生成对抗逃逸和分布外检测用于无人机网络攻击

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

***Abstract*— 无人机日益融入民用空域，增强了对弹性及智能入侵检测系统（IDS）框架的需求，因为传统的异常检测方法往往难以识别新型威胁。一个常见的策略是将不熟悉的攻击视为离群样本（OOD），因此，不足的缓解响应会使得系统易受攻击，给予对手造成潜在损害的能力。此外，常规的 OOD 探测器通常无法区分隐蔽的对抗性攻击与 OOD 样本。本文提出了一种基于条件生成对抗网络（cGAN）的框架，专门用于设计能够有效规避 IDS 机制的隐蔽对抗性攻击。首先，我们构建了一个稳健的多类分类器作为 IDS，该分类器将良性无人机遥测数据从已知的网络攻击类型中区分出来，包括拒绝服务（DoS）、虚假数据注入（FDI）、中间人（MiTM）和重放攻击。利用这个分类器，我们的提议 cGAN 战略上扰动已知的攻击特征，生成经过精心设计以通过良性误分类来逃避检测的复杂对抗样本。然后，迭代优化生成的隐蔽对抗样本，保持其与离群样本（OOD）统计上的相似性，同时实现较高的攻击成功率。为了有效检测这些隐蔽的对抗性扰动，我们实施了一个条件变分自编码器（CVAE），使用负对数似然作为区分对抗样本和真实 OOD 样本的指标。基于 CVAE 的后悔分析与传统的马氏距离探测器之间的比较表明，CVAE 的负对数似然在检测来自 OOD 样本的隐蔽对抗性攻击方面显著优于传统方法。我们的研究结果强调了高级概率建模技术对于可靠地检测并使现有的 IDS 适应新型、基于生成模型的隐蔽网络威胁的必要性。**

I. 介绍

无人驾驶飞行器（UAVs）越来越多地被用于民用领域，包括空中监视、精准农业和物流等领域 [1]。作为新兴城市空中交通（UAM）系统的关键组成部分，无人机依赖于网络通信和机载自主性，使它们容易受到复杂网络攻击的威胁 [2]。对手可能通过未经授权访问通信渠道截获遥测数据并注入虚假信息，从而妨碍 UAM 运营商验证是否符合计划飞行路径的能力，如图 1 所示。未能检测到敌对攻击作为一般异常可能导致恶意行为者绕过安全措施并对空域操作造成严重危害。这可能会导致无人机进入受限或未经授权的区域，从而危及其他空域用户并增加被劫持或无意中泄露敏感信息的风险。因此，提高无人机系统的安全性至关重要——尤其是在动态、时间关键的操作环境中。传统安全方法如密码学和基于异常检测的 IDS 面临诸如延迟、高误报率以及对新威胁差泛化能力等挑战 [3]。

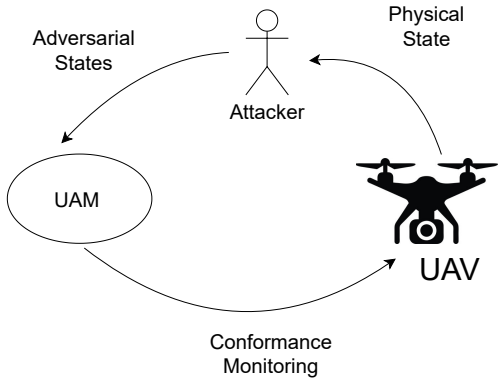

Fig. 1. 攻击者可以操纵无人机发送给无人机交通管理操作员的遥测信息，这会影响对无人机的一致性监控。

生成对抗网络（GANs）已经成为一种强大的工具，用于创建能够逃避检测的多态网络攻击。先前的工作已经展示了它们在生成模仿合法数据分布但损害分类器性能的敌对样本方面的应用 [4, 5]。然而，区分这种隐形敌对样本与自然异常值（OOD）样本文本仍然是一个重大挑战。对手可以利用这一弱点通过生成类似于良性异常的扰动来绕过基于异常检测的 IDS。无法区分分布外（OOD）样本与对抗攻击会削弱传统滤波方法的效果，如扩展卡尔曼滤波器（EKF）[6]，由于这些对抗输入在统计上类似于假设的 OOD 特征，因此可能难以检测到这种隐蔽的对抗输入。

统计方法用于对抗检测 [7, 8] 经常假设固定的数据分布，并且缺乏对高维隐蔽攻击的适应性。此外，它们无法建模潜在空间中的漏洞，限制了其在检测未见过或精心设计的对抗输入方面的有效性。为了克服这些局限性，已提出使用诸如变分自编码器（VAEs）这样的深度生成模型 [9]。VAEs 捕获概率潜在表示以支

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

持更好的异常检测。然而，在存在多种攻击类型的情况下，它们的泛化能力受到限制。条件 VAEs（CVAEs）[10] 将攻击标签纳入学习过程，增强了判别能力和鲁棒性——但其在检测和表征多样化的隐蔽攻击中的应用仍需进一步探索。为了解决这些空白，本文介绍了一个新框架，将生成对抗逃避与使用 CVAE 的概率检测相结合，用于无人机网络物理系统。关键贡献总结如下：

- **隐身对抗攻击生成：**我们设计了一个条件生成对抗网络，扰动来自已知网络攻击（DoS、FDI、MiTM 和重放）的特征，以产生模仿良性行为从而逃避多类入侵检测系统的对抗输入。我们迭代优化这些隐蔽的对抗攻击，确保与 OOD 样本具有统计相似性的同时实现高逃逸成功率。
- **概率检测通过 CVAE：**检测来自 OOD 样本的此类隐蔽攻击，我们实现了一个基于 CVAE 的检测系统。使用负对数似然作为检测分数，我们的方法在区分隐蔽攻击与真实的 OOD 样本方面优于传统的 Mahalanobis 基检测器 [7] 和先进的基于遗憾的 CVAE 基准 [11]。

II. 总体方法论和数据集描述

A. 方法

如图 2 所示，无人机网络物理数据集包含良性特征集和攻击特征集，分别用 X_{ben} 和 X_{att} 表示。一个设计为深度前馈神经网络 $f_{\text{IDS}}(\cdot)$ 的 IDS 被训练和评估为一个多类分类器。随后，设计了一个 cGAN，由一个条件生成器网络 G 和一个判别器 D 组成。生成器 G 学习对输入攻击特征进行对抗性扰动，使得分类器 $f_{\text{IDS}}(\cdot)$ 错误地将其标记为良性样本。同时，判别器 D 作为二元分类器，区分真实的良性样本与对抗生成的样本。这种对抗互动迫使生成器产生越来越隐蔽且难以检测的扰动。扰动在 N_{ref} 步骤中迭代优化，以生成统计上与 OOD 特征相似但具有高攻击成功率的对抗样本 X_{adv}。CVAE 在良性数据和带有攻击标签的数据上预训练后，用于计算对抗 X_{adv} 和 OOD 样本 X_{ood} 的负对数似然（NLL）。根据负对数似然的分布将隐秘的对抗样本从 OOD 特征中分类出来。

B. 网络攻击数据集

本研究使用的无人机网络攻击数据集基于 Hassler 等人开发的入侵检测框架 [12]，其中采用了全面的实验设置。该数据集由多种类型网络攻击后的无人机遥测数据组成，简要描述如下：

- **去认证：**欺骗性断开连接导致安全模式触发。
- **重放：**记录的命令信号被重放以引起无人机行为异常。
- **恶双胞胎：**一个流氓接入点使中间人攻击成为可能。
- **虚假数据注入：**传感器状态操纵误导无人机控制器。

这些攻击在网络和物理领域都表现出明显的特征。关于网络特征，只保留了所有攻击类型共有的属性以确保一致性。这些包括：时间戳、帧号、帧长度、通过无线介质的传输时长、帧序列号、控制类型及子类型以及 WLAN 片段编号。相比之下，物理特征则涵盖了各种无人机遥测数据，如高度；沿 x,y,z 轴（俯仰、滚转、偏航）的速度；飞行距离和时间；压力；电池状态；以及相对于 Tello Pad 沿 x,y,z 轴的空间距离。实验的细节见 [12]。

鉴于此标记数据集的可用性，设计一个能够区分良性行为和恶意行为以及不同攻击类型的多类分类模型变得可行。然而，如果对手获得了标记的攻击数据，他们可以利用这些信息构建基于条件生成对抗网络（cGAN）的模型来生成隐蔽的对抗样本。这些从已知攻击特征合成的例子是精心制作的以规避现有的入侵检测系统（IDS）。

III. 使用 GAN 的隐秘对抗攻击

本节概述了一个条件生成对抗网络（cGAN）框架的设计，该框架旨在针对基于神经网络的入侵检测系统（IDS）$f_{\text{IDS}}(\cdot)$ 生成对抗攻击。cGAN 架构包括两个神经网络：一个生成器 $G(\mathfrak{z})$，它从注入的噪声中合成对抗扰动；以及一个判别器 $D(x)$，它学习区分良性输入和对抗性扰动输入。这两个网络在一个对抗（极小极大）优化框架下进行训练，其中生成器试图欺骗判别器，而判别器同时提高其区分真实样本和对抗样本的能力。目标函数由以下给出，

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

这里，x 代表良性数据样本，$\mathfrak{z}$ 代表随机噪声向量，$G(\mathfrak{z})$ 生成对抗扰动，$D(x)$ 输出 x 是良性的概率。为了求解在 (1) 中描述的目标，条件生成器和判别器将在下一小节中进行描述。

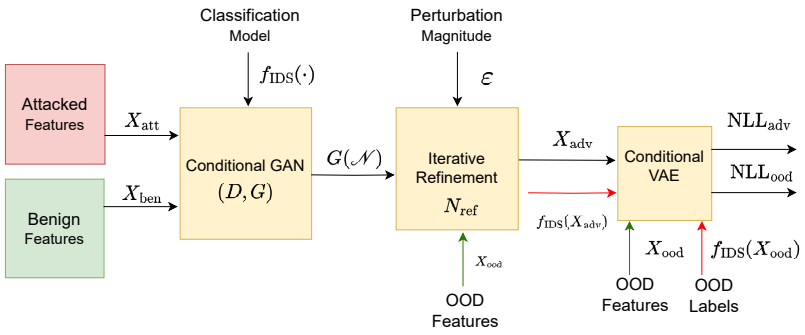

Fig. 2. 获取负对数似然（NLL）以检测来自 OOD 样本的隐秘对抗攻击的示意图。

A. 条件生成器

生成器被建模为一个前馈神经网络，旨在将噪声向量转换成对抗特征表示。具体来说，它将输入的噪声映射到能够被入侵检测系统（IDS）误分类为良性的对抗特征上。形式上，生成器定义为函数 $G:\mathbb{R}^{d_{\text{noise}}}\to\mathbb{R}^{d_{\text{input}}}$，其中 d_{noise} 表示输入噪声向量 $\mathfrak{z}$ 的维度，d_{input} 对应于无人机系统的网络物理特征空间的维度。生成器有两个主要目标：(1) 引起目标 IDS 的误分类，从而最大化其预测误差；(2) 确保生成的对抗特征在特征空间中与合法良性样本相似度高。令 $\delta=G(\mathfrak{z})$ 表示从噪声向量 $\mathfrak{z}$ 生成的对抗扰动。对抗输入 X_{adv} 是通过用 2-范数扰动原始（被攻击）特征向量构建的。令 $f_{\text{IDS}}(X_{\text{adv}})$ 表示当 IDS 接收到对抗输入时产生的分类概率，令 y_{ben} 表示良性样本的真实标签分布。因此，cGAN 生成器的总损失由下式给出，

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

其中 λ_{cls},λ_{st} 和 λ_G 分别表示控制误分类损失、隐蔽损失和生成器损失的超参数。(2) 中的第一项指的是 Kullback-Leibler(KL) 散度损失，它引导生成器朝向产生类似良性的预测。第二项引入了特征相似性损失，鼓励对抗特征在欧氏空间中保持与良性特征的相似性。第三项是标准 cGAN 公式的一部分，有助于生成更多类似的良性样本。

B. 判别器

判别器 D 被训练以区分良性样本和对抗样本，因此使用二元交叉熵损失来区分它，如下所示

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

在 cGAN 的训练过程中，我们假设攻击者可以通过外部 API 访问 IDS。cGAN 可以按照标准方式进行训练，如 [13] 所示。

C. 迭代细化的隐秘攻击

迭代细化策略类似于投影梯度下降 (PGD) 攻击 [14] 中的策略，其中攻击以多步注入的方式进行，使得整体扰动保持在一定范围内。目标是使对抗攻击 X_{adv} 接近由正态分布噪声生成的 OOD 样本 X_{ood}，噪声尺度为 ρ，定义为 $X_{\text{ood}}=X_{\text{att}}+\mathcal{N}\left(0,\rho^2 I\right)$。因此，为了确保对抗攻击的秘密性，$X_{\text{adv}}$ 需要在分布上与 X_{ood} 类似，并具有较高的攻击成功率。因此，我们将隐秘的对抗攻击定义如下：

定义 *1:* 令 $X\subset\mathbb{R}^d$ 表示用于 IDS 的无人机遥测数据输入特征。我们定义 P_{ben}、P_{ood}、P_{adv} 分别为良性无人机、OOD 和对抗攻击特征的分布。我们定义 $\mathscr{W}(\cdot,\cdot)$ 为衡量分布距离的 Wasserstein 距离，因此我们定义隐蔽的对抗攻击如果满足以下条件：

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

如果攻击者想要诱导最大幅度为 ε 的扰动，则迭代细化下的最终对抗攻击特征 X_{adv} 的定义如下所示：

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

如果 N_{att} 和 $N_{\text{adv}}^{\text{ben}}$ 分别表示被攻击的数量和被视为良性的对抗特征数量，那么攻击成功率由 $\eta_{\text{succ}} = N_{\text{adv}}^{\text{ben}}/N_{\text{att}}$ 给出。因此，为了计算根据定义 1 的最优隐蔽攻击的细化 N_{ref}，目标函数可以构架为，

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

使用迭代细化生成攻击的整个算法如下：

Algorithm 1 迭代改进隐蔽的对抗样本

Require: 对抗输入 X_{att}，生成器 $G(z,c)$，边界 ε，细化步骤 N_{ref}

初始化 $X_{\text{adv}}^{(0)} \leftarrow X_{\text{att}}$

for $i = 1$ to N_{ref} **do**

采样噪声 $z \sim \mathscr{N}(0, I)$，标签 c

生成扰动：$\delta = G(z,c)$

应用有界更新：

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

end for

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

IV. 用于隐蔽攻击检测的条件变分自编码器

CVAEs[10] 表示一类广泛应用于各种实际领域的强大的深度概率生成模型。这些模型包含一个潜在变量 z，从先验分布 $p(z)$ 中采样，并以条件标签 c 为依据的条件分布 $p_\theta(x|z,c)$ 来生成观测变量 x。然而，在高维设置中直接获得边际似然 $p_\theta(x)$ 在计算上是不可行的，因为需要对潜在空间进行积分。为了解决这一挑战，采用了变分推断来近似后验分布，并导出观测数据对数似然的一个可处理下界。这导致了众所周知的证据下界（ELBO），它作为训练模型的目标函数：

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

这里，$q_\phi(z|x,c)$ 表示对真实后验分布 $p_\theta(z\,|\,x)$ 的变分近似，其中 ϕ 和 θ 分别代表编码器和解码器网络的参数。这些组件通常使用神经网络实现。CVAE 通过最小化训练数据上的变分目标 $\mathscr{L}(x;\theta,\phi)$ 进行训练。然而，CVAE 并不直接最大化真实似然；相反，它优化定义在（7）中的 ELBO。这种近似可能导致学习到的表示过于受限或表达能力有限，因为只有少量样本可能落在近似后验概率高的区域。为了解决这一限制，采用了重要性加权自编码器（IWAE）[15]。IWAE 通过使用一个 $k-$ 样本的重要性加权估计引入了边缘对数似然 $\log p(x)$ 的一个更紧密的下界，从而使得生成模型更具表达力。

$$\mathscr{L}_k(x) = \mathbb{E}_{z_1,\cdots,z_k \sim q(z|x,c)}\left[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p(x,z_i)}{q(z_i|x,c)}\right]. \tag{8}$$

这里 $z_1,\cdots,z_k$ 是从编码器网络中独立采样的。求和内部的项对应于联合分布未归一化的重要性权重，表示为 $w_i = p(x,z_i)/q(z_i\,|\,x)$。由 Jensen 不等式可以证明，平均重要性权重是给定的 $p(x)$ 的无偏估计量，

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

使用重要性采样计算特定数据点的权重公式为，

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

第一项表示来自潜特征 z 和条件标签 c 的可能性。第二项表示潜变量 z 的标准正态分布先验。第三个项表示给定数据 x 和条件标签 c 的潜变量 z 的变分后验分布。给定采样的 x 的负对数似然（NLL）由以下公式给出：

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

如 (11) 所示，NLL 用于计算对抗特征 X_{adv} 和 OOD 特征 X_{ood}。

V. 结果与讨论

A. 数值实现与训练结果

本研究中使用的数据集来源于 [16]，如在 [12] 中所用。所有数据均采用 min-max 归一化方法进行处理，

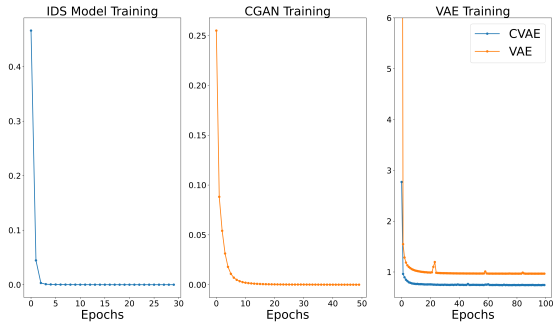

Fig. 3. 损失函数与训练周期对于（a）前馈 IDS 网络（b）cGAN 判别器训练（c）VAE 和 CVAE 训练。

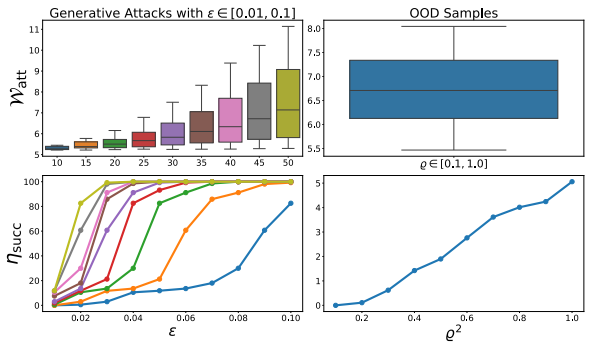

Fig. 4. (a)X_{adv} 和 X_{att} 在各种细化下的分布距离。(b)X_{ood} 和 X_{att} 在各种细化下的分布距离。(c) 不同细化的攻击成功率 (c) 对于不同 OOD 样本的攻击成功率。

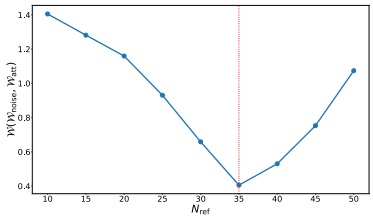

Fig. 5. 攻击样本与 OOD 样本以及攻击样本与对抗样本之间的分布距离的不同细化迭代结果

具体描述见 [12]。预处理后，最终的网络物理数据集包含 12,514 个实例，每个实例有 30 个特征。该数据集使用 80-20 分割法划分为训练子集和测试子集。为了建立基线入侵检测系统（IDS），实现了一个多类前馈神经网络分类器。该网络由三个隐藏层组成，每层包含 128 个神经元。其在训练周期中的收敛行为如图 3a 所示。经过训练的 IDS 在测试集上达到了完美的分类准确率，即 100%。

后续步骤涉及开发一个 cGAN 来生成经过多步优化的隐形对抗攻击。cGAN 的目标是在与攻击样本相关的特征中引入对抗性扰动，从而使 IDS 将其误分类为良性。cGAN 架构包括一个生成器和一个判别器，每个组件都实现为具有每层 256 个神经元的两隐藏层网络。两个组成部分均使用 Adam 优化器进行训练，学习率为 0.001。生成器损失函数包含三个目标的加权组合：分类损失、隐形损失和 GAN 损失。相应的权重被经验性地设置为 $\{\lambda_{\text{stealth}}, \lambda_{\text{GAN}}, \lambda_{\text{cls}}\} = \{10, 0.1, 1\}$。cGAN 的训练性能如图 3b 所示。

为了检测隐身的对抗攻击，VAE 和 CVAE 都被训练用于从网络物理特征中提取潜在表示。每个模型都被设计为一个三层前馈神经网络，并具有 200 个潜在维度。性能比较如图 3c 所示，表明 CVAE 的重建损失低于标准 VAE，这表明它在建模基于类别标签的网络物理联合分布方面具有改进的能力。这些网络训练完成后，最后阶段涉及通过改变细化步骤数量和扰动幅度来系统地描述对抗攻击策略。此分析旨在识别输入空间中对抗示例表现出最大隐身性的区域，从而有效逃避入侵检测系统的检测。

B. 隐蔽攻击特征分析

我们通过将使用不同细化步骤生成的对抗特征与 OOD 特征进行比较来评估它们（如 (8) 所述）。噪声尺度 ρ 从 0.1 变化到 1，迭代细化的最大扰动 ε 则从 0.01 变化到 0.1。如果细化步骤太少，则特征类似于 OOD 特征，并且无法始终误导入侵检测系统，因为它们可以被扩展卡尔曼滤波器轻易检测到。此外，如图 4 所示，步骤过低会导致攻击成功率降低，因此不符合定义 1 中的隐蔽性要求。然而，过多的细化步骤会使特征远离原始特征分布，同时减少隐蔽性并引入更高的扰动幅度。如图 4 所示，超过 30 步时攻击成功率急剧增加，而在 25 到 40 步之间对抗样本和 OOD 特征之间的分布距离保持最小。然而，对于 OOD 特征而言 IDS 失败率仍然很低，与生成式攻击不同的是，后者通过较少的步骤就能达到近乎完美的误分类效果。为了找到最佳隐蔽性细化程度，我们最小化良性特征与 OOD 及对抗样本特征间的分布距离，如图 5 所示，确定 35 步为理想值。如果我们观察图 4c，对于 $N_{\text{ref}} = 35$ 而言，在扰动幅度最大仅为 0.03 和 0.04 时攻击成功率分别接近 80%和 100%。

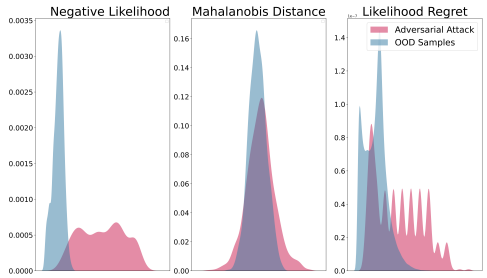

Fig. 6. 用于从 OOD 样本中检测对抗样本的度量标准（a）对数似然比（b）马氏距离（c）似然遗憾。

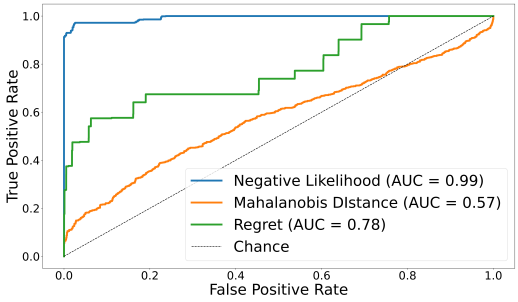

Fig. 7. 检测来自域外样本的对抗样本的 ROC 曲线。

C. 隐蔽攻击检测

在本小节中，我们评估了不同检测指标在区分隐秘生成对抗攻击与分布外 (OOD) 样本方面的有效性。如图 6 所示，负对数似然 (NLL) 相较于其他指标（例如似然遗憾 [11] 和马氏距离 [7]）在区分对抗样本与 OOD 样本方面具有更优的鉴别能力。虽然似然遗憾能够在一定程度上区分良性输入与 OOD 样本，但它无法考虑不同 OOD 输入的不同特征。这一局限性源于似然遗憾依赖于一个相对评分机制，在不同的 OOD 分布性质下，参考似然可能会变得不稳定。因此，它对 OOD 分布微妙变化的敏感度降低。类似地，马氏距离假设每个类内的特征来自表现良好的高斯分布。然而，对于 CVAE 而言，这一假设并不成立，其学习到的潜在表示通常偏离严格的高斯性。经验上，该观察结果得到了曲线下面积 (AUC) 得分的支持，在图 7 中 NLL 达到了接近 0.99 的值。这表明 NLL 提供了区分 OOD 和对抗输入更强大且可靠的信号，在这种设置中优于另外两个指标。

VI. 结论

在本研究中，我们提出了一种基于条件生成对抗网络的对抗攻击框架，旨在产生能够逃避无人机环境中入侵检测系统的隐蔽扰动。为了提高所生成对抗样本的真实性和隐匿性，我们引入了一种迭代优化机制，该机制战略性地最小化良性样本与 OOD 和对抗变体之间的分布距离。尽管 OOD 样本与对抗样本相对于良性数据的分布相近，但只有对抗样本能够实现高攻击成功率——这强调了生成式攻击所带来的独特威胁。至关重要的是，我们证明了从条件变分自编码器（CVAE）衍生出的负对数似然（NLL）分数提供了一种可靠的方法来检测此类隐蔽的对抗输入。相比之下，基于 CVAE 的可能性遗憾和马氏距离指标对于区分对抗扰动与 OOD 特征是不够的。这些结果突显了传统评分方法的局限性，并强调了基于可能性的检测方法在确保安全无人机系统中强大的 OOD 检测的重要性。

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].