计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200189

基于 GAN 的对抗网络流量生成研究

杨　琳[1]　林宏刚[1,2]
1 成都信息工程大学网络空间安全学院(芯谷产业学院)　成都 610225
2 先进密码技术与系统安全四川省重点实验室　成都 610225
(linyang_24330@163.com)

摘　要　对抗网络流量在设备隐私保护和网络安全等领域扮演着重要角色,然而目前对抗网络流量生成方法缺乏对质量的约束,导致生成的流量偏离原始流量特性,在实际应用中丧失其对抗能力。因此,提出一种基于 GAN 的对抗网络流量生成方法,改进生成器设计,以卷积神经网络提取原始流量特征的抽象表示,经基础迭代算法生成扰动,确保扰动保持原始流量的特性;优化生成器损失函数,实现生成流量与原始流量之间的最小差异;引入干扰器模块,利用网格搜索算法为扰动分配权重并优选参数组合,保证生成流量的多样性。为了综合考虑特征空间距离差异与相对变化速率对生成质量的影响,提出相对差异扰动量指标,能更准确地评估对抗网络流量与原始流量之间的差异。实验结果表明,在有效扰动范围内,相较于其他方法,该方法生成的对抗网络流量对目标分类模型保持高欺骗率的同时,产生的 L_∞ 扰动量与相对差异扰动量均更小,与原始流量的相似性更高,有效提高了对抗网络流量的生成质量。

关键词:生成对抗网络;对抗网络流量;相对差异扰动;相似性保持目标函数;生成质量控制
中图分类号　TP393

Research on Generating Adversarial Network Traffic Based on Generative Adversarial Network

YANG Lin[1] and LIN Honggang[1,2]
1 College of Cyberspace Security, Chengdu University of Information Technology, Chengdu 610225, China
2 Sichuan Key Laboratory of Advanced Cryptography and System Security, Chengdu 610225, China

Abstract　Adversarial network traffic plays a crucial role in fields such as device privacy protection and network security. However, current adversarial network traffic generation methods lack constraints on quality, resulting in generated traffic that deviates from the original traffic characteristics, thereby losing its adversarial capability in practical applications. Therefore, this paper proposes a GAN-based adversarial network traffic generation method, which improves the generator design. The convolutional neural network is employed to extract abstract representations of original traffic features, and perturbations are generated through basic iterative algorithms to ensure that the perturbations maintain the characteristics of the original traffic. The generator loss function is optimized to achieve minimal differences between the generated traffic and the original traffic. Additionally, a perturber module is introduced, utilizing a grid search algorithm to assign weights to perturbations and optimize parameter combinations, ensuring the diversity of the generated traffic. To comprehensively consider the impact of feature space distance differences and relative change rates on generation quality, a relative difference disturbance metric is proposed to more accurately evaluate the differences between adversarial network traffic and the original traffic. Experimental results show that, within an effective perturbation range, compared to other methods, the adversarial network traffic generated by this method maintains a high deception rate for target classification models while producing smaller L_∞ disturbance and relative difference disturbance values, and exhibiting higher similarity to the original traffic, effectively improving the generation quality of adversarial network traffic.

Keywords　Generative adversarial network, Adversarial network traffic, Relative differential disturbance, Similarity preservation objective function, Generative quality control

1　引言

随着机器学习技术的发展,攻击者仅通过分析网络通信的流量,即可获取网络设备各种有价值的信息[1],进而对其进行恶意活动。为了有效应对这一威胁,研究者提出了对抗网络流量生成技术[2],即通过混淆网络流量的真实特征,从而保护设备身份信息,规避公开漏洞攻击[3]。因此高质量的生成对抗网络流量对于增强网络安全,保护设备隐私至关重要。

生成对抗网络(Generative Adversarial Network, GAN)[4]因其强大的生成能力,近年来在对抗网络流量生成

基金项目:国家 242 信息安全计划项目(2021-037);四川省自然科学基金项目(2024NSFSC0515)
This work was supported by the National 242 Information Security Plan Project(2021-037) and Sichuan Provincial Natural Science Foundation Project(2024NSFSC0515).
通信作者:林宏刚(linhg@cuit.edu.cn)

领域备受关注。在传统基于 GAN 的对抗网络流量生成技术中，预先训练好的分类模型使判别器在训练初期便展现出了强大的识别能力，而生成器模拟真实数据能力较弱，导致梯度更新有限，训练困难，可能引发模式崩溃。为缓解此现象，研究者通过增大输入数据的噪声添加幅度以增加样本的多样性。与此同时，也出现新的问题：模型仅关注提升对抗网络流量的多样性和丰富程度，忽略了对对抗网络流量生成过程的可控性，导致引入一些与原始流量不符的特征或模式，使得对抗网络流量在实际应用中很容易被检测或识别出来，从而丧失其对抗能力。

针对上述问题，本文提出了一种新的对抗网络流量生成方法：可控生成对抗网络(Controllable Generative Adversarial Network，Control-GAN)。该方法首先利用卷积神经网络从原始流量中提取特征表示，随后通过基础迭代算法(Basic Iterative Method，BIM)对其进行梯度优化，生成基础扰动；接着，采用干扰器为其分配多组权重，与原始流量融合，通过最小化均方误差，选择与原始流量差异最小的对抗网络流量；最终，将对抗网络流量输入目标分类模型，利用相似性保持目标函数，确保对抗网络流量与原始流量中心特征的距离最小化。

综上所述，本文的贡献主要有：

1)改进生成器设计：结合卷积神经网络和 BIM 算法，通过学习原始流量的特征表示并应用梯度优化来生成基础扰动；改进生成器的损失函数，采用生成目标函数与相似性保持目标函数双重控制，使对抗网络流量既能有效干扰判别器，又接近原始流量中心特征，提高其生成质量。

2)改进生成对抗网络的结构，增加干扰器模块：干扰器在训练阶段负责为基础扰动分配权重并优选参数组合，确保对抗网络流量在保持原始流量基础特性的同时，引入足够的差异性和复杂性，从而保证对抗网络流量的多样性。

3)提出了一种新的扰动量计算指标——相对差异扰动量(Relative Differential Disturbance，RDD)：改进传统扰动量计算方式，通过计算相对差异变化速率替代各特征平方和的平方根，降低了大值的敏感性，更精确地评估对抗网络流量与原始流量之间的差异。

本文第 2 章概述了对抗网络流量生成的研究现状；第 3 章详细介绍了基于 Control-GAN 的对抗网络流量生成方法；第 4 章通过对比实验验证了本文方法的有效性；最后总结全文。

2 相关工作

对抗网络流量通过干扰算法向原始流量中添加扰动，改变流量的特征，使目标模型难以准确识别或分类流量，从而保护设备安全。现有的主流对抗网络流量生成方法主要分为两种：基于梯度的对抗网络流量生成和基于 GAN 的对抗网络流量生成[5]。

2.1 基于梯度的对抗网络流量生成

基于梯度的对抗网络流量生成方法通过增大梯度反方向上的扰动来生成对抗网络流量。Zhang 等[6]首次探讨了将对抗样本用于规避流量分析攻击的方法，通过快速梯度符号方法(Fast Gradient Sign Method，FGSM)计算梯度生成扰动信息，并利用卷积神经网络(Convolutional Neural Network，CNN)分类模型对对抗网络流量进行识别，使用梯度上升的方式快速地生成扰动。然而，该方法只能进行单次梯度计算，并且无法充分捕捉模型的非线性特性。Hu 等[7]在 Moore 数据集[8]上验证了 FGSM，Deepfool 和 C&W 3 种不同的扰动算法生成的对抗网络流量对分类模型的总体欺骗效果和单类欺骗效果。实验结果表明：以 LeNet-5 深度卷积神经网络作为分类模型，FGSM 方法可以达到 99% 的总体欺骗率；在 email 类型中，FGSM 可以实现 96% 的单类欺骗率。总体而言，3 种扰动方法中，FGSM 对分类模型的欺骗效果最佳，但是该方法产生的扰动量也更大。Rigaki 等[9]评估了 FGSM 和基于雅可比的显著性图攻击(Jacobian-based Saliency Map Attack，JSMA)两种方法对目标模型的干扰结果，目标模型包括决策树、随机森林、线性支持向量机等。实验结果表明：在 NSL-KDD 数据集上，所有分类器都受到了影响，其中影响最严重的是线性支持向量机模型。Huang 等[10]同样使用 FGSM 和 JSMA 干扰算法，不同的是，其目标模型更换为多层感知机(Multilayer Perceptron，MLP)、CNN 和长短期记忆网络(Long Short-Term Memory，LSTM) 3 种深度学习模型。两者的实验结果都表明，与 FGSM 相比，JSMA 需要更多的时间来生成对抗网络流量，但生成的对抗网络流量对目标模型的干扰效果更好。Ibitoye 等[11]使用了 FGSM、BIM、投影梯度下降(Projected Gradient Descent，PGD) 3 种扰动方法。对于 FGSM 对抗样本，模型的预测精度从 95.1% 降低到 24%，用 BIM 和 PGD 对抗样本重复实验，精度分别降低了 18% 和 31%。由于 BIM 算法具有迭代优化的特点，其对目标模型的欺骗效果表现更好。

2.2 基于 GAN 的对抗网络流量生成

基于 GAN 的对抗网络流量生成方法通过生成器生成对抗网络流量，并借助判别器对其进行区分。生成器和判别器通过对抗训练相互学习，持续提升对抗网络流量的逼真度。

Li 等[12]第一次提出基于 GAN 的对抗网络流量生成方法，即通过提取目标流量集的特征生成变形流量集，并采用随机掷硬币的策略发送流量。该方法对攻击者具有一定的迷惑效果，但其结果具有随机性，无法实现实时动态对抗网络流量生成。CHOUGULE 等[13]提出了 SCAN-GAN，用于生成合成的 CAN 数据集，解决了现有 CAN 数据集稀缺的问题。通过与原始数据集的对比，发现生成数据集在多个参数和分类算法下表现更佳，证明了 GAN 在数据生成中的有效性和适应性。Anande 等[14]实现了两种生成对抗网络模型，通过数据转换，生成的数据在均值和标准差的对数值、主成分等方面与真实数据高度相似，且约 85% 的生成流量特征可无差别地替代真实数据。为了提高对抗网络流量的生成质量，Hui 等[15]通过知识图谱引入了各种物联网设备的语义知识和网络结构知识，再采用 LSTM 和注意力机制来捕获流量序列的时间相关性，最后通过 GAN 生成大规模的物联网流量。该模型能够直接生成具有综合掩蔽能力的对抗网络流量，但需要进行复杂的先验知识的学习，并且缺乏动态伪装能力。Huang 等[16]通过集成门控循环单元(Gated Recurrent Unit，GRU)模型作为编码器和解码器，对嵌入层转换后的特征嵌入向量序列进行建模，学习特征序列的远距离依赖关系，生成扰动。这种方法丰富了原始流量中的特征表示。Li 等[17]通过输入原始流量与随机噪声生成对抗网络流量，并在训练过程中采用梯度反向传播进行优化。该方法对分类模型具有较强的欺骗能力，但无法控制扰动的添加过程。

综上所述，基于梯度的对抗网络流量生成算法实现简单，但依赖于单一的梯度更新方式，降低了对抗网络流量的多样性，基于 GAN 的对抗网络流量生成方法训练过程更具探索空间。然而，现有的基于 GAN 的对抗网络流量生成方法主要依赖随机噪声，缺乏对对抗网络流量生成质量以及模型训练过程的控制。

3　基于 Control-GAN 的对抗网络流量生成方法

针对现有对抗网络流量生成方法的不足，本文设计了一种新的对抗网络流量生成模型 Control-GAN，其由生成器、干扰器和判别器三部分组成，如图 1 所示。与现有 GAN 模型相比，Control-GAN 通过引入干扰器模块和相似性保持目标函数，优化了 GAN 的网络结构和生成器的损失函数，增强了模型训练稳定性，减小了对抗网络流量与原始流量在特征空间中的分布差异。

生成器结合卷积神经网络和 BIM 算法，通过学习原始流量的特征表示并应用梯度优化来生成基础扰动；干扰器则通过网格搜索为其分配权重，与原始流量融合后选择最小化均方误差的参数组合，从而输出对抗网络流量；预先训练的判别器分类模型用于识别对抗网络流量。在当前的 Control-GAN 架构中，训练模式调整为固定判别器仅优化生成器。生成器的损失函数包含生成目标函数和相似性保持目标函数，前者旨在增强判别器对对抗网络流量的不可区分性，后者则确保两者在中心特征空间中的类内相似性。整个损失函数的设计目标在于，对抗网络流量能够干扰判别器的分类结果，同时保持与原始流量之间的最小差异。

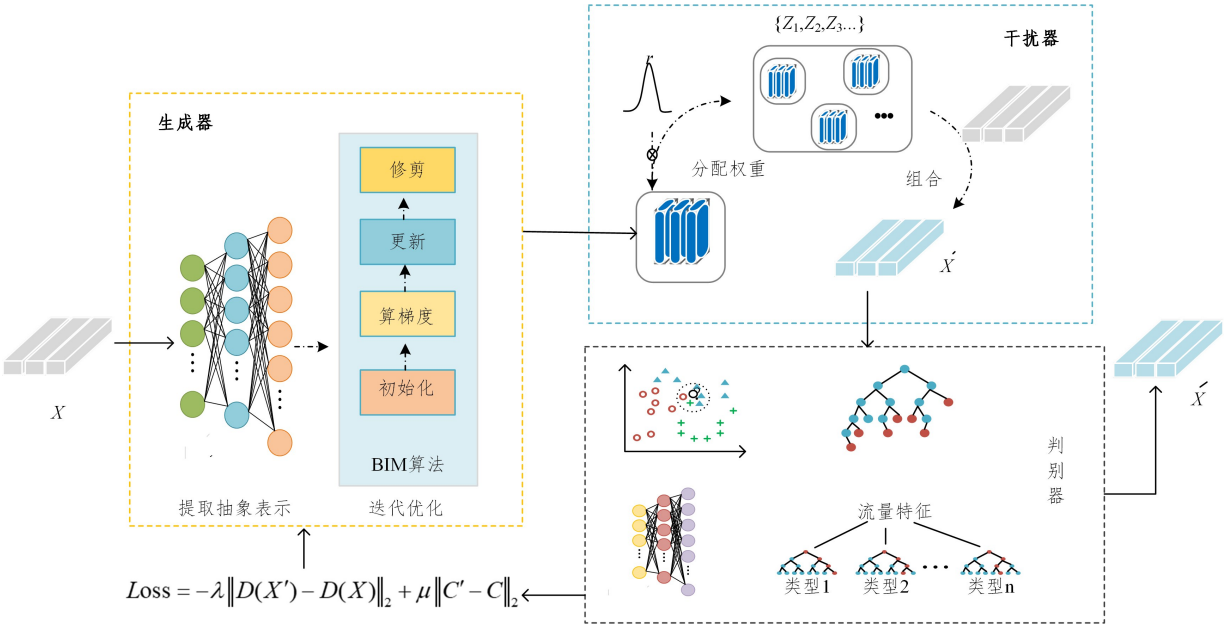


图 1　Control-GAN 对抗网络流量生成模型

Fig. 1　Control-GAN adversarial network traffic generation model

3.1　基于 BIM 算法的生成器设计

传统生成器以随机噪声生成对抗网络流量，侧重欺骗目标分类模型而增大与原始流量的差异。为此，本文生成器的主要任务是生成基础扰动，该基础扰动为原始流量特征的抽象表示，保留了原始流量的基础特性，提高了对抗网络流量的生成质量。

基于 BIM 算法的生成器迭代训练包括两步：第一步，通过卷积神经网络学习并提取原始流量特征的抽象表示；第二步，将该抽象表示传入 BIM 算法中进行梯度优化，生成基础扰动。

3.1.1　抽象表示提取

如图 2 所示，为避免简单地添加随机噪声可能破坏数据内在结构的问题，利用抽象表示取代随机噪声保持原始流量的基本结构，这使得数据在经过处理后仍然具有一定的相似性和连续性。

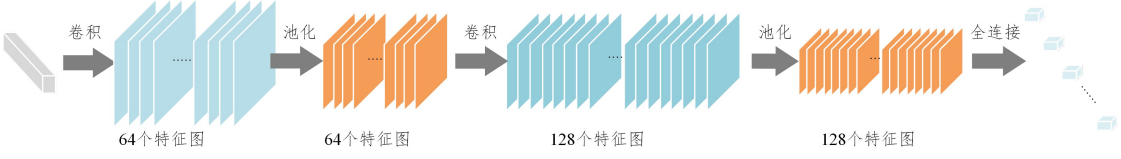


图 2　特征抽象表示提取过程

Fig. 2　Feature abstraction representation extraction process

输入原始流量 x，一维卷积层使用一组卷积核对 x 进行卷积操作，生成特征图：

$$y[i]=\sum_{j=0}^{M-1}x[i+j]\cdot w[j]+b \tag{1}$$

其中，$y[i]$是特征图的第 i 个元素，$x[i]$是原始流量的第 i 个特征值，$w[j]$是卷积核的第 j 个权重，b 是偏置项，M 是卷积核的长度。

通过池化层进行最大池化操作，减少特征图的维数：

$$y[i]=\max(x[i\cdot s:(i+1)\cdot s]) \tag{2}$$

其中，$y[i]$是池化后的特征图的第 i 个元素，$[i\cdot s:(i+1)\cdot s]$是流量的第 i 个池化窗口。最后，使用全连接输出层，通过

tanh 函数将输出控制在[−1,1]。

3.1.2 基础扰动生成

生成基础扰动的过程主要分为 4 步。

1)初始化参数

初始化设置迭代次数 T、迭代步长 α、剪枝参数 ε,原始基础扰动 z,初始化噪声 $\eta=z$。

2)计算梯度

将噪声 η 添加到原始基础扰动 z 上,计算 η 引起的梯度变化:

$$\nabla J(\theta, z+\eta, z) \tag{3}$$

其中,J 为计算梯度的函数,θ 为参数设置。梯度变化的方向和大小提供了关于基础扰动 z 如何变化的信息。

3)梯度更新

对梯度进行更新:

$$\alpha \cdot \text{sign}(\nabla z' J(\theta, z', z)) + z \tag{4}$$

其中,符号函数 sign()的表达式如下:

$$\text{sign}(gradient)=\begin{cases}-1, & gradient<0\\ 0, & gradient=0\\ 1, & gradient>0\end{cases} \tag{5}$$

通过符号函数转化与迭代步长 α 相乘,得到更新后的基础扰动。

4)基础扰动修剪

为了避免算法产生的基础扰动过大,采用修剪函数对其进行缩放和修剪:

$$z_{ij}'=\begin{cases}clip\_min, & z_{ij}'<clip\_min\\ z_{ij}', & clip\_min z_{ij}'\leqslant clip\_max\\ clip\_max, & z_{ij}'>clip\_max\end{cases} \tag{6}$$

其中,z_{ij}' 为扰动量中第 i 行第 j 列的元素,$clip\_min$ 为最小剪切值,$clip\_max$ 为最大剪切值。

基础扰动修剪函数将所有元素限制在指定范围内,小于最小剪切值的元素被替换为最小剪切值,大于最大剪切值的元素被替换为最大剪切值。通过修剪,可以保证基础扰动中的每一个元素 z_{ij}' 都限制在闭区间[$clip\_min$,$clip\_max$]内。

3.1.3 目标函数

为了使对抗网络流量能够干扰判别器的分类结果,同时保持与原始流量之间的最小差异,本文对传统生成器的损失函数进行了改进。具体地,生成器的损失函数包括生成目标函数 L_G 与相似性保持目标函数 L_S,其中,生成器目标函数用于增大对抗网络流量与原始流量之间的差异,提高对判别器的干扰能力,具体表达式如下:

$$L_G=-\|D(X')-D(X)\|_2 \tag{7}$$

其中,$D(X')$ 与 $D(X)$ 分别为对抗网络流量与原始流量在隐空间的特征表示;L_G 越小表明对抗网络流量与原始流量在特征空间中的距离越大,生成的对抗网络流量对判别器的干扰效果越强。因此,本文通过最小化 L_G 来提高生成器的生成能力。

在使判别器实现误分类的同时,为了控制对抗网络流量的生成质量,本文还引入了相似性保持目标函数 L_S,具体表达式如下:

$$L_S=\|C'-C\|_2 \tag{8}$$

其中,C' 与 C 分别为对抗网络流量与原始流量的中心特征,约束 L_S 可以缩小对抗网络流量与原始流量中心特征的分布距离。其中,中心特征 C 的提取过程如算法 1 所示。

算法 1 中心特征提取算法

输入:原始流量特征 X,真实标签 y,特征数 n

输出:中心特征 C

```
1. 开始
2. FOR i IN range(n) DO
       P(X_i,y) //计算联合概率分布
       P(X_i) //计算特征边缘概率分布
       P(y) //计算 y 边缘概率分布
       I(X_i;y)=P(X_i,y) * log2(P(X_i,y) /(P(X_i) * P(y)))
       //计算互信息
3. END FOR
4. QuickSort(I(X_i;y))//进行排序
       pivot=I(X_i;y)[len(I(X_i;y)) // 2]
       left=[x for x in I(X_i;y) if x < pivot]
       middle=[x for x in I(X_i;y) if x== pivot]
       right=[x for x in I(X_i;y) if x > pivot]
       return QuickSort(left)+middle+ QuickSort(right)
5. C=Select(k,I(X_i;y))//选择中心特征
6. 结束
```

为了使对抗网络流量能够干扰判别器的分类结果,同时保持与原始流量之间的最小差异,生成器的损失函数 $Loss$ 表示为 L_G 与 L_S 的加权和:

$$Loss=\lambda L_G+\mu L_S \tag{9}$$

其中,λ 和 μ 均为系数。当判别器正确分类时,$\lambda=1$ 且 $\mu=0$。此时的对抗网络流量还未能成功干扰判别器的识别,为了提高对抗网络流量的欺骗性,需要增大对抗网络流量与真实流量在特征空间中的距离。反之,当判别器对对抗网络流量误分类时,$\lambda=0$ 且 $\mu=1$。为了优化对抗网络流量的生成质量,通过约束 C' 与 C 之间的距离以最小化类内差异,减小对抗网络流量与原始流量之间的差异。

3.2 基于网格搜索算法的干扰器设计

经生成器生成的基础扰动已具备原始流量特性。然而,为了进一步赋予对抗网络流量足够的差异性和多样性,需对基础扰动进行赋权处理。在此过程中,为确保对抗网络流量的生成质量,必须控制扰动的幅度,以实现权重分配与扰动幅度的双重管理。因此,本文干扰器通过权重分配融合生成器产生的基础扰动与原始流量,并优化参数组合,最终输出对抗网络流量。

$$x'=(r+t)\cdot z+x \tag{10}$$

其中,r 为扰动因子,t 为步长,z 为生成器生成的基础扰动,x 为原始流量,x' 为受干扰后的对抗网络流量。由式(10)可以看出,干扰器旨在调控扰动范围,因此扰动因子与步长的选择至关重要。为简化算法实现,本文将两者均设为整数。针对扰动因子大小选择的问题,采用基尼不纯度特征重要性公式进行初步预估,公式如下。

$$I(X_i)=\frac{\sum\limits_{t\in all\_trees}\Delta Gini(t,X_i)}{\sum\limits_{t\in all\_trees}\sum\limits_{k\in all\_features}\Delta Gini(t,X_k)} \tag{11}$$

其中,$\Delta Gini(t,X_i)$ 表示在树 t 节点分裂时,使用特征 X_i 进行分裂所带来的基尼不纯度减少量。在本文所使用的数据集中,特征 Entropy 的重要性最高,评分分别为 0.27 与 0.17。由实验分析显示,该特征的最大值与均值间绝对差约为 20,且数据多集中于均值之下。因此,将扰动因子限制在 20 内,有助于生成对抗网络流量更贴近实验数据集分布,并避免过

大扰动使流量过于偏离实际数据。这一限制能够确保生成的对抗流量既具有足够的多样性，又能保持其原始流量的特征，从而提高生成的流量在网络中的自然性和逼真性。由于本文选择的参数空间较小且都是离散的值，为了快速找到最优解，使用网格搜索算法遍历扰动因子与步长的组合。这种设置使得模型在训练中能够注重基础扰动的生成，而非仅依赖于增大扰动强度来提高对分类模型的欺骗率。网格搜索通过遍历各个组合，可以在较小的参数空间内高效找到最优的扰动因子与步长，从而确保在有限的时间和计算资源下，生成的对抗流量既能有效地欺骗分类模型，又不会过度偏离原始流量的分布，最终提高对抗网络流量的生成质量。网格搜索的过程如图 3 所示。

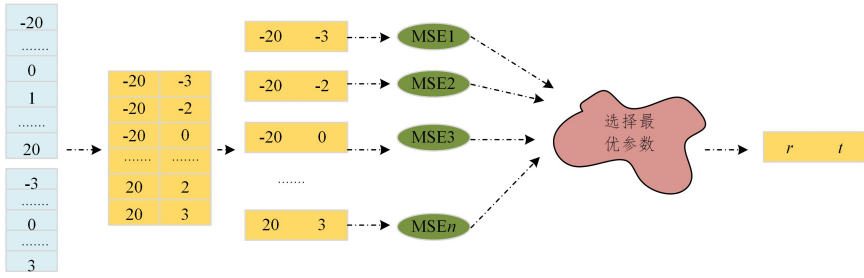


图 3　网格搜索算法过程

Fig. 3　Grid search algorithm process

网格搜索算法由 4 个步骤组成。

1）定义网格参数

定义扰动因子 r 的范围为 $R=[-20,20]$，步长 t 的范围为 $T=[-3,3]$，并且 $r+t\neq 0$。

2）网格搜索

使用笛卡尔积在参数网格上遍历所有的参数组合：

$$R\times T=\{(r,t)\mid r\in R,t\in T,r+t\neq 0\} \tag{12}$$

3）结果评估

选择均方误差 MSE 值计算对抗网络流量特征向量与原始流量特征向量之间的误差值：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(x_i'-x_i)^2 \tag{13}$$

其中，n 是流量中特征的数量，x_i 是原始流量特征向量中的第 i 个元素，x_i' 是对抗网络流量特征向量中的第 i 个元素。

4）选择最优参数组合

选择 MSE 值最小的一组对应的扰动因子 r 与步长 t 的参数组合。

3.3　判别器识别

判别器模型为事先训练好的分类模型。判别器模型的分类任务训练过程由 2 个步骤组成。

1）数据预处理

将流量数据的特征与标签进行分离，随后应用 MinMaxScale 归一化算法对特征进行处理。该算法通过线性变换，将每个特征的值域缩放到[0，1]，其中特征的最小值被转换为 0，最大值被转换为 1。

2）分类

使用交叉熵损失函数对流量进行分类：

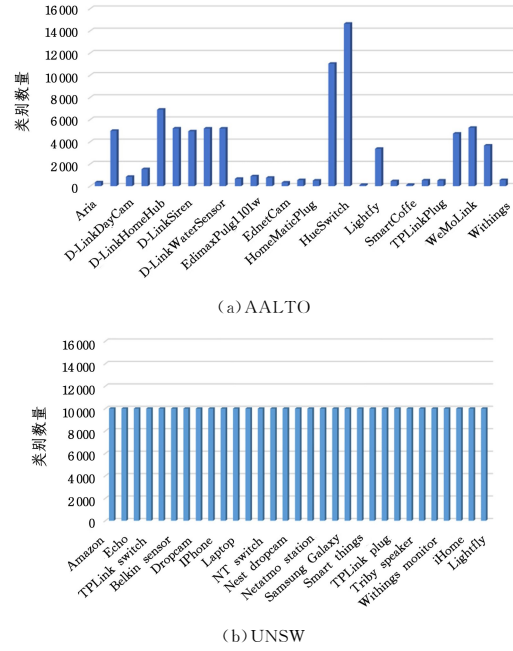
$$J(W,b)=-\frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n}y_{i,j}\log_2(\hat{y}_{i,j}) \tag{14}$$

其中，m 是样本数量，n 是类别数量，$y_{i,j}$ 是真实标签，$y_{i,j}^{\wedge}$ 是模型输出。

4　实验与分析

4.1　数据集

在实验中，使用了 AALTO[18] 和 UNSW[19] 两个数据集，如图 4 所示。AALTO 和 UNSW 数据集包含来自网关、传感器等多种设备的流量数据，以 pcap 格式文件作为数据载体。



(a) AALTO

(b) UNSW

图 4　数据集中的类别数量分布

Fig. 4　Distribution of category counts across datasets

AALTO 数据集包含 27 种不同设备类型的流量数据，涉及 D-Linksensor 和 HueSwitch 等设备类型，由图 4(a)可知，每一种类型的流量样本数量均不一致，最多的类别数量有 14649 个，最少的仅 121 个。

UNSW 数据集包含 31 种不同设备类型的流量数据，涉及 TPLink Router Bridge LAN 和 HP Printer 等设备类型，由图 4(b)可知，每一种设备类型均有 10000 条流量样本。

4.2　评价指标

选定的实验评估指标包括欺骗率、扰动量及相对差异扰动量、FID、SSIM。

欺骗率 FR 为：

$$FR=\frac{FP+FN}{TP+TN+FP+FN} \tag{15}$$

其中，TP 为预测结果为阳性类的阳性样本数；TN 为预测结果为阴性类的阴性样本数；FP 为预测结果为阳性类的阴性

样本数；FN 为预测结果为阴性类的阳性样本数。

扰动量 L_p 为：

$$L_p = \| x \|_p = \left(\sum_{i=1}^{n} \| x_i \|^p\right)^{\frac{1}{p}} \tag{16}$$

其中，L_0 用于计算原始流量中受到扰动的特征数量；L_2 用于计算对抗网络流量与原始流量之间的欧氏距离；L_∞ 用于计算受扰动特征中的最大扰动值。

由上述公式可以看出，L_2 范数通过计算所有特征值绝对差异的平方和的平方根来衡量特征值之间的总体差异。然而，这种计算方式存在一个潜在的问题：即使两个特征值之间的绝对差异很小，但如果这个差异相对于一个较大的原始特征值而言并不显著，L_2 范数仍可能将其视为一个极其重要的差异。

为克服 L_2 范数在评估对抗网络流量与原始流量特征差异时的局限，提出了一种以原始特征值为基准的相对变化评估指标。该指标综合考虑绝对差异与原始特征值的比例，通过计算变化率来精细评估特征变化，避免特征值大小的不公平影响，实现更公平准确的评估。相对差异扰动量 RDD_p 为：

$$RDD_p = (1-p)\frac{\sum_{i=0}^{n}\frac{\| x_i' - x_i \|}{x_i + \varepsilon}}{n} + p \max_i\left(\frac{\| x_i' - x_i \|}{x_i + \varepsilon}\right) \tag{17}$$

其中，n 是特征的数量；x_i 是原始流量特征向量中的第 i 个元素，x_i' 是对抗网络流量特征向量中的第 i 个元素；ε 为一个接近于零的数；RDD_0 用于计算对抗网络流量与原始流量之间的相对差异扰动均值；RDD_1 用于计算最大相对差异扰动值。

为了量化对抗网络流量与原始流量在灰度图表示上的差异，本文采用以下两个评估指标。

Fréchet 感知距离 FID 为：

$$FID = \| \mu_x - \mu_{x'} \|^2 + \mathrm{tr}(\Sigma x + \Sigma x' - 2\Sigma x \Sigma x')^{\frac{1}{2}} \tag{18}$$

结构相似性指数 SSIM 为：

$$SSIM(x, x') = \frac{(2\mu_x\mu_{x'} + C_1)(2\sigma_{xx'} + C_2)}{(\mu_x{}^2 + \mu_{x'}^2 + C_1)(\sigma_x{}^2 + \sigma_{x'}^2 + C_2)}$$

其中，μ_x 为图像的均值；σ 为图像的方差；C 为常数；$tr(\cdot)$ 为矩阵的迹。

4.3 实验设计

为了验证本文方法的有效性，设计了如下 3 个实验。

1）评估对抗网络流量的有效扰动范围：通过变化 L_∞ 范数下的扰动强度，分析 FGSM[7]，Vanilla GAN[14]，PGD[11]，GAN[17] 及 Control-GAN 5 种方法生成的对抗网络流量对 KNN 分类模型的欺骗率变化，并应用 Hurst 指数验证对抗网络流量与原始流量之间的统计显著性差异。

2）验证对抗网络流量对目标分类模型的干扰效果：在有效扰动范围内，对比分析 Control-GAN 与 FGSM[7]，Vanilla GAN[14]，PGD[11] 和 GAN[17] 等方法在不同数据分布下对多种分类模型（KNN、随机森林、决策树、CNN＋GRU）的欺骗效果。其中，UNSW 中每一种类型的流量样本数量均相同；AALTO 中每一种类型的流量样本分布不均衡；4 种分类模型在 UNSW 测试集上的总体准确率分别为 91.78%，92.49%，92.43% 和 88%；在 AALTO 测试集上的准确率表现分别为 85.6%，95.28%，95.74% 和 90%。

3）评估对抗网络流量的生成质量：在较高欺骗率下，对比分析 Control-GAN 与 FGSM[7]，Vanilla GAN[14]，PGD[11] 以及 GAN[17] 生成的对抗网络流量质量。通过计算 L_0，L_2 及 L_∞ 指标，反映流量在特征空间中的绝对距离差异；通过计算 RDD_0 及 RDD_1 指标，衡量流量在特征空间中的相对距离变化速率；利用灰度图可视化不同方法生成的对抗网络流量与原始流量之间的差异，计算每组灰度图的 FID 和 SSIM 指标，评估其相似性。

4.4 实验结果与分析

为了有效评估不同对抗网络流量对分类模型的欺骗效果与生成质量，首先需要界定各自的有效扰动界限。超出此界限，对抗网络流量与原始流量间的统计相关性显著减弱甚至消失，使得对抗网络流量丧失有效性。

4.4.1 对抗网络流量的有效扰动范围实验结果与分析

为了减少模型差异导致的误差，增强实验结果的可靠性和可重复性，进行了预实验。实验结果显示，KNN 分类模型在评估 5 种对抗网络流量模型生成的对抗网络流量时表现出更高的稳定性，因此选定 KNN 作为分类基准，其中 K 值为 10，距离度量方式为 Minkowski。鉴于 L_2 扰动量受 L_∞ 扰动量调控，本实验通过调整 L_∞ 范数扰动强度来探究影响。Hurst 指数是流量数据中最显著的统计特征，用于描述自相似性[20]。本文运用经典的 R/S 分析方法计算该指数，实验结果如表 1 所列。

表 1 实验结果

Table 1 Experimental results

(a) UNSW experimental results

L_∞	方法（欺骗率，原始流量 Hurst exponent，对抗网络流量 Hurst exponent）				
	FGSM	Vanilla GAN	PGD	GAN	Control-GAN
10	(0.76, 0.24, 0.11)	(0.9, 0.24, 0.1)	(0.83, 0.24, 0.18)	(0.8, 0.24, 0.12)	(0.93, 0.24, 0.2)
20	(0.82, 0.24, 0.19)	(0.91, 0.24, 0.24)	(0.86, 0.24, 0.22)	(0.83, 0.24, 0.25)	(0.97, 0.24, 0.26)
50	(0.9, 0.24, 0.23)	(0.95, 0.24, 0.3)	(0.9, 0.24, 0.28)	(0.9, 0.24, 0.48)	(0.97, 0.24, 0.47)
100	(0.96, 0.24, 0.5)	(0.96, 0.24, 0.5)	(0.94, 0.24, 0.5)	(0.94, 0.24, 0.5)	(0.96, 0.24, 0.55)
200	(0.98, 0.24, 0.58)	(0.97, 0.24, 0.69)	(0.98, 0.24, 0.61)	(0.96, 0.24, 0.56)	(0.97, 0.24, 0.67)

(b) AALTO experimental results

L_∞	方法（欺骗率，原始流量 Hurst exponent，对抗网络流量 Hurst exponent）				
	FGSM	Vanilla GAN	PGD	GAN	Control-GAN
10	(0.62, 0.19, 0.06)	(0.74, 0.19, 0.19)	(0.67, 0.19, 0.18)	(0.58, 0.19, 0.27)	(0.8, 0.19, 0.09)
20	(0.68, 0.19, 0.17)	(0.79, 0.19, 0.24)	(0.72, 0.19, 0.26)	(0.7, 0.19, 0.49)	(0.87, 0.19, 0.17)
50	(0.8, 0.19, 0.5)	(0.8, 0.19, 0.3)	(0.77, 0.19, 0.34)	(0.8, 0.19, 0.55)	(0.87, 0.19, 0.45)
100	(0.82, 0.19, 0.54)	(0.82, 0.19, 0.5)	(0.82, 0.19, 0.5)	(0.82, 0.19, 0.69)	(0.87, 0.19, 0.68)
200	(0.89, 0.19, 0.62)	(0.86, 0.19, 0.76)	(0.83, 0.19, 0.6)	(0.83, 0.19, 0.82)	(0.87, 0.19, 0.74)

从上述实验结果可知，随着 L_∞ 扰动强度的增加，上述 5 种不同的对抗网络流量生成模型所生成的对抗网络流量对 KNN 分类模型的欺骗率逐渐增加，但对抗网络流量的 Hurst 指数与原始流量差异亦逐渐扩大。当对抗网络流量的 Hurst 指数超过 0.5 时，对抗网络流量呈现随机性，与原始流量无相关性，两者间的统计差异显著。

根据表 1 数据，在 L_∞ 扰动量设定为 50 的条件下，本文方法生成的对抗网络流量在两个数据集上的 Hurst 指数表现均

未超 0.5，欺骗率却几乎保持不变，因此本文将 L_∞ 有效扰动强度上限设定为 20。为便于后续实验进行有效对比，表 2 列出了 5 种对抗网络流量生成方法在 UNSW 和 AALTO 数据集上的有效扰动范围上限值。

表 2　有效扰动范围实验结果

Table 2　Experimental results of effective perturbation range

方法	L_∞ 有效扰动强度上限值	
	UNSW	AALTO
FGSM	100	50
Vanilla GAN	100	100
PGD	100	100
GAN	100	50
本文方法	20	20

4.4.2　对抗网络流量的欺骗率对比实验结果与分析

当样本数量分布均衡时，本文方法生成的对抗网络流量与其他 4 种方法生成的对抗网络流量对不同分类模型的欺骗率实验结果如图 5 所示。

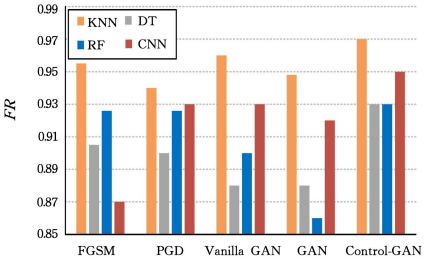


图 5　对抗网络流量欺骗率对比

Fig. 5　Comparison of adversarial network traffic fraud rates

在未引入扰动的原始状态下，4 种分类算法在 UNSW 数据集上展现出卓越的识别性能，其中随机森林达到 92.49% 的准确率。然而，如图 5 所示，当原始流量经过 FGSM，Vanilla GAN，PGD，GAN 及本文方法的对抗训练后，4 种分类模型的表现均遭受显著影响。KNN、决策树、随机森林和神经网络模型的准确率分别下降至 3%～6%，7%～12%，7%～14% 以及 7%～13%。

当样本数量分布不均衡时，图 6 给出了 5 种不同方法生成的对抗网络流量对不同分类模型的总体欺骗率结果。

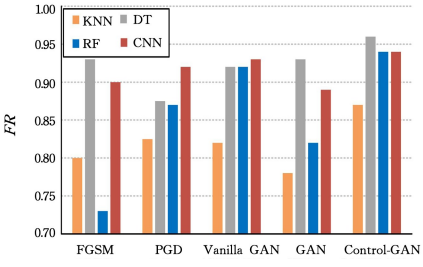


图 6　对抗网络流量欺骗率对比

Fig. 6　Comparison of adversarial network traffic fraud rates

当流量样本分布不均衡时，4 种分类模型对原始流量的分类准确率分别为 85.6%，95.28%，95.74% 和 90%。然而，在进行干扰后，4 种分类模型的识别准确率均降至 20% 左右。在 KNN 模型上，本文方法生成的对抗网络流量达到了 87.12% 的欺骗率，与 FGSM，Vanilla GAN，PGD，GAN 4 种方法相比，提升了 7% 左右。在随机森林模型上，本文方法的优势更为明显，总体欺骗率分别提升了 21.26%，2%，7%，12.38%。

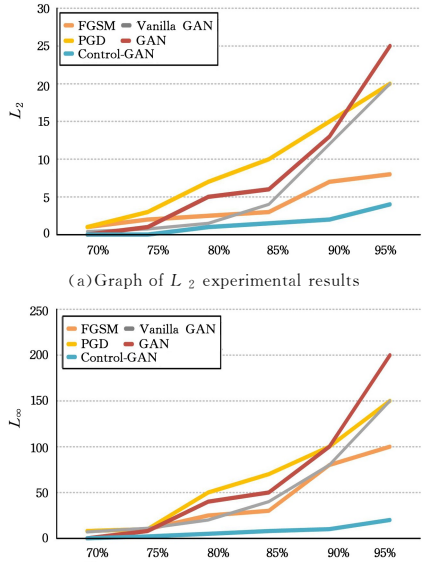
综上，当数据集分布均衡时，本文模型能够全面学习并表征各类数据的特征，使得生成的对抗网络流量能够广泛覆盖特征空间，从而使对抗样本的干扰效果更加稳定。然而，当数据集出现不均衡现象时，FGSM，Vanilla GAN，PGD，GAN 4 种方法往往倾向于过度关注多数类数据，导致对少数类数据的特征学习不足，模型的泛化能力在少数类上受限。而本文方法利用卷积神经网络深入学习和理解原始流量的内在结构和分布特征，即使在数据分布不均衡的情况下，也能够有效地捕捉并聚焦于少数类别数据的特征学习，从而增强模型的整体泛化能力和鲁棒性。在 5 种对抗网络流量生成方法中，本文方法展现出了最优的对抗性能。具体而言，与 FGSM 方法相比，本文方法通过 BIM 算法的迭代优化策略，克服了初始点敏感性问题，这种迭代优化的方式使得本文方法能够更准确地捕捉流量数据的复杂特征。尽管 PGD 方法也采用了迭代优化，但本文方法更注重权重的合理分配与融合，从而实现了更高效的优化。与 Vanilla GAN 及 GAN 方法相比，本文方法通过调整生成器训练过程并融入权重分配与融合机制，解决了 GAN 在训练稳定性上的问题。

4.4.3　对抗网络流量的生成质量对比实验结果与分析

在 KNN 分类模型下，采用 GAN，FGSM，Vanilla GAN，PGD 及本文方法对原始流量实施干扰，生成对抗网络流量。随着欺骗率从 70% 增至 95%，5 种方法所生成的对抗网络流量在 L_p 及 RDD_p 下展现出不同的变化趋势，具体实验结果与分析如下。

1）L_0，L_2 与 L_∞ 扰动量结果对比实验

由于 5 种方法均全面扰动所有特征，因此各自的 L_0 值均与特征总数一致。图 7 给出了 5 种不同对抗网络流量生成方法在不同欺骗率水平下所产生的 L_2 范数与 L_∞ 范数扰动量的变化趋势。



(a)Graph of L_2 experimental results

(b)Graph of L_∞ experimental results

图 7　L_p 实验结果

Fig. 7　Experimental results of L_p

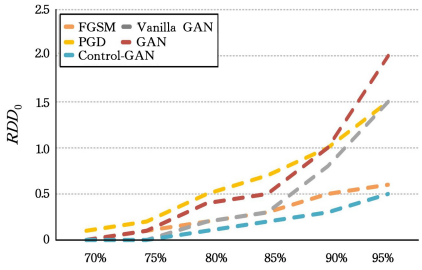
由图 7(a)可知，在提高欺骗率的同时，5 种方法对于单个特征的 L_2 扰动量均呈上升趋势，其中，本文方法在每一个阶段对单个特征产生的 L_2 扰动量均是最小的。当分类模型的欺骗率达到 95%时，相较于 FGSM 和 PGD 方法，本文方法通过精细迭代，优化步长选择以及优化的目标函数，降低了单个

特征上的 L_2 扰动量，减少量分别达到了 4.68，4.7 和 16.54。传统的 GAN 模型及 Vanilla GAN 在生成对抗网络流量时，缺乏对生成质量的精细控制。本文提出的双目标约束函数方法通过确保对抗网络流量与原始流量在中心特征空间中的类内相似性，使得引入的扰动量最小。这种方法提高了对抗网络流量的自然性，所需的 L_2 扰动量分别减少了 21.7 与 16。因此，本文方法以更小的 L_2 扰动量实现了相同的欺骗率，整体表现更佳。其次，从不同方法产生的 L_2 扰动量曲线变化幅度可以看出，当欺骗率增加时，本文方法的 L_2 扰动量变化更为平缓。

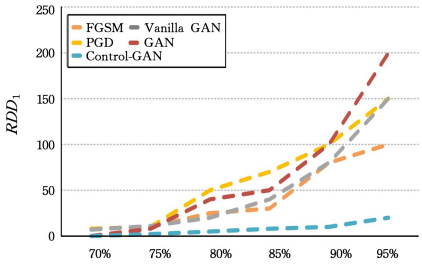
从图 7(b)中可以看出，对分类模型的总体欺骗率逐渐增加时，不同方法对单个特征干扰前后产生的 L_∞ 扰动量存在较大差异。本文方法采用相似性保持目标函数作为约束，增强了对抗网络流量生成质量的可控性。该方法避免了过度增加扰动量以提升欺骗率，从而减小了对抗网络流量的扰动量变化，确保单个特征的 L_∞ 扰动量始终处于预定范围内。

2)相对差异扰动量 RDD_p 结果对比实验

图 8 给出了 5 种方法在不同欺骗率增长情况下，所生成的对抗网络流量对应的 RDD_0 与 RDD_1 扰动量指标的变化情况。



(a)Graph of RDD_0 experimental results



(b)Graph of RDD_1 experimental results

图 8 RDD_p 实验结果

Fig. 8 Experimental results of RDD_p

根据图 8(a)的数据，本文方法在 RDD_0 指标下展现出的 RDD_0 扰动量值均低于其他 4 种方法，这表明本文方法生成的对抗网络流量与原始流量在特征上的差异更小且更均衡，整体上未在特征中引入过大扰动。相较于仅衡量特征距离变化的 L_2 扰动量，RDD_0 通过评估特征的距离变化速率提供了更为全面的评估。同时，在 RDD_1 指标下，该指标聚焦于选取单个特征最大扰动变化速率。鉴于本文实验数据集中包含特征值等于 1 的样本，RDD_1 所反映的相对差异扰动量变化与 L_∞ 扰动量在整体趋势上呈现一致性。实验结果显示，在上述两个评估指标下，本文方法的整体表现均优于其他方法，其生成的对抗网络流量质量更高。

3)基于灰度图的相似性对比

为直观展现扰动增加后对抗网络流量与原始流量的可视化差异，本文借鉴文献[7]的方法，采用灰度图对比。在分类模型达到最高欺骗率时，图 9 给出了 5 种方法生成的对抗网络流量灰度图对比结果。

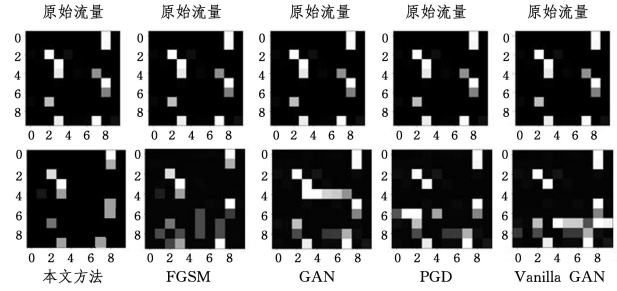


图 9 对抗网络流量灰度图

Fig. 9 Grayscale image of adversarial network traffic

表 3 列出了上述 5 种方法的对抗网络流量灰度图与原始流量灰度图之间的相似性指标量化结果。

表 3 生成相似性量化结果

Table 3 Quantitative results of generation similarity

方法	*FID*	*SSIM*
FGSM	3.502	0.064
Vanilla GAN	2.372	0.042
PGD	7.638	0.086
GAN	6.390	0.066
本文方法	1.137	0.464

由表 3 可知，本文方法生成的对抗网络流量灰度图与原始流量的 SSIM 值为 0.464，相较于其他 4 种方法中的最佳表现，本文方法提升了 0.422，表明两者的结构相似性指数更高；在 FID 指标下，本文方法得到的 1.137 均低于其他 4 种方法，表明两者在特征空间的分布差异更小，生成质量更高。与 FGSM，Vanilla GAN，PGD 以及 GAN 4 种方法相比，本文模型通过引入相似性保持目标函数，有效减小了对抗网络流量与原始流量的差异，从灰度图对比来看，其在 FID 和 SSIM 两种指标下表现更加优异。这表明本文方法在实现有效欺骗的同时，维持了对抗网络流量与原始流量的较高相似性。

结束语 本文提出一种对抗网络流量生成模型 Control-GAN，模型集成了卷积神经网络和 BIM 算法来生成基础扰动，并设计干扰器为基础扰动分配权重，与原始流量进行加权和，从而生成对抗网络流量。通过引入相似性保持目标函数，优化了生成器的损失反馈，减小了对抗网络流量与原始流量之间的差异。Control-GAN 有效解决了传统基于 GAN 的对抗网络流量生成模型生成的对抗网络流量质量不佳的问题，与现有的对抗网络流量生成方法相比，本文方法的欺骗率、扰动量大小等指标都得到了有效提高。在后续的研究中，将进一步尝试优化和完善现有模型。

参 考 文 献

[1] ZHANG W, LIU Y, FENG Y, et al. A Comprehensive Review of Network Reconnaissance and Defense Technologies [J]. Communications Technology, 2022, 55(10): 1247-1256.

[2] HUI S, WANG H, WANG Z, et al. Knowledge Enhanced GAN for IoT Traffic Generation[C]// Proceedings of the ACM Web Conference. 2022: 3336-3346.

[3] JIA Z P, FANG B X, LIU C G, et al. Overview of Network De-

ception Techniques [J]. Journal of Communications, 2017, 38(12):128-143.

[4] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets[C]// 28th Conference on Neural Information Processing Systems(NIPS). 2014.

[5] WANG J, LU B, ZHU Y F. A Survey on the Generation and Application of Adversarial Network Traffic [J]. Computer Science, 2022, 49(S2):651-661.

[6] ZHANG X, HAMM J, REITER M K, et al. Statistical Privacy for Streaming Traffic[C]// Proceedings of the 26th ISOC Symposium on Network and Distributed System Security. 2019.

[7] HU Y J, GUO Y B, MA J, et al. Method for Generating Network Deceptive Traffic Based on Adversarial Samples [J]. Journal of Communications, 2020, 41(9):59-70.

[8] MOORE A, ZUEV D, CROGAN M. Discriminators for use in flow－based classification[EB/OL]. https://www.cl.cam.ac.uk/～awm22/publication/RR-05-13.pdf.

[9] RIGAKI M. Adversarial Deep Learning Against Intrusion Detection Classifiers[C]// 2017 NATO IST-152 Workshop on Intelligent Autonomous Agents for Cyber Defence and Resilience (IST-152 2017). 2017.

[10] HUANG C H, LEE T H, CHANG L, et al. Adversarial Attacks on SDN-based Deep Learning IDS System [C]// Mobile and Wireless Technology 2018: International Conference on Mobile and Wireless Technology(ICMWT 2018). Springer, 2019: 181-191.

[11] IBITOYE O, SHAFIQ O, MATRAWY A. Analyzing Adversarial Attacks Against Deep Learning for Intrusion Detection inIoT Networks[C]// 2019 IEEE Global Communications Conference (GLOBECOM). IEEE, 2019: 1-6.

[12] LI J, ZHOU L, LI H, et al. Dynamic Traffic Feature Camouflaging Via Generative Adversarial Networks [C]// 2019 IEEE Conference on Communications and Network Security(CNS). IEEE, 2019: 268-276.

[13] CHOUGULE A, AGRAWAL K, CHAMOLA V. Scan-gan: Generative Adversarial Network Based Synthetic Data Generation Technique for Controller Area Network[J]. IEEE Internet of Things Magazine, 2023, 6(3):126-130.

[14] ANANDE T J, AL-SAADI S, LEESON M S. Generative Adversarial Networks for Network Traffic Feature Generation[J]. International Journal of Computers and Applications, 2023, 45(4): 297-305.

[15] HUI S, WANG H, WANG Z, et al. Knowledge Enhanced GAN forIoT Traffic Generation[C]// Proceedings of the ACM Web Conference 2022. 2022: 3336-3346.

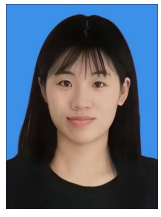
[16] HUANG Y, CHEN Y, ZHANG Y, et al. TFHM: A Traffic Feature Hiding Scheme Based on Generative Adversarial Networks [C]// 2022 7th IEEE International Conference on Data Science in Cyberspace(DSC). IEEE, 2022: 175-182.

[17] LI J, ZHOU L, LI H X, et al. Network Traffic Feature Camouflage Technology Based on Generative Adversarial Networks [J]. Computer Engineering, 2019, 45(12): 119-126.

[18] SIVANATHAN A, SHERRATT D, GHARAKHEILI H H, et al. Characterizing and Classifying IoT Traffic in Smart Cities and Campuses[C]// 2017 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). IEEE, 2017: 559-564.

[19] MIETTINEN M, MARCHAL S, HAFEEZ I, et al. IoT Sentinel: Automated Device-type Identification for Security Enforcement in IoT[C]// 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017: 2177-2184.

[20] LI J, WANG D, LI S, et al. Deep Learning Based Adaptive Sequential Data Augmentation Technique for the Optical Network Traffic Synthesis [J]. Optics Express, 2019, 27 (13): 18831-18847.



YANG Lin, born in 1999, postgraduate. Her main research interests include adversarial network traffic generation and network security.

LIN Honggang, born in 1976, Ph.D, professor. His main research interests include cloud computing, big data security, artificial intelligence and cybersecurity.