital perturbations (e.g., tampered video feeds) [43]. In addition, their communication channels are susceptible to attacks (e.g., jamming), which can prevent them from performing their intended functions, especially in collaborative drone systems [44].

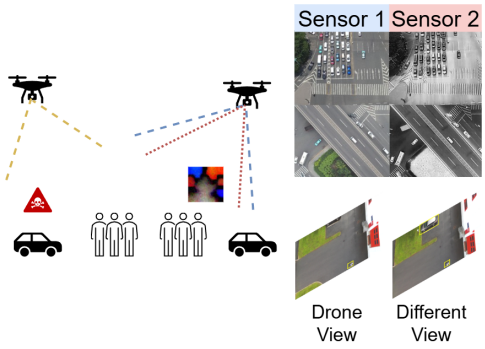

Fig. 4. Use case 1 shows a malicious agent using adversarial patches or digital attacks. Defense strategies include lightweight spatio-temporal methods, multi-sensor fusion, and a module collecting predictions from multiple drones.

GuardAI addresses these challenges by leveraging its core technological pillars (Fig. 4). The *adversarially resilient toolkit* introduces lightweight defenses that purify spatio-temporal data streams before detection, reducing susceptibility to manipulated inputs. It augments that with a network defense layer that protects the communication channels from jamming attacks. The *context-enhanced inference engine* leverages multi-sensor fusion (e.g., camera and infrared) and cross-agent collaboration across multiple drones to enforce spatial and temporal consistency, thereby maintaining robustness even when individual sensors are compromised. Finally, the *defense recommender system* selects suitable defense strategies based on the operational context and threat profile, enabling on-board adaptation under evolving conditions.

Evaluation scenarios will involve UAVs monitoring populated areas under both normal and adversarial conditions. Attacks will be simulated by inserting physical artifacts and digital perturbations, and jamming the drones' communication channels. At the same time, GuardAI modules will be tasked with detecting, mitigating, and recovering from these disturbances. Performance will be measured using: (i) response time from detection to mitigation, (ii) robustness index (successful defense rate), (iii) false positives, and (iv) false negative rates.

B. Protecting decentralized 5G network analytics

This Use Case focuses on the deployment and evaluation of AI-enabled defense mechanisms within decentralized 5G environments. The objective is to investigate resilience of network analytics and intrusion detection against adversarial behavior in multi-domain 5G infrastructures.

GuardAI builds upon an Attack Generation Engine (AGE) and AI-Enabled Intrusion Detection System (AI-IDS), previously developed and validated in real 5G environments, to generate, detect, and mitigate diverse network attacks [45]. The experimentation follows a three-phase process: (i) *attack generation and dataset creation*, where the AGE simulates realistic threats such as Denial-of-Service and protocol-based attacks to

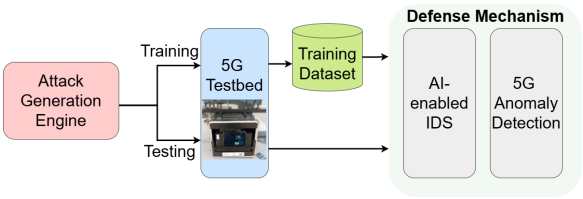

Fig. 5. Use Case 2 for resilient 5G network analytics, combining automated attack generation, AI-based intrusion detection, and adaptive mitigation mechanisms across distributed edge domains.

produce labeled datasets; (ii) *AI-enabled intrusion detection*, where deep learning models are trained on the generated traffic and deployed within the 5G testbed for real-time inference; and (iii) *mitigation and anomaly response*, where detection alerts trigger automated countermeasures to maintain network continuity and integrity. Complementing this setup, a 5G testbed (Fig. 5) is employed that generates both benign and malicious traffic metrics for the training and validation of AI-based anomaly detection models. The resulting models perform real-time inference on live data streams, with detected anomalies visualized through an interactive dashboard to assist operators in decision-making. Within GuardAI, testbeds form a federated experimentation environment that supports the assessment of context-aware AI defenses across distributed 5G infrastructures. Evaluation metrics include: (i) detection accuracy and false alarm rate, (ii) latency between attack occurrence and detection, (iii) mitigation response time, and (iv) system robustness under adversarial stress.

C. Resilient Perception in Connected Autonomous Vehicles

This Use Case tackles the core challenge of learning and operating safely with distributed, scarce, and heterogeneous data generated across remote interacting agents. In real traffic scenarios, the amount of local data is not sufficient to derive resilient and robust learning modalities, while onboard sensors (e.g., camera, radar, LiDAR, ultrasonics, GNSS) degrade under rain, darkness, occlusions, or high speed, and are additionally exposed to cyber-physical attacks (e.g., camera/radar spoofing, LiDAR saturation, GNSS jamming/spoofing, and adversarial perturbations) that can corrupt perception at its source.

GuardAI addresses these challenges by leveraging the *adversarially resilient toolkit* to deliver efficient, model-based defenses for attack detection and mitigation on visual sensors, preventing spoofing and perturbation attempts before they cascade into downstream tasks. Additionally, the *context-enhanced inference engine* leverages multimodal, multi-agent, and cross-view processing (CAV/UAV/satellite) fusing heterogeneous measurements with AI insights from collaborating agents, while its secure collaborative learning framework strengthens the confidentiality of multimodal federated learning by integrating homomorphic encryption and differential privacy. Together, these capabilities: (a) counter single-sensor weaknesses or failures (whether environmental or attack-induced), (b) extend sensing beyond line-of-sight to detect occluded objects, and (c) enhance overall safety, resilience, and robustness via shared, context-aware perception.

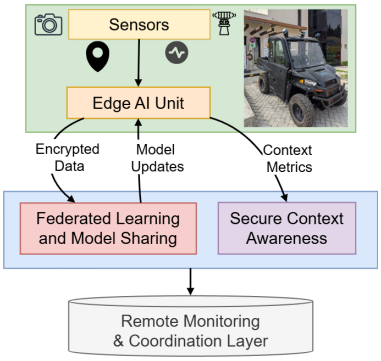

Fig. 6. Use Case 3 with the Polaris Ranger EV equipped with an embedded edge AI unit and multi-sensor suite, used to evaluate secure perception and adaptive defense mechanisms in autonomous operations.

Validation combines real-world and simulated assets, including fully autonomous test vehicles (e.g., Polaris Ranger EV; Fig. 6) equipped with sensing, processing and communication infrastructure, as well as a ROS-integrated CAV simulation stack integrating diverse sensors, realistic traffic environments, and dynamic agent interactions, with the added capability of modelling and simulating multiple attack surfaces (perception attacks, and cyber-physical threats). Indicative scenarios include distributed multi-modal perception, robust cooperative localization, multi-object detection and tracking, and cross-view geo-localization. Performance will be measured using metrics such as Mean Time to Detect, False Positives, and False Negatives.

VI. Conclusion and Future Work

This work provides an overview of the GuardAI project and its technologies designed to improve robustness, security, and trustworthiness in edge AI systems. GuardAI combines adversarial resilience, context-enhanced inference, and security-by-design paradigm into a unified stack that spans the sensing, inference, and defense layers. Its technologies demonstrate how resilience can be effectively embedded in resource-constrained environments. Through representative Use Cases involving UAV-based monitoring, decentralized 5G analytics, and CAVs, GuardAI shows that adaptive robustness and efficient inference are not mutually exclusive. Instead, they can develop together when defense mechanisms are co-designed with perception and control pipelines. This convergence of robustness, efficiency, and security creates a foundation for pursuing certifiable AI operation in safety-critical settings. Future work will focus on expanding GuardAI's adaptive defense orchestration using reinforcement learning to enable dynamic strategy selection across heterogeneous edge devices. We will develop evaluation protocols that connect resilience metrics with compliance requirements of emerging AI frameworks (EU AI Act, ISO/IEC JTC 1/SC 42, CEN/CLC JTC 21), and establish benchmarks for edge AI security across the three use cases.

Acknowledgment

This work was supported the European Union's Horizon Europe research and innovation programme under grant agreement 101168067 (GuardAI - Enhancing Robustness and Security of Edge AI Systems for Safety-Critical Applications).

REFERENCES

[1] T. Ahmad, A. Morel, N. Cheng, K. Palaniappan, P. Calyam, K. Sun, and J. Pan, "Future UAV/Drone Systems for Intelligent Active Surveillance and Monitoring," *ACM Comput. Surv.*, vol. 58, Sept. 2025.

[2] M. Kamal, C. Kyrkou, N. Piperigkos, A. Papandreou, A. Kloukiniotis, J. Casademont, N. P. Mateu, D. B. Castillo, R. D. Rodriguez, N. G. Durante, P. Hofmann, P. Kapsalas, A. S. Lalos, K. Moustakas, C. Laoudias, T. Theocharides, and G. Ellinas, "A comprehensive solution for securing connected and autonomous vehicles," in *2022 Design, Automation & Test in Europe Conference & Exhibition (DATE)*, pp. 790–795, 2022.

[3] X. Wang, B. Wang, Y. Wu, Z. Ning, S. Guo, and F. R. Yu, "A Survey on Trustworthy Edge Intelligence: From Security and Reliability to Transparency and Sustainability," *IEEE Communications Surveys & Tutorials*, vol. 27, no. 3, pp. 1729–1757, 2025.

[4] C. Kyrkou, A. Papachristodoulou, A. Kloukiniotis, A. Papandreou, A. Lalos, K. Moustakas, and T. Theocharides, "Towards Artificial-Intelligence-Based Cybersecurity for Robustifying Automated Driving Systems Against Camera Sensor Attacks," in *2020 IEEE Computer Society Annual Symposium on VLSI (ISVLSI)*, pp. 476–481, 2020.

[5] J. Peck, B. Goossens, and Y. Saeys, "An Introduction to Adversarially Robust Deep Learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 4, pp. 2071–2090, 2024.

[6] H. Zhang, L. Xiao, X. Cao, and H. Foroosh, "Multiple Adverse Weather Conditions Adaptation for Object Detection via Causal Intervention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 3, pp. 1742–1756, 2024.

[7] G. Pang, C. Shen, L. Cao, and A. V. D. Hengel, "Deep Learning for Anomaly Detection: A Review," *ACM Comput. Surv.*, vol. 54, Mar. 2021.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," 2014.

[9] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2019.

[10] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," 2020.

[11] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, "Adversarial training for free!," 2019.

[12] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, "Pruning and quantization for deep neural network acceleration: A survey," 2021.

[13] Y. Fukuda, K. Yoshida, and T. Fujino, "Evaluation of model quantization method on vitis-ai for mitigating adversarial examples," *IEEE Access*, vol. 11, pp. 87200–87209, 2023.

[14] P. Austria, E. Begoli, and A. Sadovnik, "The effects of compounded model size reductions on adversarial robustness," in *2025 IEEE 18th Dallas Circuits and Systems Conference (DCAS)*, pp. 1–6, 2025.

[15] K. Nguyen, T. Fernando, C. Fookes, and S. Sridharan, "Physical adversarial attacks for surveillance: A survey," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 12, pp. 17036–17056, 2024.

[16] W. Nie, B. Guo, Y. Huang, C. Xiao, A. Vahdat, and A. Anandkumar, "Diffusion models for adversarial purification," in *Proceedings of the 39th International Conference on Machine Learning* (K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, eds.), vol. 162 of *Proceedings of Machine Learning Research*, pp. 16805–16827, PMLR, 17–23 Jul 2022.

[17] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[18] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *International Conference on Learning Representations*, 2021.

[19] M. Hill, J. Mitchell, and S.-C. Zhu, "Stochastic security: Adversarial defense using long-run dynamics of energy-based models," *arXiv preprint arXiv:2005.13525*, 2020.

[20] A. Gkillas, A. S. Lalos, and D. Ampeliotis, "An efficient deep unrolling super-resolution network for lidar automotive scenes," in *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 1840–1844, 2023.

[21] Y. Zhang, J. Hou, and Y. Yuan, "A Comprehensive Study of the Robustness for LiDAR-Based 3D Object Detectors Against Adversarial Attacks," *Int. J. Comput. Vision*, vol. 132, p. 1592–1624, Nov. 2023.

[22] S. Seoni, V. Jahmunah, M. Salvi, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023)," *Computers in Biology and Medicine*, vol. 165, p. 107441, 2023.

[23] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *International Journal of Computer Vision*, vol. 132, no. 12, pp. 5635–5662, 2024.

[24] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine learning*, vol. 110, no. 3, pp. 457–506, 2021.

[25] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, pp. 1050–1059, PMLR, 2016.

[26] H. Chen, Z. Huang, H. Lam, H. Qian, and H. Zhang, "Learning prediction intervals for regression: Generalization and calibration," in *International Conference on Artificial Intelligence and Statistics*, pp. 820–828, PMLR, 2021.

[27] C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther, "Ladder variational autoencoders," *Advances in neural information processing systems*, vol. 29, 2016.

[28] A. N. Angelopoulos, S. Bates, *et al.*, "Conformal prediction: A gentle introduction," *Foundations and Trends® in Machine Learning*, vol. 16, no. 4, pp. 494–591, 2023.

[29] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," *Advances in neural information processing systems*, vol. 31, 2018.

[30] D. Bethell, S. Gerasimou, and R. Calinescu, "Robust uncertainty quantification using conformalised monte carlo prediction," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 20939–20948, 2024.

[31] A. Angelopoulos, S. Bates, J. Malik, and M. I. Jordan, "Uncertainty sets for image classifiers using conformal prediction," *arXiv preprint arXiv:2009.14193*, 2020.

[32] Y. Romano, E. Patterson, and E. Candes, "Conformalized quantile regression," *Advances in neural information processing systems*, vol. 32, 2019.

[33] C. Barker, D. Bethell, and S. Gerasimou, "Guided uncertainty learning using a post-hoc evidential meta-model," *arXiv preprint arXiv:2509.24492*, 2025.

[34] M. Shen, Y. Bu, P. Sattigeri, S. Ghosh, S. Das, and G. Wornell, "Post-hoc uncertainty learning using a dirichlet meta-model," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 9772–9781, 2023.

[35] C. Barker, D. Bethell, and S. Gerasimou, "Quantifying adversarial uncertainty in evidential deep learning using conflict resolution," *arXiv preprint arXiv:2506.05937*, 2025.

[36] Z. Liu, J. Liu, B. Zhang, L. Ma, X. Fan, and R. Liu, "Paif: Perception-aware infrared-visible image fusion for attack-tolerant semantic segmentation," in *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, (New York, NY, USA), p. 3706–3714, Association for Computing Machinery, 2023.

[37] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[38] A. Altaweel, H. Mukkath, and I. Kamel, "Gps spoofing attacks in fanets: A systematic literature review," *IEEE Access*, vol. 11, pp. 55233–55280, 2023.

[39] Y. Shi and H. Li, "Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17010–17020, 2022.

[40] C. Anagnostopoulos, A. Gkillas, N. Piperigkos, and A. S. Lalos, "Personalized federated learning for cross-view geo-localization," in *2024 IEEE 26th International Workshop on Multimedia Signal Processing (MMSP)*, pp. 1–6, 2024.

[41] J. Fan, E. Zheng, Y. He, and J. Yang, "A cross-view geo-localization algorithm using uav image and satellite image," *Sensors*, vol. 24, no. 12, 2024.

[42] A. Duke, R. Shimmon, R. Wiseman, H. Zheng, D. Yue, A. Seghers, E. Anastassacos, C. McMaster, A. Mendez, P. Moorcroft, S. Laouici, T. Tideswell, and C. Lyons, "Drones-as-a-service for efficient critical national infrastructure operations: Reducing the time from image capture to insight generation," in *2024 International Conference on Unmanned Aircraft Systems (ICUAS)*, pp. 421–429, 2024.

[43] S. M. K. A. Kazmi, N. Aafaq, M. A. Khan, M. Khalil, and A. Saleem, "From pixel to peril: Investigating adversarial attacks on aerial imagery through comprehensive review and prospective trajectories," *IEEE Access*, vol. 11, pp. 81256–81278, 2023.

[44] J.-P. Yaacoub, H. Noura, O. Salman, and A. Chehab, "Security analysis of drones systems: Attacks, limitations, and recommendations," *Internet of Things*, vol. 11, p. 100218, 2020.

[45] S. Kalafatidis, G. Agrafiotis, K. Giapantzis, A. Lalas, and K. Votis, "Experiments with digital security processes over sdn-based cloud-native 5g core networks," in *2024 27th Conference on Innovation in Clouds, Internet and Networks (ICIN)*, pp. 97–99, IEEE, 2024.