基于生成对抗网络的网络攻击图像仿真与防御测试方法

白小龙　卞锦豪　刘梦琦
西安市物联网应用实验室　陕西西安　710000

摘　要：本文提出了一种基于生成对抗网络（GAN）的网络攻击图像仿真与防御测试方法。通过构建深度卷积生成对抗网络（DCGAN），实现了对钓鱼邮件、恶意网页等攻击场景的高仿真图像生成，并结合动态特征注入技术确保攻击载荷的有效性。在防御测试方面，设计了混合虚拟化测试环境，支持自动化攻击注入与响应分析，并提出了基于检测率、误报率和阻断时效的量化评估指标。进一步融合对抗攻击技术生成扰动样本，通过对抗训练增强了防御模型的鲁棒性。实验表明，该方法生成的攻击图像在视觉真实性和攻击特征完整性上均达到较高水平，且防御测试系统能有效评估安全设备的性能，为网络安全攻防研究提供了可扩展的实践平台。

关键词：生成对抗网络（GAN）；网络攻击仿真；防御测试；对抗样本；虚实融合

一、攻击图像生成技术框架

（一）生成对抗网络核心结构

生成对抗网络（GAN）通过双模块博弈机制实现数据合成，其核心由生成器与判别器构成。生成器采用深度卷积结构，通过转置卷积层逐步上采样潜在噪声向量，构建具备空间相关性的像素矩阵。判别器则利用多层卷积核提取图像特征，输出真伪判别概率。在训练阶段，两者通过最小化对抗损失函数进行零和博弈：生成器试图伪造逼近真实攻击样本的图像，判别器则持续提升对合成数据的鉴别能力。实际应用中，采用谱归一化技术稳定训练过程，避免梯度消失或模式崩溃。卷积层参数通过迁移学习初始化，加速攻击场景特征的学习效率。

（二）攻击场景特征编码

针对典型网络攻击形态，构建结构化特征编码方案。在钓鱼邮件场景中，提取邮件头部的发件人伪造字段、正文中的超链接结构以及附件文件的MIME类型等关键要素。对于恶意网页攻击，重点捕获DOM树异常节点、JavaScript代码混淆模式及HTTP请求重定向链。通过动态特征注入机制，将攻击载荷嵌入生成流程：在GAN训练阶段，将特征编码转换为条件向量输入生成器，控制特定攻击元素的生成概率。例如，在生成伪造银行登录页面时，通过调整URL长度、表单字段名称等参数，实现域名仿冒程度与界面布局的精细化控制。

（三）仿真图像生成流程

数据预处理阶段对真实攻击样本进行解析，提取屏幕截图、网络流量转储等原始素材。采用自适应直方图均衡化增强图像对比度，消除不同采集设备带来的亮度差异。噪声输入优化模块通过傅里叶变换将高斯噪声映射至频域，保留低频分量以控制生成图像的全局结构。在图像合成阶段，生成器输出经双线性插值调整分辨率，并通过风格迁移网络增强纹理细节。后处理环节引入可微分渲染器，模拟浏览器渲染引擎的字体抗锯齿效果及CSS样式解析偏差，使生成的恶意网页截图在视觉层面与真实设备捕获结果保持像素级一致。最终通过多维度校验机制，确保生成图像包含完整的攻击载荷且不触发常规反病毒引擎的静态检测规则。

二、高仿真攻击场景构建方法

（一）多场景攻击模板设计

基于公开威胁情报库构建标准化攻击模板体系，覆盖钓鱼邮件、水坑攻击、供应链投毒等典型攻击向量。模板设计采用模块化架构，将攻击载荷分解为静态元素（如邮件正文模板、网页框架代码）与动态参数（如域名生成算法、社会工程学话术）。通过参数化配置引擎实现场景定制：攻击者可调整C2服务器地址轮询策略、载荷加密方式及传播渠道权重。例如，在钓鱼邮件模板中，支持正则表达式驱动的发件人域名伪造规则，以及基于Markov链生成的逼真话术段落。模板库采用版本控制机制，定期集成CVE漏洞利用代码及新型攻击战术，确保

基金项目：智能峰顶导航登山杖（编号：S202411080103）
作者简介：白小龙（2004.09-），男，汉，陕西省榆林市，本科生，研究方向：网络安全。

仿真场景与实际威胁演进保持同步。

（二）动态行为模拟

构建攻击行为时序引擎，将APT攻击链拆解为初始访问、横向移动、数据窃取等阶段，每个阶段关联预定义行为脚本。采用行为树架构管理攻击流程，支持条件分支（如根据防御设备响应选择继续渗透或终止攻击）与循环操作（如定期尝试权限提升）。在用户交互层面，开发基于Selenium的浏览器自动化模块，模拟真实用户点击伪造按钮、填写表单等操作轨迹。针对钓鱼页面，实现鼠标悬停动态效果、表单提交后的重定向链等细节仿真。时序引擎集成网络延迟模拟功能，通过调整TCP握手时间、数据包重传率等参数，复现不同网络环境下的攻击行为特征。

（三）成效果评估指标

建立三维评估体系：视觉保真度采用结构相似性指数（SSIM）与感知哈希算法，量化生成图像与真实攻击截图在像素分布及语义特征上的差异。攻击载荷完整性通过静态代码分析验证，检查生成的恶意网页是否包含预期的JavaScript漏洞利用代码，钓鱼邮件是否携带正确格式的跟踪像素。行为仿真度评估则基于攻击链覆盖率指标，统计自动化测试中成功触发的TTPs（战术、技术、程序）数量。设计可视化评估工作台，集成OpenCV图像比对工具与MITRE ATT&CK知识库，支持研究人员通过交互式界面调整评估权重，生成包含缺陷定位标注的评估报告。

三、防御测试环境与流程设计

（一）测试环境拓扑架构

构建虚实结合的混合测试网络，核心节点采用KVM虚拟化技术部署，包含模拟终端主机、软件定义路由器及开源防火墙系统。物理层通过TAP接口实现虚拟网络与真实设备的互联，例如将蜜罐传感器接入虚拟子网，捕获穿透虚拟防火墙的攻击流量。网络拓扑支持动态重构，通过Python脚本调用libvirt API，可在分钟级完成星型、网状等拓扑切换。为模拟复杂企业网络，部署多级代理架构：前端Web服务器承受攻击流量，后端数据库服务器通过VLAN隔离，中间件节点运行带漏洞的旧版Apache Struts框架。所有节点启用进程级监控，记录攻击载荷执行过程中的系统调用序列。

（二）自动化测试流程

开发基于Jenkins的持续测试管道，实现攻击图像生成、环境部署、测试执行、结果分析全流程自动化。在攻击注入阶段，采用分布式爬虫架构，从Ceph存储集群并行读取预生成的攻击图像数据集，通过修改IP头部的TTL字段实现流量分片，避免单一节点触发流量阈值警报。防御设备响应捕获模块集成Bro IDS规则引擎，实时解析NetFlow数据与PCAP全包捕获文件。开发自定义YARA规则，对日志中的恶意URL、异常出站连接等行为进行标记。测试结果通过InfluxDB时序数据库存储，支持通过Grafana面板进行可视化回溯。

（三）防御效能评估体系

设计三维评估矩阵：基础维度包含检测率（TP/（TP+FN））与误报率（FP/（FP+TN）），采用滑动窗口算法消除瞬时流量波动影响。时效维度引入MTTD（平均检测时间）与MTTR（平均响应时间）指标，通过在攻击流量中嵌入时间戳水印，精确计算防御系统的事件处置延迟。业务维度则聚焦攻击阻断对正常服务的影响，通过SYN重传率、HTTP响应延迟等参数，量化旁路检测机制与主动防御策略的适用场景。最终评估报告生成模块采用Jinja2模板引擎，可自动生成包含攻击路径热力图、设备性能衰减曲线的定制化报告，辅助安全团队定位防御体系薄弱环节。

四、对抗样本生成与防御增强

（一）对抗攻击技术融合

在攻击图像生成流程中集成快速梯度符号法（FGSM），通过计算输入图像像素值的梯度方向，生成微小扰动噪声。该技术模拟攻击者对防御模型弱点的探测过程，在邮件正文图片的特定区域嵌入人眼难以察觉的对抗纹理，使光学字符识别（OCR）系统产生字符分类错误。进一步开发多扰动叠加算法，将FGSM生成的噪声与基于生成对抗网络的图像变异相结合，构建包含多种攻击变体的测试样本库。在混合训练阶段，设计交替训练策略：防御模型先通过标准攻击图像进行预训练，再引入对抗样本进行微调，增强模型对边缘案例的识别能力。

（二）防御模型优化策略

构建包含对抗样本的增强训练数据集，采用半监督学习框架扩大数据覆盖范围。首先通过无监督聚类算法自动标注未标记的攻击图像，然后利用对抗训练技术生成伪标签样本。在模型优化层面，实施集成学习策略，将多个基于不同初始化的卷积神经网络（CNN）进行投票融合。实验表明，该架构在面对未知对抗样本时，分类置信度较单一模型提升。针对模型过拟合问题，引入

弹性网络正则化项，在损失函数中同时约束L1和L2范数，平衡特征选择与权重衰减效应。

（三）动态防御机制集成

开发基于行为基线的实时监控模块，通过系统调用序列分析检测异常进程行为。该模块采用隐马尔可夫模型（HMM）对正常用户操作进行建模，当检测到偏离基线的文件访问模式或注册表修改操作时，触发防御响应流程。构建闭环反馈系统，将防御设备的日志数据实时注入攻击生成引擎，实现攻击策略的动态调整。例如，当防火墙阻断特定端口连接时，自动生成绕过该端口的变异攻击流量。在威胁响应层面，设计基于策略引擎的自动配置模块，根据攻击类型动态调整入侵防御系统（IPS）规则集，完成从检测到响应的全流程自动化。

五、工程化实践与案例分析

（一）系统实现与部署

基于微服务架构构建测试平台，核心组件采用容器化部署方案。攻击生成模块与测试执行模块通过RESTful API解耦，前者封装GAN模型推理服务，后者集成Scapy流量重放引擎。资源调度层采用Kubernetes编排框架，根据测试负载动态调整生成器与判别器的副本数量。为解决虚拟化环境性能瓶颈，开发基于DPDK的高性能网卡驱动，将攻击流量转发效率提升至线速水平的85%。在数据持久化方面，采用分布式时序数据库存储测试元数据，结合MinIO对象存储服务管理生成的攻击图像与日志文件。部署脚本通过Ansible自动化工具实现，支持单节点验证环境与跨云平台大规模测试集群的快速构建。

（二）典型测试案例

在钓鱼邮件防御测试中，生成器产出包含动态跟踪像素的仿真邮件，成功绕过某开源邮件网关的静态特征检测。防御测试显示，传统规则引擎的检测率从72%降至19%，而集成AI检测模型的商用设备仍保持91%的拦截率。在恶意网页篡改场景测试中，动态特征注入技术生成包含零日漏洞利用代码的仿真页面，导致某UTM设备在300秒内CPU占用率持续超过90%。进一步分析发现，该设备未对CSS表达式执行进行沙箱隔离。针对工业控制系统，构建包含Modbus协议混淆载荷的测试用例，验证某工业防火墙对非常规端口通信的监测盲区，推动厂商发布专项规则补丁。

（三）性能优化与扩展性

开发混合精度推理引擎，将GAN模型的浮点运算转换为INT8量化计算，在保持生成质量的前提下，单张NVIDIA T4显卡的吞吐量提升。引入分布式训练框架，通过参数服务器架构实现多GPU卡间的梯度同步，将百万级攻击图像的训练周期缩短。针对多节点并行测试需求，设计基于消息队列的负载均衡机制，利用Kafka实现测试任务与结果数据的异步解耦。在扩展性验证实验中，系统成功支撑1000并发测试会话，生成流量峰值达，未出现数据包乱序或服务中断。资源监控模块显示，CPU预留策略使关键服务在95%负载下仍保持响应时间稳定。

参考文献

[1]雷刚，王和，李少朋，罗炜.考虑路径点约束的攻击时间与攻击角度控制制导方法[J].中国惯性技术学报，2025，33（2）：172-178.

[2]章海亮，王宇，胡梅，张译之，张晶，詹白勺，刘雪梅，罗微.基于Sigmoid函数归一化的多光谱指数特征在油茶林遥感识别和样本迁移研究中的应用[J].光谱学与光谱分析，2025，45（4）：1159-1167.

[3]罗诗语，李倩.为什么威力无穷的导弹会起名叫“响尾蛇”?[J].学与玩，2025（3）：34-35.

[4]楼琦凯，陈蓓，丁勐，魏莹.基于自适应滑模策略的直流微电网安全控制[J].控制工程，2025，32（3）：416-424.

[5]孙旭，张文琼，龙显忠，李云.基于分层结构的多任务对抗样本归因方法[J].网络与信息安全学报，2025，11（1）：92-105.

[6]颜丽蓉，赵泽荣.基于改进BiLSTM的ADS-B信号欺骗检测方法研究[J].计算机测量与控制，2025，33（3）：54-62.

[7]夏晨旭，郝群，张一鸣，张韶辉，李凡飞，杨智慧，孙建坤.基于结构光投影三维重建的人脸特征检测[J].激光与光电子学进展，2023，60（22）：178-184.

[8]许喆，王志宏，单存宇，孙亚茹，杨莹.基于重构误差的无监督人脸伪造视频检测[J].计算机应用，2023，43（5）：1571-1577.