In: 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 05-08 October 2025, Vienna, Austria, pp. 6295-6300
DOI: 10.1109/SMC58881.2025.11343517

Generative Adversarial Evasion and Out-of-Distribution Detection for UAV Cyber-Attacks

Deepak Kumar Panda[1] and Weisi Guo[1]

***Abstract*— As UAVs are increasingly integrated into civilian airspace, the need for resilient intrusion detection system (IDS) frameworks grow, as traditional anomaly detection methods often struggle to detect novel threats. A common strategy is to treat the unfamiliar attacks as out-of-distribution (OOD) samples; hence, inadequate mitigation responses can leave systems vulnerable, granting adversaries the capability to cause potential damage. Furthermore, conventional OOD detectors frequently fail to discriminate the stealthy adversarial attacks from OOD samples. This paper proposes a conditional generative adversarial network (cGAN)-based framework specifically designed to craft stealthy adversarial attacks that effectively evade IDS mechanisms. Initially, we construct a robust multi-class classifier as IDS which classifies the benign UAV telemetry data from known cyber-attack types, including Denial of Service (DoS), false data injection (FDI), man-in-the-middle (MiTM), and replay attacks. Leveraging this classifier, our proposed cGAN strategically perturbs known attack features, generating sophisticated adversarial samples engineered to evade detection through benign misclassification. Then, the generative stealthy adversarial samples are refined to match the distribution of the OOD samples while ensuring high attack success. To effectively detect these stealthy adversarial perturbations, a conditional variational autoencoder (CVAE) is implemented, using negative log-likelihood as a metric to distinguish adversarial samples from genuine OOD samples. Comparative analyses between CVAE-based regret analysis and traditional Mahalanobis distance-based detectors demonstrate that the CVAE's negative log-likelihood significantly outperforms in detecting stealthy adversarial attacks from OOD samples. Our findings highlight the necessity of advanced probabilistic modeling techniques to reliably detect and adapt the existing IDS against novel, generative-model-based stealthy cyber threats.**

I. INTRODUCTION

Unmanned Aerial Vehicles (UAVs) are increasingly employed in civilian applications, including aerial surveillance, precision agriculture, and logistics [1]. As critical components of emerging urban air mobility (UAM) systems, UAVs rely on networked communication and onboard autonomy, rendering them vulnerable to sophisticated cyber-attacks [2]. Adversaries may intercept telemetry data and inject falsified information by gaining unauthorized access to communication channels, thereby obstructing the UAM operator's ability to verify compliance with planned flight paths as shown in Fig. 1. Failing to detect adversarial attacks as general anomalies can allow malicious actors to bypass security measures and inflict serious harm on airspace operations, leading to breaching of restricted or unauthorized zones. Such violations can endanger other airspace users and increase the risk of hijacking or unintentional exposure of sensitive information. Therefore, enhancing the security of UAV systems is crucial—especially in dynamic, time-critical operational environments. Traditional security methods such as cryptography and anomaly-based IDS struggle with latency, high false positives, and poor generalization to novel threats [3]. Generative adversarial networks (GANs) have emerged as a powerful tool for crafting polymorphic cyber-attacks that evade detection. Prior work has demonstrated their use in generating adversarial examples that mimic legitimate data distributions while compromising classifier performance [4, 5]. However, distinguishing such stealthy adversarial samples from natural out-of-distribution (OOD) samples remains a major challenge. Adversaries can exploit this weakness by generating perturbations that resemble benign anomalies, thereby bypassing anomaly detection-based IDS. An inability to differentiate out-of-distribution (OOD) samples from adversarial attacks can undermine the effectiveness of conventional filtering methods, such as the extended Kalman filter (EKF) [6], which may struggle to detect stealthy adversarial inputs due to their statistical resemblance to the assumed OOD features.

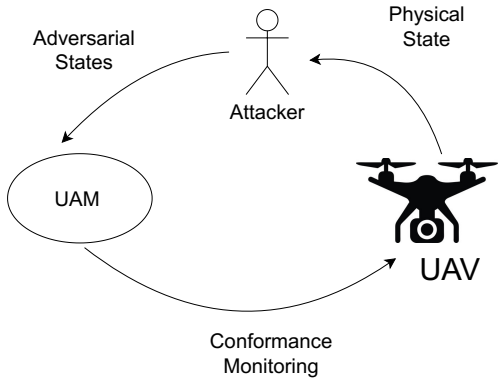

Fig. 1. The attacker can manipulate the telemetry information sent by UAV to the UAM operator, which can influence the conformance monitoring for the UAV.

Statistical methods for adversarial detection [7, 8] often assume fixed data distributions and lack adaptability to high-dimensional stealthy attacks. Moreover, they fail to model vulnerabilities in the latent space, limiting their effectiveness in detecting unseen or well-crafted adversarial inputs. To

This work was supported by the Royal Academy of Engineering and the Office of the Chief Science Adviser for National Security under the UK Intelligence Community Postdoctoral Research Fellowship programme

[1]The authors are with Faculty of Engineering and Applied Sciences, Cranfield University, MK43 0AL Cranfield, U.K Deepak.Panda@cranfield.ac.uk, weisi.guo@cranfield.ac.uk

Published by IEEE. This is the Author Accepted Manuscript issued with: Creative Commons Attribution License (CC:BY 4.0).
The final published version (version of record) is available online at DOI:10.1109/smc58881.2025.11343517. Please refer to any applicable publisher terms of use.

overcome these limitations, deep generative models such as variational autoencoders (VAEs) have been proposed [9]. VAEs capture probabilistic latent representations that support better anomaly detection. However, their generalization ability is limited in the presence of multiple attack types. Conditional VAEs (CVAEs) [10], which incorporate attack labels into the learning process, enhance discriminative power and robustness. Unlike prior works that use GANs and CVAEs separately, we present a unified framework that crafts stealthy perturbations aligned with attack features and detects them using negative log-likelihood–based distributional distinctions. The key contributions are summarized as:

- **Stealthy Adversarial Attack Generation:** We design a conditional GAN that perturbs features from known cyberattacks (DoS, FDI, MiTM and replay) to produce adversarial inputs that evade a multi-class IDS by mimicking benign behavior. We optimize the stealthy adversarial attacks iteratively to ensure statistical similarity with OOD samples while achieving high evasion success.
- **Probabilistic Detection via CVAE:** To detect such stealthy attacks from OOD samples, we implement a CVAE-based detection system. Using negative log-likelihood as a detection score, our method outperforms traditional Mahalanobis-based detectors [7] and advanced regret-based CVAE [11] baselines in distinguishing stealthy attacks from genuine OOD samples.

II. Overall Methodology and Dataset Description

A. Methods

As illustrated in Fig. 2, the UAV cyber-physical dataset comprises benign and attacked feature sets, denoted by X_{ben} and X_{att} respectively. An IDS, designed as deep feedforward neural network $f_{\text{IDS}}(\cdot)$, is trained and evaluated as a multi-class classifier. Subsequently, a cGAN is designed, consisting of a conditional generator network G and a discriminator D. The generator G learns to craft adversarial perturbations to the input attack features such that the classifier $f_{\text{IDS}}(\cdot)$ incorrectly labels them as benign. Meanwhile, the discriminator D functions as a binary classifier, distinguishing real benign samples from adversarially generated ones. This adversarial interplay compels the generator to produce increasingly stealthy and harder-to-detect perturbations. The perturbations are iteratively refined in N_{ref} steps, to produce adversarial samples X_{adv} which is statistically similar to the OOD features, but with high attack success rate. CVAE, pre-trained on both benign and attack-labeled data, is then used to compute the negative log-likelihood (NLL) for the adversarial X_{adv} and OOD samples X_{ood}. Based on the negative log-likelihood distrbution, we classify stealthy adversarial samples from OOD.

B. Cyber Attack Dataset

The UAV cyber-attack dataset utilized in this study is based on the intrusion detection framework developed by Hassler et al. [12], where a comprehensive experimental setup was employed. The dataset is comprised of UAV telemetry obtained after several types of cyberattacks, briefly described as follows:

- **De-authentication:** Spoofed disconnects cause safety mode triggers.
- **Replay:** Recorded command signals are replayed to cause erratic UAV behavior.
- **Evil Twin:** A rogue access point enables MiTM attacks.
- **False Data Injection:** Sensor state manipulation misleads UAV controllers.

These attacks exhibit clear manifestations across both cyber and physical domains. Regarding the cyber features, only those attributes common across all attack types were retained to ensure consistency. These include: timestamp, frame number, frame length, transmission duration over the wireless medium, frame sequence number, control type and subtype, and the WLAN fragment number. In contrast, the physical features encompass various UAV telemetry data, such as altitude; velocities along the x, y, z axes pitch, roll, yaw; flight distance and time; pressure; battery status; and spatial distances along the x, y, z axes relative to the Tello Pad. Details of the experiment is provided in [12].

Given the availability of this labeled dataset, it becomes feasible to design a multi-class classification model capable of distinguishing both benign and malicious behavior and distinct attack types. However, if an adversary gains access to the labeled attack data, they can leverage this information to construct conditional GAN-based models that generate stealthy adversarial examples. These examples, synthesized from known attack features, are strategically crafted to evade existing IDS.

III. Stealthy Adversarial Attack using GAN

This section outlines the design of a conditional generative adversarial network (cGAN) framework aimed at generating adversarial attacks against a neural network-based (IDS) $f_{\text{IDS}}(\cdot)$. The cGAN architecture comprises two neural networks: a generator $G(\mathfrak{z})$, which synthesizes adversarial perturbations from injected noise, and a discriminator, $D(x)$ which learns to differentiate between benign and adversarially perturbed inputs. These networks are trained via adversarial (min-max) optimization, where the generator seeks to fool the discriminator, while the discriminator simultaneously improves its ability to distinguish between authentic and adversarial samples. The objective function is given by,

$$\min_G \max_F \mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)] + \mathbb{E}_{\mathfrak{z}\sim p_{\mathfrak{z}}}[\log(1 - D(G(\mathfrak{z})))]. \tag{1}$$

Here, x represents benign data sample, $\mathfrak{z}$ represents the random noise vector, $G(\mathfrak{z})$ generates adversarial perturbations and $D(x)$ outputs the probability that x is benign. For solving the objective described in (1), the conditional generator and discriminator are described in the next subsection.

A. Conditional Generator

The generator is modeled as a feedforward neural network designed to transform a noise vector into adversarial

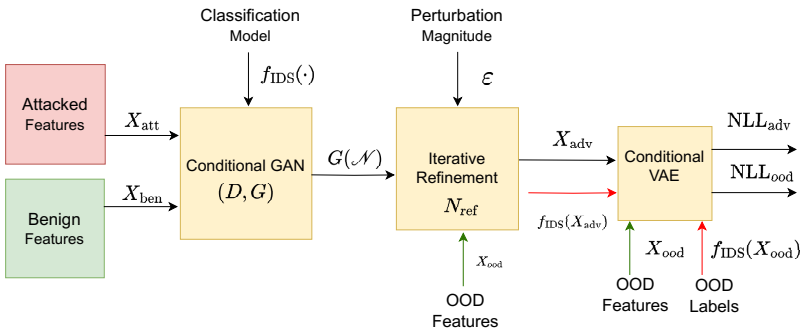

Fig. 2. Schematic for obtaining negative likelihood (NLL) to detect stealthy adversarial attacks from OOD samples.

feature representations. Specifically, it maps input noise to adversarial features that are misclassified as benign by the intrusion detection system (IDS). Formally, the generator is defined as a function $G : \mathbb{R}^{d_{\text{noise}}} \rightarrow \mathbb{R}^{d_{\text{input}}}$, where d_{noise} denotes the dimensionality of the input noise vector $\mathfrak{z}$, and d_{input} corresponds to the dimensionality of the cyber-physical feature space of the UAV system. The generator has two primary objectives: (1) to induce misclassification in the target IDS, thereby maximizing its prediction error, and (2) to ensure that the generated adversarial features closely resemble legitimate benign samples in feature space. Let $\delta = G(\mathfrak{z})$, represent the adversarial perturbation generated from a noise vector $\mathfrak{z}$. The adversarial input X_{adv} is then constructed by perturbing the original (attacked) feature vector with 2-norm. Let $f_{\text{IDS}}(X_{\text{adv}})$ denote the classification probabilities produced by the IDS when presented with adversarial input, and let y_{ben} represent the ground-truth label distribution for benign samples. Hence, the total loss for the cGAN generator is given by,

$$\begin{aligned}\mathscr{L} &= \lambda_{\text{cls}} \cdot \mathscr{L}_{\text{cls}} + \lambda_{\text{st}} \cdot \mathscr{L}_{\text{st}} + \lambda_G \cdot \mathscr{L}_G, \\ &= \lambda_{\text{cls}} \cdot \frac{1}{N} \sum_{i=1}^{N} f_{\text{IDS}}(X_{\text{adv}})_i \log \frac{f_{\text{IDS}}(X_{\text{adv}})_i}{y_{\text{ben}}} + \\ &\lambda_{\text{st}} \cdot \frac{1}{N} \sum_{i=1}^{N} \|X_{\text{adv}} - X_{\text{ben}}\|_2^2 - \lambda_G \cdot \mathbb{E}\left[\log\left(D(X_{\text{adv}})\right)\right].\end{aligned} \tag{2}$$

where λ_{cls}, λ_{st} and λ_G represent hyperparameters controlling the misclassification loss, stealth loss and generator loss. The first term in (2), refers to the Kullback-Leibler (KL) divergence loss, which guides the generator towards producing benign-like predictions. The second terms, introduces a feature similarity loss, which encourages the adversarial features to remain similar to the benign features in the Eucleadian space. The third term is a part of standard cGAN formulation to help produce more benign-like samples.

B. Discriminator

The discriminator D is trained to distinguish between benign and adversarial samples, hence binary cross entropy loss is used to differentiate it, which is presented as follows

$$\mathscr{L}_D = -\frac{1}{2}\left(\mathbb{E}_{x \sim X_{\text{ben}}}[\log D(x)] - \mathbb{E}_{x_{\text{adv}} \sim G}[\log(1 - D(x_{\text{adv}}))]\right). \tag{3}$$

During the cGAN training, we assume that the attacker has the access to the IDS via an external API. The cGAN can be trained in a standard manner as shown in [13].

C. Iterative Refinement for Stealthy Attack

Iterative refinement strategy is analogous to the one in projected gradient descent (PGD) attacks [14], where attack is injected in multi-steps so that the overall perturbations stay within a certain bound. The aim is to make the adversarial attack X_{adv} close to the OOD sample X_{ood} generated from normally distributed noise with the noise scale ρ, defined as $X_{\text{ood}} = X_{\text{att}} + \mathcal{N}\left(\mathbf{0}, \rho^2 \mathbf{I}\right)$. Hence, to ensure the stealth nature of the adversarial attack, X_{adv} needs to be distributionally similar to X_{ood}, with high attack success rate. Hence, we define the stealthy adversarial attacks as follows:

Definition 1: Let $X \subset \mathbb{R}^d$ denote the input feature representing UAV telemetry data for IDS. Let us define P_{ben}, P_{ood}, P_{adv} as the distribution of the benign UAV , OOD and adversarial attacked features respectively. Let us define $\mathscr{W}(\cdot,\cdot)$ as the Wasserstein distance measuring the distributional distance, hence we define the stealthy adversarial attack if it satisfies the following conditions:

$$\begin{aligned}&\mathscr{W}(P_{\text{adv}}, P_{\text{ben}}) \approx \mathscr{W}(P_{\text{ood}}, P_{\text{ben}}), \\ &\mathbb{P}\left[f_{\text{IDS}}(X_{\text{adv}}) \neq f_{\text{IDS}}(X_{\text{ben}})\right] \gg \mathbb{P}\left[f_{\text{IDS}}(X_{\text{ood}}) \neq f_{\text{IDS}}(X_{\text{ben}})\right].\end{aligned} \tag{4}$$

If the attacker, wants to induce perturbations of maximum magnitude ε, the definition of final adversarial attack features X_{adv} under iterative refinement is given as:

$$\begin{aligned}X_{\text{adv}}^{N_{\text{ref}}} &= X_{\text{att}} + \sum_{i=0}^{N_{\text{ref}}} \min\left(\max\left(X_{\text{adv}}^i + G(z,c) - X_{\text{att}}, -\varepsilon\right), \varepsilon\right), \\ X_{\text{adv}}^0 &= X_{\text{att}}.\end{aligned} \tag{5}$$

If N_{att} and $N_{\text{adv}}^{\text{ben}}$ represent the number of attacked and number of adversarial features classified as benign respectively, then the attack success rate is given by, $\eta_{\text{succ}} = N_{\text{adv}}^{\text{ben}} / N_{\text{att}}$. Hence, to compute the refinements N_{ref} for optimum stealthy attack as per Definition 1, the objective function can be framed as,

$$\begin{aligned}&\arg\min_{N_{\text{ref}}} \mathscr{W}\left[\mathscr{W}(X_{\text{adv}}, X_{\text{att}}), \mathscr{W}(X_{\text{ood}}, X_{\text{att}})\right], \\ &\text{s.t. } \eta_{\text{succ}} \geq \eta_{\max}.\end{aligned} \tag{6}$$

The whole algorithm for generating the attack using iterative refinement is given as follows:

Algorithm 1 Iterative Refinement of Stealthy Adversarial Samples

Require: Attack input X_{att}, generator $G(z,c)$, bound ε, refinement steps N_{ref}

Initialize $X_{\text{adv}}^{(0)} \leftarrow X_{\text{att}}$
for $i = 1$ to N_{ref} **do**
  Sample noise $z \sim \mathcal{N}(0, I)$, label c
  Generate perturbation: $\delta = G(z,c)$
  Apply bounded update:

$$X_{\text{adv}}^{(i)} \leftarrow \text{clip}\left(X_{\text{adv}}^{(i-1)} + \delta, X_{\text{att}} - \varepsilon, X_{\text{att}} + \varepsilon\right)$$

end for
return $X_{\text{adv}}^{(N_{\text{ref}})}$

IV. Conditional Variational Autoencoder for Stealthy Attack Detection

CVAEs [10] represent a powerful class of deep probabilistic generative models widely applied across various practical domains. These models incorporate a latent variable $\mathbf{z}$, sampled from prior distribution $p(\mathbf{z})$, and conditional distribution $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{c})$ with respect to conditional label $\mathbf{c}$ to generate the observed variable $\mathbf{x}$. However, obtaining the marginal likelihood $p_\theta(\mathbf{x})$ directly, becomes computationally intractable in high-dimensional settings due to the integration over the latent space. To address this challenge, variational inference is employed to approximate the posterior distribution and derive a tractable lower bound on the log-likelihood of the observed data. This results in the well-known evidence lower bound (ELBO), which serves as the objective function for training the model:

$$\begin{aligned}\log p_\theta(\mathbf{x}) &\geq \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z},\mathbf{c})\right] - D_{\text{KL}}\left[q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})\|p(\mathbf{z})\right] \\ &\triangleq \mathscr{L}(\mathbf{x};\theta,\phi).\end{aligned} \tag{7}$$

Here, $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{c})$ denotes the variational approximation to the true posterior distribution $p_\theta(\mathbf{z}\,|\,\mathbf{x})$, where ϕ and θ represent the parameters of the encoder and decoder networks, respectively. These components are typically implemented using neural networks. The CVAE is trained by minimizing the variational objective $\mathscr{L}(\mathbf{x};\theta,\phi)$ over the training data. However, the CVAE does not directly maximize the true likelihood; instead, it optimizes the ELBO, as defined in (7). This approximation can lead to learned representations that are overly constrained or limited in expressiveness, as only a limited subset of samples are likely to fall within regions of high posterior probability under the approximate posterior. To address this limitation, the importance-weighted autoencoder (IWAE) [15] is employed. The IWAE introduces a tighter lower bound on the marginal log-likelihood $\log p(\mathbf{x})$ by using a $k-$ sample importance-weighted estimate, thereby enabling a more expressive generative model.

$$\mathscr{L}_k(\mathbf{x}) = \mathbb{E}_{\mathbf{z}_1,\cdots,\mathbf{z}_k\sim q(\mathbf{z}|\mathbf{x},\mathbf{c})}\left[\log \frac{1}{k}\sum_{i=1}^{k}\frac{p(\mathbf{x},\mathbf{z}_i)}{q(\mathbf{z}_i|\mathbf{x},\mathbf{c})}\right]. \tag{8}$$

Here $\mathbf{z}_1,\cdots,\mathbf{z}_k$ are sampled independently from the encoder network. The term inside the sum corresponds to unnormalized importance weights for joint distribution, denoted as $w_i = p(\mathbf{x},\mathbf{z_i})/q(\mathbf{z}_i\,|\,\mathbf{x})$. From Jensen's inequality, it can be shown that average importance weights are an unbiased estimator of $p(\mathbf{x})$ given as,

$$\mathscr{L}_k = \mathbb{E}\left[\log \frac{1}{k}\sum_{i=1}^{k} w_i\right] \leq \log \mathbb{E}\left[\frac{1}{k}\sum_{i=1}^{k} w_i\right] = \log p(\mathbf{x}). \tag{9}$$

The weight computation using importance sampling for a particular datapoint is given by,

$$\begin{aligned} w &= \log p(x|z,c) + \log p(z) - \log q(z|x,c). \\ &= \sum_{i=1}^{b}\sum_{j=1}^{d} x_{i,j}\log \hat{x}_{i,j} + \log p(z) - \sum_{j=1}^{d}\left(\frac{z_j^2}{2} + \frac{\log(2\pi)}{2}\right) \\ &\quad - \sum_{j=1}^{d}\left(\frac{(z_j-\mu_j)^2}{2\sigma_j^2} + \frac{\log(2\pi)}{2} + \frac{\log\sigma_j^2}{2}\right). \end{aligned} \tag{10}$$

The first term represents the likelihood from the latent feature z and conditional label c. The second term, represents the standard normally distributed prior for the latent variables z. The third term, represents the variational posterior for the latent variables z, given data x and conditional label c. The negative log-likelihood (NLL) for the given sample $\mathbf{x}$ is given by:

$$\text{NLL} = -\left(\log \mathbb{E}_{q(z|x,c)}\left[e^{w-\max(w)}\right] + \max(\text{w})\right). \tag{11}$$

The NLL, as shown in (11), is used to compute for the adversarial features X_{adv} and OOD features X_{ood}.

V. Results and Discussion

A. Numerical Implementation and Training Results

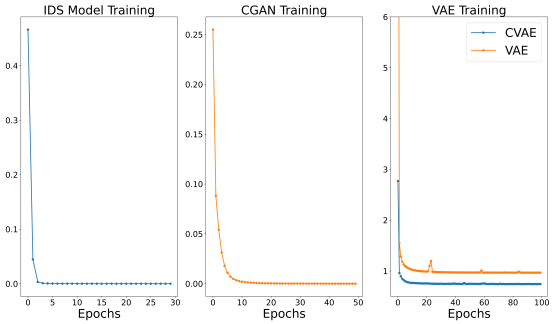

Fig. 3. Loss function with the training epochs for the (a) Feedforward IDS network (b) cGAN discrimator training (c) VAE and CVAE training.

We used the dataset from [16], previously utilized in [12]. All data were normalized using the min-max scaling approach as described in [12]. After preprocessing, the final cyber-physical dataset comprises 12,514 instances, each with

30 features. This dataset was partitioned into training and testing subsets using an 80-20 split. To establish a baseline IDS, a multi-class feedforward neural network classifier was implemented. The network consists of three hidden layers, each containing 128 neurons. Its convergence behavior over training epochs is illustrated in Figure 3a. The trained IDS achieved a perfect classification accuracy of 100% on the test set.

The subsequent step involved developing a cGAN to generate stealthy adversarial attack with multi-step refinement. The objective of the cGAN is to introduce adversarial perturbations into the features associated with attack samples, such that the IDS misclassifies them as benign. The cGAN architecture comprises a generator and a discriminator, each implemented as a two-hidden-layer network with 256 neurons per layer. Both components were trained using the Adam optimizer with a learning rate of 0.001. The generator optimizes a weighted sum of classification loss, stealth loss, and GAN loss as described before with weights empirically set to $\{\lambda_{\text{stealth}}, \lambda_{\text{GAN}}, \lambda_{\text{cls}}\} = \{10, 0.1, 1\}$. The training performance of the cGAN are shown in Figure 3b.

To detect the stealthy adversarial attack, both VAE and CVAE were trained to extract latent representations from the cyber-physical features. Each model was designed as a three-layer feedforward neural network with 200 latent dimensions. Performance comparison, illustrated in Figure 3c, indicates that the CVAE exhibits lower reconstruction loss than the standard VAE, suggesting improved capability to model the joint distribution of cyber-physical conditioned on class labels. Following the training of these networks, the final phase involves systematically characterizing adversarial attack strategies by varying the number of refinement steps and perturbation magnitudes. This analysis aims to identify regions in the input space where adversarial examples exhibit maximal stealth, thus effectively evading detection by the IDS.

B. Stealthy Attack Characterization

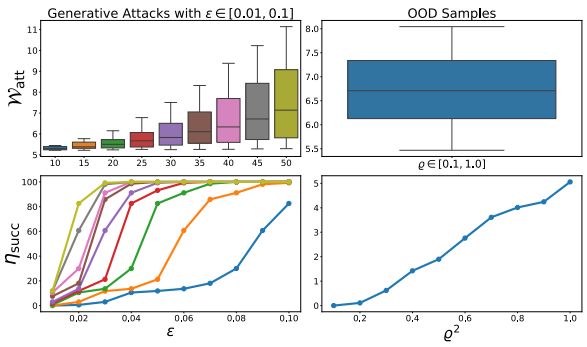

Fig. 4. (a) Distributional distance between X_{adv} and X_{att} with various refinements. (b) Distributional distance between X_{ood} and X_{att} with various refinements. (c) Attack success rate for different refinements (c) Attack success rates for different OOD samples.

We assess adversarial features generated with varying refinement steps by comparing them to OOD features as

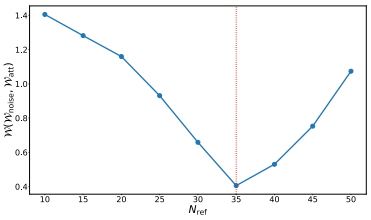

Fig. 5. Distributional Distance between distributional distance between attacked and OOD samples and attacked and adversarial samples for different refinement iterations

per (8). The noise scale ρ is varied from 0.1 to 1, and the maximum perturbation ε for the iterative refinement is from 0.01 to 0.1. With too few refinement steps, features resemble OOD features and fail to consistently mislead the IDS as they can be easily detected by extended Kalman filter. Moreover, as per Fig 4, low steps causes low attack success rate, hence it does not satisfy the stealthy nature of the attack as per Definition 1. Excessive steps for refinement, however, pushes features away from the original feature distribution, while reducing stealthy nature while introducing higher perturbation magnitude. As shown in Figure 4, attack success sharply increases beyond 30 steps, while the distributional distance between adversarial and OOD features remains minimal between 25 and 40 steps. Yet, IDS failure remains low for OOD features, unlike generative attacks, which achieve near-perfect misclassification with fewer steps. We identify 35 steps as optimal, based on minimal distributional distance and high attack success rate as per Fig. 5. If we observe Fig. 4c, for $N_{\text{ref}} = 35$, the attack success rate is close to 80% and 100% with mere maximum perturbation magnitude of 0.03 and 0.04 respectively.

C. Stealthy Attack Detection

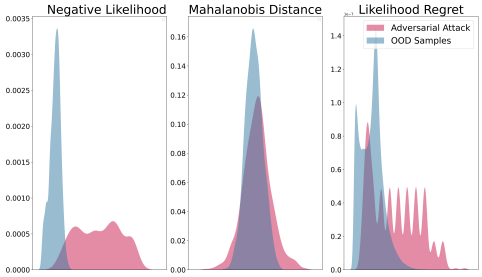

Fig. 6. The metrics to detect adversarial samples from OOD samples (a) NLL (b) Mahalanobis Distance (c) Likelihood regret.

In this subsection, we evaluate the effectiveness of different detection metrics in distinguishing stealthy generative adversarial attack from out-of-distribution (OOD) sample. As illustrated in Figure 6, negative log-likelihood (NLL) offers superior discriminatory power between adversarial and OOD samples compared to alternative metrics such as likelihood regret [11] and Mahalanobis distance [7]. While likelihood regret is capable of differentiating benign inputs from OOD samples to some extent, it fails to account for the varying

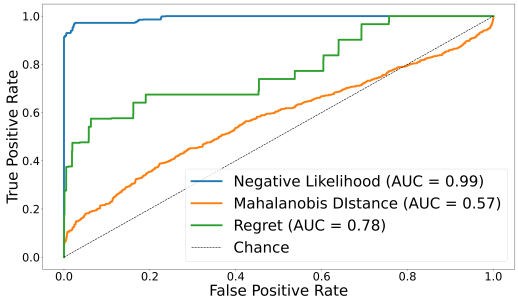

Fig. 7. ROC curve for detecting adversarial samples from OOD samples.

characteristics of distinct OOD inputs. This limitation arises because likelihood regret relies on an relative scoring mechanism, where the reference likelihood can become unstable under various nature of OOD distribution. As a result, its sensitivity to nuanced variations in OOD distributions is reduced. Similarly, the Mahalanobis distance assumes that features within each class are drawn from well-behaved Gaussian distributions. However, this assumption does not hold in the case of CVAE, whose learned latent representations often deviate from strict Gaussianity. Empirically, this observation is supported by the area under the curve (AUC) scores, where NLL achieves a value close to 0.99 as shown in Fig. 7. To validate the significance of detection performance, we computed 95% confidence intervals for the AUC scores using bootstrapping (1,000 resamples). The CVAE-based NLL achieved an AUC of 0.989 ± 0.004, significantly outperforming Mahalanobis ($p < 0.01$) and likelihood regret ($p < 0.001$) via paired Wilcoxon signed-rank tests. This indicates that NLL provides a more robust and reliable signal for distinguishing both OOD and adversarial inputs, outperforming the other two metrics in this setting.

VI. Conclusion

In this study, we proposed a conditional GAN-based adversarial attack framework designed to generate stealthy perturbations capable of evading intrusion detection systems (IDS) within UAV environments. We improve the fidelity and stealth of adversarial examples through an iterative refinement process, strategically minimizes the distributional distance between benign samples and both OOD and adversarial variants. Although the distributional proximity between OOD and adversarial samples relative to benign data appears similar, only the adversarial examples achieve high attack success rates—emphasizing the distinct threat posed by generative attacks. Crucially, we demonstrated that negative log-likelihood (NLL) scores derived from a conditional variational autoencoder (CVAE) offer a reliable means of detecting such stealthy adversarial inputs. In contrast, CVAE-based likelihood regret and Mahalanobis distance metrics proved insufficient for distinguishing adversarial perturbations from OOD features. Our findings highlight the importance of likelihood-based detection methods for uncovering stealthy cyber-attacks that evade conventional UAV intrusion detection systems. Future work will focus on integrating online detection with adaptive navigation to mitigate the real-time impact of adversarial attacks on UAV operations.

References

[1] H. Shakhatreh *et al.*, "Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges," *IEEE Access*, vol. 7, pp. 48 572–48 634, 2019.

[2] D. K. Panda and W. Guo, "Action robust reinforcement learning for air mobility deconfliction against conflict induced spoofing," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 12, pp. 21 343–21 355, 2024.

[3] R. A. AL-Syouf, R. M. Bani-Hani, and O. Y. AL-Jarrah, "Machine learning approaches to intrusion detection in unmanned aerial vehicles (uavs)," *Neural Computing and Applications*, vol. 36, no. 29, pp. 18 009–18 041, 2024.

[4] R. Chauhan and S. S. Heydari, "Polymorphic adversarial ddos attack on ids using gan," in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, IEEE, 2020, pp. 1–6.

[5] Y. Gu and K. Chen, "Gan-based domain inference attack," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 14 214–14 222.

[6] J. Xiao and M. Feroskhan, "Cyber attack detection and isolation for a quadrotor uav with modified sliding innovation sequences," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 7, pp. 7202–7214, 2022.

[7] T. Pang, C. Du, Y. Dong, and J. Zhu, "Towards robust detection of adversarial examples," *Advances in neural information processing systems*, vol. 31, 2018.

[8] L. Smith and Y. Gal, "Understanding measures of uncertainty for adversarial example detection," *arXiv preprint arXiv:1803.08533*, 2018.

[9] A. Takiddin, M. Ismail, U. Zafar, and E. Serpedin, "Variational auto-encoder-based detection of electricity stealth cyber-attacks in ami networks," in *2020 28th European Signal Processing Conference (EUSIPCO)*, IEEE, 2021, pp. 1590–1594.

[10] A. A. Pol, V. Berger, C. Germain, G. Cerminara, and M. Pierini, "Anomaly detection with conditional variational autoencoders," in *2019 18th IEEE international conference on machine learning and applications (ICMLA)*, IEEE, 2019, pp. 1651–1657.

[11] Z. Xiao, Q. Yan, and Y. Amit, "Likelihood regret: An out-of-distribution detection score for variational auto-encoder," *Advances in neural information processing systems*, vol. 33, pp. 20 685–20 696, 2020.

[12] S. C. Hassler, U. A. Mughal, and M. Ismail, "Cyber-physical intrusion detection system for unmanned aerial vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 6, pp. 6106–6117, 2023.

[13] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling tabular data using conditional gan," *Advances in neural information processing systems*, vol. 32, 2019.

[14] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[15] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015.

[16] *UAVs-Dataset-Under-Normal-and-Cyberattacks:* https://github.com/uamughal/UAVs-Dataset-Under-Normal-and-Cyberattacks, [Accessed 23-03-2025].

https://dspace.org/

Generative adversarial evasion and out-of-distribution detection for UAV cyber-attacks

Panda, Deepak Kumar; Guo, Weisi

2025-10-05
Institute of Electrical and Electronics Engineers (IEEE)
https://doi.org/10.1109/smc58881.2025.11343517
Panda DK, Guo W. (2025) Generative adversarial evasion and out-of-distribution detection for UAV cyber-attacks. In: Proceeding of the 2025 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 5-8 Oct 2025, Vienna, Austria, pp. 6295-6300

Downloaded from a DSpace repository. https://dspace.org/