NFT 仿冒欺诈的测量与检测技术

廖 鹏[1] 方滨兴[1),2)] 刘潮歌[3)] 王 志[3)] 张云涛[1)] 崔 翔[3)]

[1)](北京邮电大学可信分布式计算与服务教育部重点实验室 北京 100876)

[2)](广州大学网络空间安全学院 广州 510006)

[3)](中关村实验室 北京 100094)

摘 要 近年来非同质化代币(Non-Fungible Token,NFT)繁荣发展,但安全问题也日益凸显,尤其是 NFT 的仿冒问题.在去中心化的环境下,仿冒已有的 NFT 作品变得相对容易,而辨别真伪却尤其困难.本文围绕仿冒 NFT 的测量与仿冒检测方法的评估进行了系统深入的研究.建立了包括形式化定义、仿冒过程和仿冒特征在内的 NFT 仿冒威胁模型,给出了 NFT 仿冒定义,分析了 NFT 仿冒方式,给出了判定仿冒 NFT 的一般性方法.大规模采集了全球最大的 NFT 交易平台 OpenSea 上 50000 个 NFT 项目的智能合约地址和历史交易数据,并从以太坊区块链上采集了这些 NFT 项目的名称、创建时间、元数据以及链下存储的 NFT 图像数字载体,从中选取 668 个交易量排名靠前的 NFT 项目围绕 NFT 仿冒问题开展了测量工作,结果表明其中 95 个项目被仿冒 248 次,交易金额超过 2600 万美元,足见 NFT 生态所面临的仿冒欺诈问题之严重.本文采用了 22 种图像数据增强方法,构造了 5000 个扰动较小的攻击测试样本数据集,评估了 OpenSea 和知名的第三方商业检测平台 Fnftf 对仿冒 NFT 检测的鲁棒性,测试结果表明有 6 种图像数据增强方法构造的攻击测试样本能够轻易绕过检测,揭示了 NFT 行业仿冒欺诈检测产品的脆弱性.为提高对仿冒 NFT 检测的鲁棒性,本文提出并实现了一种基于深度学习的 NFT 图像仿冒检测模型,实验表明其 AUC 值相较于 Fnftf 提升了 15.9%.

关键词 非同质化代币;区块链;以太坊;深度学习;对抗攻击

中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2024.01065

Measurement and Detection Techniques for NFT Counterfeiting Fraud

LIAO Peng[1)] FANG Bin-Xing[1),2)] LIU Chao-Ge[3)] WANG Zhi[3)] ZHANG Yun-Tao[1)] CUI Xiang[3)]

[1)](*Key Laboratory of Trustworthy Distributed Computing and Service (Beijing University of Posts and Telecommunications), Ministry of Education, Beijing* 100876)

[2)](*School of Cyber Science and Technology, Guangzhou University, Guangzhou* 510006)

[3)](*Zhongguancun Laboratory, Beijing* 100094)

Abstract With the boom in non-fungible tokens (NFT), security issues are becoming increasingly prominent, one of which is the phenomenon of NFT counterfeiting. In the decentralized environment, it becomes relatively easy to counterfeit other creators' NFT artworks, while it is very difficult to identify the authenticity, and identifying fakes and counterfeits requires a high level of blockchain technology background, The most prominent security problem of the current NFT is the threat of legitimacy due to the lack of regulation of NFT projects, and the NFT security problems of infringement, counterfeiting, plagiarism are difficult to solve only by the blockchain's own security mechanisms. In this paper, a systematic and in-depth research is conducted on measuring

收稿日期:2023-06-30;在线发布日期:2024-02-02.**廖 鹏**,博士研究生,主要研究方向为 Web3 安全、恶意代码.E-mail:liaopeng@bupt.edu.cn.**方滨兴**,博士,教授,中国工程院院士,主要研究领域为计算机体系结构、计算机网络、信息安全.**刘潮歌**(通信作者),博士,副研究员,主要研究方向为 Web 安全、僵尸网络.E-mail:liucg@zgclab.edu.cn. **王 志**,博士,主要研究方向为僵尸网络、AI 安全.**张云涛**,博士,主要研究方向为二进制安全、AI 安全.**崔 翔**,博士,教授,主要研究领域为区块链、网络安全.

NFT counterfeiting and evaluating counterfeiting detection methods. A NFT counterfeiting threat model including the formal definition, counterfeiting process and counterfeiting features are established, NFT counterfeiting definition is given, NFT counterfeiting methods are analyzed, and a general NFT counterfeiting detection method is given. The smart contract addresses and historical trading data of 50 000 NFT projects on OpenSea, the largest NFT marketplace, were collected on a large scale, and the names, creation times, metadata, and digital payload of NFT images stored off-chain were collected from the Ethereum blockchain, from which 668 top NFT projects in terms of trading volume were selected to conduct measurements around the problem of NFT counterfeiting, and the results showed that 95 of them had been counterfeited 248 times, with the trading volume of more than $26 million, which is a clear indication of the seriousness of the counterfeiting and fraud problem facing the NFT ecosystem. Twenty-two image data augmentation methods were used to construct 5000 less perturbed attack test sample datasets, and the OpenSea online real-time counterfeit fraud detection system and the well-known third-party commercial detection platform Fnftf were tested for robustness through the black-box testing method. The test samples generated using the data augmentation method, the smaller the perturbation added, the better, but it is expected that the added perturbation can bypass the image similarity detection. Therefore, on the question of how to evaluate whether the test samples can deceive the visual judgment of NFT buyers, this paper evaluates the effectiveness of the test samples through Perceptual Hash Algorithm. The results of the tests showed that test samples of attacks constructed with six image data augmentation methods were able to easily bypass their counterfeit fraud detection system, and revealing the vulnerability of counterfeit fraud detection products in the NFT industry. In this paper, we address the problems of existing commercialized detection platforms in NFT counterfeit detection, such as poor robustness and vulnerability to bypass, a NFT image counterfeiting detection model is proposed and implemented, and the performance of the model is comparatively evaluated, and the AUC is improved by 15.9% compared to Fnftf, and then the inherent defects of image-based NFT counterfeit detection methods are discussed. Finally, the existing related literature has been thoroughly investigated to demonstrate that the research in this paper is a further refinement of the existing NFT fraud research results.

Keywords non-fungible token; blockchain; Ethereum; deep learning; adversarial attack

1 引 言

非同质化代币(Non-Fungible Token,NFT)[1]是基于区块链技术构建的数字权益证明,具有不可复制和不可篡改的特性. NFT 的核心价值是对图像、声音或文字等数字内容资产化,能够将独一无二的数字物品通过区块链持久存储于数字空间,并可与个人数字身份绑定,从而实现去中心化环境下的数字身份标识、数字资产确权和权益证明.

2023 年之前,基于以太坊(Ethereum,ETH)区块链构建的 NFT 处于绝对主导地位,虽然基于 Solana① 和币安智能链②等其他区块链的 NFT 的发展也很迅速,但只是对以太坊 NFT 的简单模仿. 2023 年,出现了基于比特币(Bitcoin,BTC)区块链的 NFT,先后发布了基于 Ordinals(序数)协议③的比特蛙(Bitcoin Frogs)NFT 项目和基于 Atomicals(原子)协议④的鳄鱼(AtomToothy)NFT 项目,其中代表着纯比特

① Solana. https://solana.com/
② Binance. https://bscscan.com/
③ Ordinals. https://docs.ordinals.com/
④ Atomicals. https://docs.atomicals.xyz/

币链上、真正不可篡改的 AtomToothy NFT 受到了社区的广泛关注，业界有声音认为比特币 NFT 可能会对以太坊 NFT 一直以来的主导地位构成挑战. 以太坊 NFT 存在时间长、生态系统健全、仿冒问题突出，是本文研究的重点，因此如无特别说明，本文后续部分所提及的 NFT 均指以太坊 NFT.

以太坊 NFT 作为数字领域的新兴概念，展现发展势头惊人，不仅依托底层公链、智能合约、数字钱包和存储协议等基础设施，建立了完整的生态系统，还通过引入代币经济学为 NFT 的艺术创作和知识产权保护提供了激励机制，为 NFT 交易提供了支付手段，促进了 NFT 流通[2]. 以太坊联合创始人 Buterin 等人[3]发布的“灵魂绑定通证”是一种不可转让的 NFT，理论上可作为连接数字空间和物理空间的桥梁. 随之而来的是以太坊 NFT 市场交易量所呈现出的爆发式增长，包括数字头像、音乐、体育、游戏道具等在内的各类 NFT 作品，吸引了大量 NFT 投资者. 据统计，2022 年全年 NFT 总交易量达 555 亿美元，环比增长 175%[4]. 其中，由 Larva Labs 于 2017 年创建的 CryptoPunks 是具有代表性的 NFT 项目，其灵感来自伦敦朋克场景、赛博朋克运动和电子音乐艺术. 截至 2023 年 6 月 1 日，CryptoPunks 的累计交易量已达到 25.8 亿美元，其中最贵的单品 CryptoPunk ＃5822 于 2022 年 2 月以大约 2370 万美元的价格售出.

NFT 的繁荣发展也带来了一系列的安全问题，最突出的就是由于 NFT 项目缺乏监管而导致的合法性威胁[1]，其主要表现是因故意仿冒 NFT 作品而导致的欺诈，这使得 NFT 投资者遭受巨大损失. NFT 仿冒欺诈问题泛滥的原因较多：第一，仿冒 NFT 作品成本很低. 在去中心化环境中，仿冒者只需要复制他人的艺术作品或在复制基础上稍加改动，并重新发布在区块链上即可，相比于重新创作 NFT 省去了大量的原创成本. 第二，投资者辨别真品仿品困难. 缺乏区块链基础知识的普通 NFT 投资者根本无法完成诸如 NFT 智能合约地址甄别、智能合约代码审计等一系列技术操作，这导致仿冒 NFT 具有广阔的市场. 第三，几乎无法对仿品追责. 即使创作者和投资者发现了仿冒 NFT，也没有技术办法阻止仿冒者继续销售其作品，更糟糕的是，仿冒者可以通过区块链上的清洗操作，使非法所得代币的追踪变得更加困难. 因此，无论创作者还是投资者只能默默承受仿品带来的损失. 一则典型的 NFT 仿冒获利案例如下：Ripps 团队曾仿冒了著名的 BAYC 项目①，推出了 RR/BAYC②，BAYC 创始团队 Yuga Labs 于 2022 年 6 月提起诉讼，指控 Ripps 及其合作者侵犯版权. 但直到 2023 年 4 月，美国加州法院才裁定 RR/BAYC 为盗版，侵犯了 Yuga Labs 的版权. 这个裁决过程长达 10 个月，在此期间 RR/BAYC 已经获得了巨额利润.

学术界和工业界都已经意识到了 NFT 仿冒问题的严重性，并且开始尝试从技术角度解决这一问题，但对此问题的研究都处于起步阶段. 目前，学术界对 NFT 生态安全问题的研究并不多. 其中，Das 等人[1]的工作最具代表性，不仅归纳和揭示了包括 NFT 仿冒在内的 NFT 生态所面临的安全问题(NFT 交易市场的安全风险、外部实体的安全问题和用户恶意行为，NFT 仿冒问题被归类为用户恶意行为)，还从 NFT 藏品名称和 NFT 图像两个维度对当前 NFT 仿冒情况进行了测量. 在 NFT 藏品名称仿冒测量中，使用度量字符串相似性的算法 Levenshtein[5]对 52399 个 NFT 藏品名称进行评估，结果表明有 322 对集合具有相似名称，大多数仿冒的名称只是在原始名称基础上进行微小修改，例如使用复数形式或在难以察觉的位置添加空格等. 在 NFT 图像仿冒测量中，下载了 OpenSea 生成和缓存的分辨率较低的 9991013 张 NFT 图像，通过图像感知哈希算法[6]发现了不同 NFT 集合间存在 59425 对哈希碰撞. 随后随机选择了其中的 100 对，手工验证后发现 90%的图像在视觉上存在相似性. Das 等人首次系统全面揭示了 NFT 仿冒问题的严重性，但其对于 NFT 仿冒问题的研究主要集中在测量仿冒项目的数量上，并没有从交易规模等更多有意义的角度进行更加系统深入的研究，也没有讨论现有解决仿冒问题的方法的有效性等问题.

在工业界，一些 Web3 领域的初创公司也开始关注 NFT 仿冒问题，并提供了相关的检测服务，例如 Tovera③、Yakoa④、Optic⑤、CheckNFT⑥和 Doppel⑦等公司，表 1 列出了这些公司推出的 NFT 保护产品的名称及其融资情况. 这些公司检测仿冒 NFT 的方法通常是人工智能技术，并提供在线检测服务. 出于商业保密或技术保护的考虑，上述公司都没有公开披露其具体的检测原理和实现细节.

① Yuga Labs. http://www.boredapeyachtclub.com/
② RR/BAYC. https://rrbayc.com/
③ Tovera. https://fnftf.io/
④ Yakoa. https://www.yakoa.io/
⑤ Optic. https://www.optic.xyz/
⑥ CheckNFT. https://checknft.io/
⑦ Doppel. https://www.doppel.com/

表 1 工业界 NFT 仿冒检测公司情况

公司	成立时间	主要产品	融资
Tovera	2021	Fnftf NFT 仿冒检测	未披露
Yakoa	2022	区块链知识产权保护	$4.8M
Optic	2020	NFT 欺诈保护	$11M
CheckNFT	2021	Web3 与元宇宙知识产权保护、NFT 数据检索	未披露
Doppel	2022	数字资产品牌保护平台	$5M

NFT 交易平台首当其冲地意识到了仿冒检测对于 NFT 交易的重要性. 作为一个拥有大量用户基础的 NFT 交易平台,OpenSea① 于 2022 年 11 月 1 日上线了实时仿冒欺诈检测系统. 在此之前,OpenSea 主要依赖用户及社区的人工举报来应对 NFT 仿冒欺诈问题. OpenSea 采用图像识别技术,对用户上传的图像和新铸造的 NFT 进行仿冒检测,如果新铸造的 NFT 被判定为仿冒欺诈,OpenSea 会将其下架. 同时,OpenSea 还不断扩大匹配数据集,并通过持续的模型训练提升检测效果[7-9]. 此外,OpenSea 还建立了专门的人工审核团队,负责审查删除情况. 目前 OpenSea 等 NFT 交易平台和一些 Web3 的初创公司发布的 NFT 仿冒检测系统的鲁棒性尚未有详细的描述资料公开,并且安全性也没有得到充分的验证和测试.

鉴于此,本文围绕 NFT 仿冒问题进行了全面系统化的研究,通过大范围的测量进一步揭示 NFT 仿冒问题的严重性,通过样本对抗测试评估了仿冒检测系统的脆弱性,旨在提升 NFT 的安全,保护 NFT 创作者和投资者的权益,推动 NFT 生态系统的健康可持续发展. 本文的主要贡献如下:

(1)深入剖析了 NFT 仿冒问题. 建立了包括形式化定义、仿冒过程和仿冒特征在内的威胁模型,给出了 NFT 仿冒定义,分析的了 NFT 仿冒方式.

(2)通过测量了提示了以太坊生态中 NFT 仿冒问题的严重性. 从多个维度采集所需数据,包括 NFT 交易量、NFT 图像和相似 NFT 图像等数据,并使用基于数据分析的方法测量了以太坊生态中 NFT 仿冒的交易规模,取得了对该问题的全面且系统的实证研究成果.

(3)评估并验证了当前主流 NFT 仿冒检测系统的脆弱性. 通过使用图像数据增强的方法生成扰动较小的攻击测试样本,成功绕过 OpenSea 和第三方商业 NFT 仿冒检测系统的安全机制,揭示了其脆弱性.

(4)提出了一种更具鲁棒性的 NFT 图像仿冒检测模型. 该模型能够成功检测出绕过主流 NFT 仿冒检测系统的攻击样本,验证了其在对抗 NFT 仿冒方面的效果和可行性.

2 背 景

区块链是一种去中心化的分布式记账技术,可用于安全记录和验证电子投票的交易[10]. 智能合约是以太坊生态中 NFT 项目的核心,用于定义 NFT 的属性、所有权和交易行为接口,当前使用最广泛的智能合约标准是 ERC-721[11] 和 ERC-1155[12]. 通常 NFT 的创作者根据以太坊 NFT 智能合约标准协议来创建和发布自己的 NFT,并生成相应的 NFT 智能合约地址. NFT 智能合约是 NFT 代币的唯一标识,可以作为区分不同 NFT 项目的标识.

元数据是 NFT 的重要组成部分,它关联了 NFT 资产信息,包括 NFT 的名称、描述等,更重要的是它还保存了 NFT 数字资产的存储位置. 通过调用智能合约接口 *TokenURI* 并传入 *TokenID* 参数,可以读取特定 ID 的 NFT 元数据信息,从而可以获取到 NFT 数字资产的存储位置.

铸造是 NFT 的核心业务,通过铸造可以将数字资产代币化,即将图像、音频或者视频数字资产通过元数据关联到区块链上,并分配唯一 *TokenID*. 目前,智能合约协议中并没有关于 NFT 铸造的标准接口,而是由 NFT 发行方自定义编写.

BAYC是一个在以太坊上部署的NFT,其智能合约地址是0xbc4ca0eda7647a8ab7c2061c2e118a18a936f13d. 通过 EtherScan 浏览器可以查看其智能合约代码、智能合约的调用方法等. 以 BAYC 的 TokenID 为 3738 的 NFT 为例,通过调用 TokenURI 接口并传入 TokenID 为 3738 的参数,可以读取该 NFT 元数据的地址信息(ipfs://Qme***Wtq/3738). 如图 1 所示,该地址存储了以 JSON 格式描述的 NFT 元数据信息,其中 image 字段关联了 NFT 的图像信息,attributes 字段描述了该 NFT 的属性信息,例如衣服款式、颜色和面部表情等.

3 NFT 仿冒问题

3.1 NFT 发布流程

参与 NFT 项目的角色主要是 NFT 创作者和 NFT 投资者,另有 NFT 存储服务、NFT 交易市场和区块链三个基础服务设施. 其中,NFT 创作者为 NFT 项目的实际拥有者,负责内容创作与生产;NFT 投资者为购买了 NFT 的用户,即 NFT 产业链下游的消费者;NFT 存储服务主要用于存储 NFT

① OpenSea. https://opensea.io/

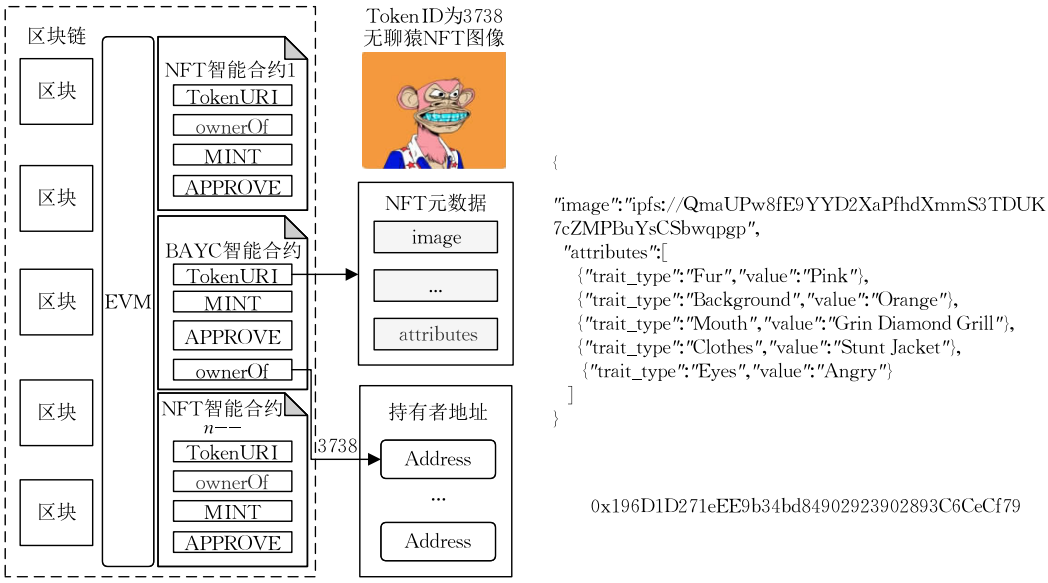

图 1　NFT 存在形态与底层构造

元数据和 NFT 数字内容，例如亚马逊云、阿里云和微软云分别推出了中心化文件存储服务和星际文件系统(Inter Planetary File System，IPFS)①、织网文件系统(Arweave，AR)②等分布式文件存储服务；NFT 交易市场为 NFT 创作者和投资者提供 NFT 发布和交易的平台，目前知名的交易市场包括 OpenSea、Blur③和 LooksRare④等，所交易的 NFT 数字物品的形态包括但不限于图像、动态图像(Graphics Interchange Format，GIF)、声音和视频，其中图像是最常见的形态. 本文的研究就聚焦于 NFT 图像.

图 2 展示了发布 NFT 项目的一般流程：(1) NFT 创作者将创作完成的数字图像上传到互联网的存储服务，并获得这些数字图像的统一资源标识符(Uniform Resource Identifie，URI)；(2) NFT 创作者在区块链上部署智能合约代码，并获得 NFT 智能合约地址，同时通过 NFT 铸造将数字图像代币化，将 NFT 元数据对应的 URI 写入区块链；(3) NFT 创作者通过铸造获得 NFT，随后将 NFT 发布到 NFT 交易市场上进行售卖，且 NFT 交易市场会获得 NFT 项目对应的智能合约地址；(4) NFT 交易市场通过 NFT 项目的智能合约地址，在区块链上获取 NFT 基础信息，例如 NFT 代币名称、发行总量、NFT 元数据 URI 等；(5) NFT 交易市场通过 URI 下载数字图像并缓存，以供 NFT 购买者浏览.

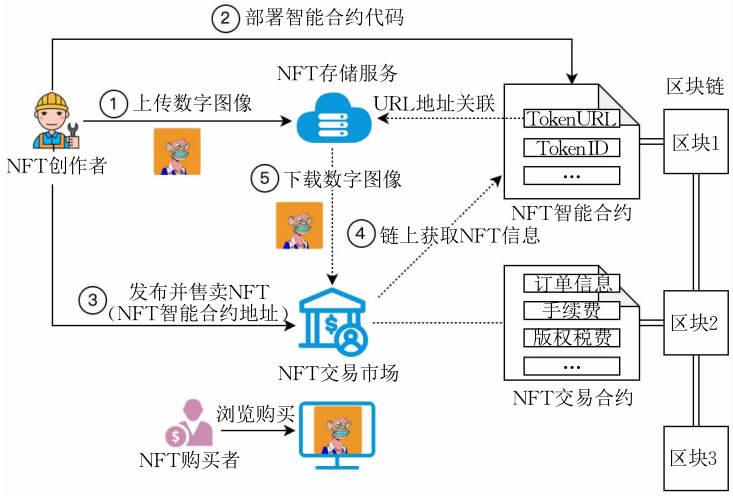

图 2　NFT 项目发布流程

① IPFS. https://ipfs.tech/
② Arweave. https://www.arweave.org
③ Blur: Fastest NFT Marketplace for Pro Traders. https://blur.io/
④ LooksRare. https://looksrare.org/

3.2 威胁模型

3.2.1 形式化定义

定义 1. NFT 由四元组构成，反映了 NFT 的唯一性、可验证性、可操作性和可追溯性等四个基本特性，其形式化描述如下：

$$NFT=\{UID,Obj,\lambda,Txs\},$$

其中

$$UID=\{ChainID,Address,TokenID\},$$

$$Obj=\{Name,Metadate,Payload\},$$

$$\lambda=\{mint,approval,transfer,burn,\cdots\},$$

$$Txs=\{tx_1,tx_2,\cdots,tx_n\}.$$

UID 反映了 NFT 的唯一性. 其中，*ChainID* 代表区块链的 ID，以区分 NFT 部署的区块链，例如比特币、以太坊和 Solana 等区块链；*Address* 代表 NFT 在区块链上的代币地址，在以太坊上用部署智能合约所生成的地址来标识 NFT 代币地址；*TokenID* 代表 NFT 唯一编号，用于区分同一 NFT 代币地址下不同的 NFT 数字载体.

Obj 用于描述 NFT 的数字对象本体. 其中，NFT 数字对象内容可在区块链上被验证，并且不能被篡改，反映 NFT 的可验证性；*Name* 代表 NFT 的名称，通常 NFT 交易市场采用该名称对 NFT 藏品进行命名；*Metadata* 代表 NFT 的元数据，描述 NFT 属性信息以及 NFT 数字载体的 URI；*Payload* 代表 NFT 的数字内容，也即直观的外在表象.

λ 用于描述 NFT 在区块链上的活动事件，反映 NFT 的可操作性. *mint* 动作用于铸造创建 NFT，*approval* 动作将 NFT 的控制权授予给其他地址，*transfer* 动作转移 NFT 所有权，*burn* 动作用于销毁 NFT.

Txs 代表 NFT 在区块链上的完整活动事件的交易记录，反映 NFT 的可追溯性.

定义 2. NFT 仿冒威胁模型（NFT Counterfeiting Threat Model，NCTM）由三元组构成，分别为 NFT、仿冒函数 $\mathcal{F}$ 和检测分类器 θ，

$$NCTM=\{NFT,\mathcal{F},\theta\}.$$

通常情况下，任意两个 NFT 资产的分类结果均不同，有

$$\theta(NFT_i)\neq\theta(NFT_j).$$

但攻击者在仿冒时，根据原有 NFT_A 生成新的 NFT_B，且二者在特征不同的前提下具有相同的视觉效果，这一过程可表示为

$$\mathcal{F}(NFT_A)\rightarrow NFT_B$$

且

$$\theta_{\text{vision}}(NFT_B)=\theta_{\text{vision}}(NFT_A),$$

$$\theta_{\text{detector}}(NFT_B)\neq\theta_{\text{detector}}(NFT_A),$$

其中，仿冒函数 $\mathcal{F}$ 反映攻击者的特征异化能力，攻击者通过修改仿冒 NFT 的部分特征，使其和原始 NFT 的差异足够大，并绕过检测引擎 θ_{detector} 的相似性判断. 检测分类器 θ_{vision} 反映攻击者的视觉误导能力，攻击者通过减小仿冒 NFT 和原始 NFT 之间的视觉差异误导投资者做出错误判断.

3.2.2 NFT 仿冒过程

本文用以太坊生态中的图像 NFT 仿冒来论述这一现象，其对应的仿冒过程如图 3 所示. 由于 NFT

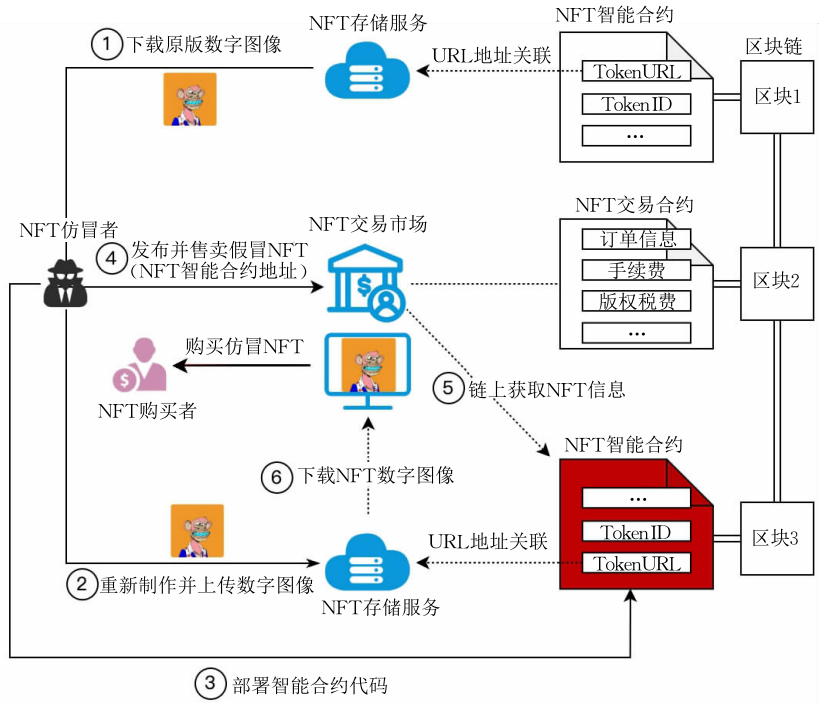

图 3 NFT 仿冒过程

具有可验证性，当一个项目 NFT_A 完成发布后，NFT 仿冒攻击者可以从区块链上获取 NFT_A 的数字对象内容，也即 Obj 中 NFT_A 的名称、元数据和数字图像等信息. 因此，攻击者首先下载原版数字图像，通过仿冒函数 $\mathcal{F}$ 构建一份特征异化的数字图像，并通过检测分类器 θ 修正数字图像，使得仿冒 NFT 具有一定的视觉误导性，并完成 NFT 数字图像的重新制作上传至存储服务. 随后，通过正常的 NFT 项目发布流程发布仿冒 NFT，并生成新的项目 $NFT_B=\{UID_B, Obj_B, \lambda, Txs\}$. 攻击者再将 NFT_B 授权给 NFT 交易市场，并将 NFT_B 在交易市场进行售卖. 当 NFT 购买者浏览到 NFT_B 时，NFT 购买者有可能被误认为是原版 NFT_A，最终诱导购买者购买仿冒的 NFT_B.

3.2.3　仿冒特征分析

NFT 的外观，即 NFT 图像和 NFT 名称高度相似，是仿冒 NFT 的重要特征. 对 NFT 投资者而言，图像和名称是最直接、最直观的视觉外观，NFT 投资者仅依赖 NFT 交易平台所展示的 NFT 藏品名称和图像便决定交易买卖. 因此图像和名称的仿冒最具误导性，而仿冒 NFT 其它维度信息的收益不大，根本没有必要. 具体而言，只需遵照 NFT 的智能合约标准 ERC721 和 ERC1155 就能完成 NFT 智能合约代码的编写、部署和发布，因此没有必要仿冒智能合约代码；NFT 元数据是 NFT 图像信息的描述，仿冒者按仿冒图像的特征提供相应的描述信息即可，其与图像仿冒绑定；NFT 智能合约地址是由区块链随机生成，是一串无自然语义的数字，不必仿冒也无法仿冒. 2023 年 4 月，在 RR/BAYC 侵犯 Yuga Labs 版权的案件中，美国加州法院公布的判决书中指出[13]，其做出判决的依据之一是 RR/BAYC 在宣传推广中使用了和 Yuga Labs 的 BAYC 几乎完全一样的商标，并且 RR/BAYC 销售的产品和 BAYC NFT 图像完全相同，从视觉的角度来诱导消费者. 由此实际判例也可知，仿冒的重点是 NFT 图像和 NFT 名称而非其它.

NFT 在区块链上部署的时间先后顺序是判断原创与仿品的最重要的依据. 若区块链上存在 NFT_A 和 NFT_B 两个项目，且 NFT_B 仿冒 NFT_A. 从它们的智能合约代码、元数据、智能合约地址、“外观相似度”和“名称相似度”等角度无法判断谁原创与仿品，这往往就需要查看部署上链的时间早晚，一般认为仿冒品 NFT_B 部署到区块链上的时间晚于原创品 NFT_A.

综上，仿冒 NFT 的特征可以概括为两点：一是仿品与原创 NFT 的图像和名称高度相似；二是仿品在区块链上的部署时间晚于原创 NFT. 在检测技术方面，NFT 名称相似度检测的本质是字符串相似度检测，现有技术已非常成熟，在本文的后续论述中不再单独讨论；NFT 图像相似度的检测方面，虽然已有一些成熟的技术，但由于涉及检测干扰等对抗性问题，使得这一问题变得复杂起来，因此该部分的内容是本文工作的重要.

3.3　NFT 仿冒定义

前文给出了仿冒 NFT 的定性特征，本节则具体给出判定仿冒 NFT 的严格定义. 假设某一区块链上或不同区块链之间，存在两款 NFT，分别用 NFT_A 和 NFT_B 表示，NFT_B 仿冒 NFT_A 的判断定义为

$$Similarity(NFT_A, NFT_B) \geq \varphi \times |NFT_A|$$

且　$Similarity(NFT_{NAME\,A}, NFT_{NAME\,B}) \geq \mu$

且　$Timestamp(NFT_A) < Timestamp(NFT_B)$，

其中，NFT_A 表示该 NFT 包含数字图像的集合 $\{fig_1, fig_2, \cdots, fig_n\}$，$|NFT_A|$ 表示集合中数字图像数量，NFT_B 表示该 NFT 包含数字图像的集合 $\{fig_1, fig_2, \cdots, fig_n\}$，$Similarity(NFT_A, NFT_B)$ 表示 NFT_A 和 NFT_B 中相似的图像数量. $NFT_{NAME\,A}$ 表示 NFT_A 的名称字符串，$NFT_{NAME\,B}$ 表示 NFT_B 的名称字符串，$Similarity(NFT_{NAME\,A}, NFT_{NAME\,B})$ 表示两个字符串的相似度，且表示两个字符串相似的阈值 μ 满足 $0<\mu\leq 1$. $Timestamp(NFT_A)$ 和 $Timestamp(NFT_B)$ 分别表示 NFT_A 和 NFT_B 智能合约创建的时间，且表示 NFT_B 仿冒 NFT_A 程度的 φ 满足 $0<\varphi\leq 1$.

上述定义可解释为，某一区块链上或不同区块链之间的 NFT 项目 NFT_A 和 NFT_B 的名称相似，图像集合中存在一定数量的相似图像，并且 NFT_B 智能合约创建时间晚于 NFT_A，则认为 NFT_B 一定程度上仿冒了 NFT_A.

3.4　NFT 仿冒方式

3.4.1　拷贝复制

拷贝复制方式的 NFT 仿冒往往发生在知名度较高的 NFT 项目上. 该类 NFT 对应的仿冒项目首先会在以太坊上重新部署一份智能合约，随后拷贝复制已有知名 NFT 项目数字图像，并将链接指向新复制的数字图像，或者直接复用原 NFT 项目的元数据.

RR/BAYC 是一个广为人知的以拷贝复制方式仿冒 BAYC 的项目. RR/BAYC 的智能合约名字是 Bored Ape Yacht Club，与 BAYC 的智能合约名字 BoredApeYachtClub 相比，只有若干个空格的差异，并且仿品的合约名字看起来更加正式；RR/BAYC

与 BAYC 智能合约的符号都是 BAYC，并且二者图像链接也相同，即 RR/BAYC 项目中所包含的数字图像与 BAYC 完全一致. 几乎没办法从上述差异中分辨出仿品和原创品，只能从进一步分析二者的智能合约创建时间中寻找真相，RR/BAYC 智能合约创建是 2022 年 5 月 13 日，而 BAYC 智能合约的创建时间则是更早的 2021 年 4 月 22 日. 由此可见，RR/BAYC 才是仿品，而且采用了拷贝复制的仿冒方式.

3.4.2 简单变形

简单变形方式的 NFT 仿冒是下载已经存在的 NFT 项目的数字图像，对其执行简单的变换处理后上传到互联网存储服务，并在区块链上重新部署一份智能合约进行发布. 常见的变换处理手段包括但不限于裁剪图像、改变像素大小和左右镜像处理等.

mfher① 是一个采用简单变形方式仿冒火柴人(mfer②)的 NFT 项目，主要依据是 mfher 的智能合约创建于 2021 年 12 月 2 日，而火柴人创建智能合约的时间是 2021 年 11 月 29 日，比 mfher 要早了一年. 本文下载了 mfher 的 10000 份 NFT 图像，并对其进行镜像处理，随后采用经典图像相似度计算方法——差异哈希算法与 mfer 相应的 10000 份 NFT 原创图进行两两计算，最终得出图像的相似度全部为 1.0. 由此可见，mfher 完全将火柴人所包含的图像镜像处理后作为自己的数字图像(如图 4 所示)，是典型的简单变形方式的仿冒.

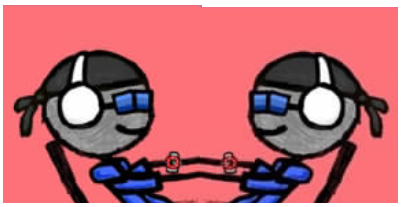

图 4 mfer 和 mfher 的 NFT 示例

4 NFT 仿冒规模测量

本文通过测量方法来分析 NFT 仿冒问题的规模与严重性. 以太坊区块链作为首个引入智能合约的平台，已成为 NFT 交易量最大、生态系统最为完善的网络[14]. 以太坊虚拟机(Ethereum Virtual Machine，EVM)在其区块链上运行，同时也为其他兼容 EVM 的区块链以及二层扩展链所采用. 鉴于此，本文的主要研究焦点为以太坊区块链上的 NFT 仿冒问题.

首先，将采集以太坊区块链上 NFT 项目的数据，分别形成 NFT 项目信息数据集、种子 NFT 图像数据集和相似 NFT 图像数据集. 随后基于此数据集对以太坊上交易量排名靠前的 NFT 项目进行仿冒问题分析.

4.1 NFT 项目信息数据采集

NFT 项目信息包括其智能合约地址、NFT 名称、元数据 URI、交易量、创建时间等静态属性. 首先利用 OpeaSea 提供的开放 API 接口采集交易量排名前 50000 的 NFT 项目智能合约地址及其交易量(截至 2022 年 3 月 27 日)，随后基于智能合约地址进一步利用以太坊区块链浏览器 Etherscan 提供的开放 API 接口采集各 NFT 项目的名称、创建时间、元数据 URI 等信息. 所采集的 NFT 项目信息构成数据集 data_nft，其内容如表 2 所示.

表 2 所采集的 NFT 项目信息

交易量排名	智能合约地址	NFT 名称	交易总量
1	0xbc4 * 13d	BoredApeYachtClub	$2680M
2	0xb47 * bbb	CryptoPunks	$2510M
3	0x60e * 7c6	MutantApeYachtClub	$2020M
…	…	…	…
50000	0x10b * 020	Lion Kingdom	$249.0M

4.2 种子 NFT 图像采集

以交易量对 data_nft 数据集中的 NFT 进行降序排序，分析 NFT 交易量的累计分布情况. NFT 交易量前 10000 的 NFT 项目累计分布如图 5 所示，其呈现出明显的长尾特征. 排名 TOP 1000 的 NFT 项目交易量占整个 TOP 10000 的 NFT 项目交易量的 90.12%，可见大部分交易集中在头部 1000 个 NFT 项目中. 这一实验结果符合常理认知，因为对于欺诈攻击者来说，仿冒交易活跃的头部 NFT 项目最容易使其利益最大化. 后文将选取交易量排名前 1000 的 NFT 项目开展进一步的深入分析.

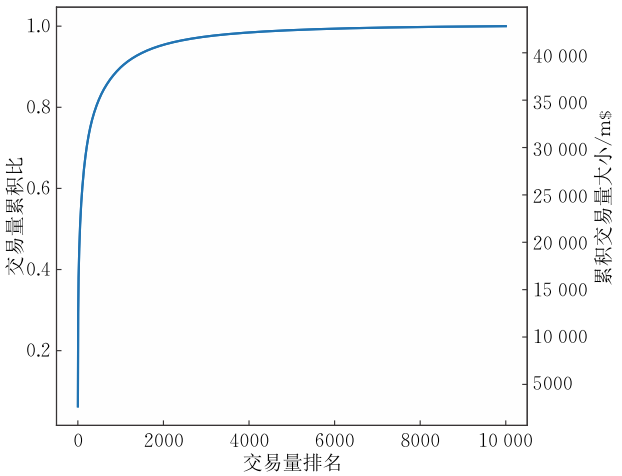

图 5 TOP 10000 的 NFT 项目交易量累积分布

① Mfhers. https://coinstats.app/nft/mfhers/
② mfers. https://mirror.xyz/sartoshi.eth/

考虑到本文的研究重点仅限于 NFT 图像的仿冒问题，因此从 TOP 1000 的 NFT 项目中排除了音频类和视频类的 NFT. 经过筛选，共有 668 个 NFT 项目被选定作为后续研究的种子 NFT 项目.

进一步采集 668 个种子 NFT 项目用于进一步分析，具体采集方法是：利用 Etherscan 的开放 API 接口获取 NFT 图像的 URI，再从基于这些 URI 从存储服务器上逐一下载图像. 实验中，从每 NFT 个项目中随机选取 10 张图像下载，共计下载到 6680 张 NFT 图像，并以这些图像为对象进一步分析这 668 个 NFT 项目的仿冒与被仿冒情况. 值得一提的是，各 NFT 项目的图像存储方式有所不同，有些项目将图像直接存储在 Web 服务器上，有些项目则将图像存储于分布式的 IPFS，具体图像获得方法本文不再赘述. 所采集的 NFT 图像将与其所属 NFT 项目关联并构成数据集 data_nft_seed_picture，其内容如表 3 所示.

表 3　所采集的 NFT 图像信息

ID	智能合约地址	NFT 名称	本地图像地址
1	0xbc4*13d	BoredApeYachtClub	./bayc/1.png
2	0xb47*bbb	CryptoPunks	./punks/1.png
3	0x60e*7c6	MutantApeYachtClub	./mayc/2.png
…	…	…	…

4.3　相似 NFT 图像采集

相似 NFT 图像是指与图像数据集 data_nft_picture 中各图像相似的 NFT 项目图像，即疑似的仿冒项目 NFT 图像. 该数据的采集依赖 Tovera 公司的在线商用系统 Fnftf 中图像相似度匹配算法和 NFT 图像数据库，具体采集方法是：将 data_nft_picture 中的图像作为“种子”，逐一上传到 Fnftf 检测引擎，采集其所返回的相似图像、相似度及其所属的 NFT 项目信息等. Fnftf 提供跨公共区块链的检测服务，实验中共采集到 259157468 张来自以太坊、Polygon 和 Solana 等区块链上的 NFT 图像，但实验所使用的 Fnftf 免费服务仅能检索到 2022 年 12 月 7 日之前的数据. 本部分所采集数据将与 data_nft_seed_picture 数据集关联形成新的相似 NFT 图像数据集 data_nft_similar_picture，其内容如表 4 所示.

表 4　所采集的相似 NFT 图像信息

智能合约地址	NFT 名称	图像相似的智能合约地址	图像相似的 NFT 名称	相似度
0xb47*bbb	CryptoPunks	0xf07*402	CryptoPhunksV2	0.93
0xbc4*13d	BoredApeYachtClub	0x2ee*91e	RR/BAYC	1.00
0x235*68b	Moonbirds	0xdb7*727	Boonmirds	0.82
…	…	…	…	…

4.4　仿冒分析

(1) 仿冒项目数量

当一张图像上传到 Fnftf 检测引擎，若其返回的检测结果显示图像相似度大于 0.7，则认为该图像属于仿冒的 NFT 项目. 本文也同样使用该阈值过滤相似度较低的 NFT 项目. 再根据“智能合约创建时间较晚的 NFT 项目为仿冒项目”这一规则，根据相似 NFT 图像数据集 data_nft_similar_picture 将 NFT 项目仿冒的情况分为两类：种子 NFT 项目内的仿冒情况和种子 NFT 项目之外仿冒种子 NFT 项目情况.

① 种子 NFT 项目内的仿冒情况. 在 668 个种子 NFT 项目中，发现了 3 个 NFT 项目是仿冒项目，涉及交易金额总量超过 1730 万美元.

② 种子 NFT 项目之外仿冒种子 NFT 项目情况. 在 668 个种子 NFT 项目之外，发现了 245 个 NFT 项目是仿冒项目，涉及交易金额总量超过 870 万美元.

综合上述两种情况，可知种子 NFT 项目被仿冒的总体情况：668 个种子 NFT 项目中共有 95 个 NFT 项目被仿冒，仿冒项目共计有 248 个，平均每个项目被仿冒 2.61 个；NFT 仿冒项目的交易总额高达 2600 万美元，也即这些仿冒 NFT 对持有者造成了超过 2600 万美元的直接经济损失. 由此可见，NFT 的仿冒行为将严重损害原创作者的品牌影响力和知识产权，同时也将打击 NFT 持有者对于 NFT 市场的信任和信心，并对 NFT 生态系统的完整性构成了严重威胁.

(2) 仿冒项目分布

图 6 展示了 248 个仿冒项目交易量累计分布特征，其中横坐标代表的是其仿冒的 NFT 项目的交易量排名，纵坐标代表仿冒交易量累积比. 从图中可以看到仿冒的交易量集中在仿冒 TOP300 的 NFT 项目中，具有头部仿冒效应.

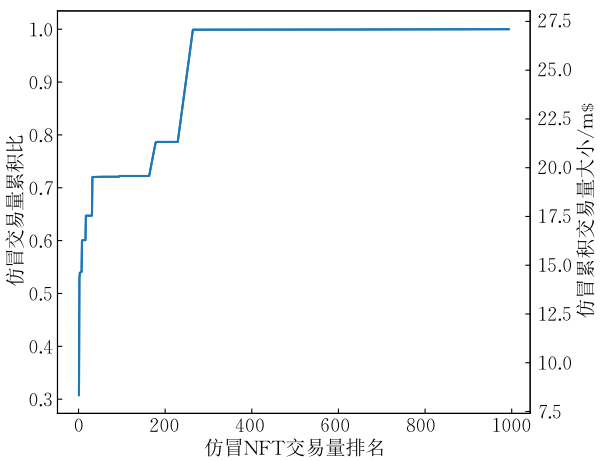

图 6　仿冒 Top1000 的 NFTs 交易量累计分布

图 7 展示了被仿冒 NFT 名称次数的词云图分布特征，从中可以看到 NFT 蓝筹项目[①]最受仿冒者们欢

① Blue Chip NFT. https://decrypt.co/resources/what-is-a-blue-chip-nft

图 7 被仿冒 NFT 名称次数的词云图分布

迎,例如Bored ApeYachtClub、CryptoPunks、Azuki 等.

(3)仿冒案例分析

RR/BAYC 是一个著名的仿冒 BAYC 的 NFT 项目,其交易量在本文测量的仿冒 NFT 交易量中排名第一,约 700 万美元. Yuga Labs 发行的另一个 NFT 项目 MAYC,也检测出一个仿盘 RRMAYC,曾经在 Opensea 上售卖过,产生的交易额约为 33 万美元. 通过调用 OpenSea 的 API 查看 RRMAYC 的状态信息,发现它已经被 OpenSea 下架.

值得一提的是,本文测量出部分仿冒项目,同时也被 OpenSea 判定仿冒. 例如仿冒 CryptoPunks 的 V3 Punks,被 OpenSea 描述为"parody meme"(山寨模仿之意),这也印证了本文测量方法的科学性.

5 NFT 仿冒检测评估

为了评估商业化的 NFT 仿冒检测系统的有效性,本文选择 OpenSea 自带的仿冒检测服务和第三方商业公司的仿冒检测系统 Fnftf 作为研究对象,以 BAYC NFT 项目为例,通过构造图像测试样本进行黑盒测试,评估两者图像仿冒检测的鲁棒性.

5.1 测试方法与样本

采用图像数据增强[15]的方法,生成一批测试样本集,评估两款 NFT 仿冒检测系统的检测能力. 图像数据增强是计算机视觉领域常用方法,通过对图像进行几何变换、颜色变换、旋转变换、缩放变换、翻转变换、裁剪变换和噪音注入等方法生成图像测试样本.

具体方法如下:首先,从 BAYC NFT 原版图像集中随机选择一张编号为 3738 的 NFT 图像,其文件 md5 哈希值为 ee20b3c863d4b4e23baa5def0040d57,像素为 631×631. 其次,使用图像数据增强技术对该图像进行变换,生成图像测试样本集. 针对同一种变换方法,选择适当的参数,尽量对图像的扰动从小到大,以便形成多组对照实验. 使用 22 种数据增强方法生成测试样本集,并根据变换方法的参数对图像扰动的大小分成若干组,共计生成 5000 个测试样本集. 表 5 展示了其中一组共 54 个测试样本,从数据增强方法所属的类别、方法名称、使用参数和生成图像样本编号等方面说明样本详细情况;图 8 则是根据上述 22 种数据增强方法生成的测试样本示例图.

表 5 生成测试样本的数据增强方法

类别	方法名称	参数名称	参数值组	样本编号
几何学	边缘填充	填充大小	1,8,100	1,2,3
	修改像素	像素大小	50,150,300	4,5,6
	中心裁剪	裁剪大小	630,610,300	7,8,9
	透视变换	失真程度	0.01,0.05,0.2	10,11,12
	角度旋转	角度大小	2,5,10	13,14,15
	弹性变换	位移大小	2,100,200	16,17,18
	随机裁剪	裁剪大小	630,620,510	19,20,21
	随机裁剪 修改像素	裁剪比例 与大小	1,0.9,0.4 与 600	22,23,24
	水平翻转	—	—	25
	垂直旋转	—	—	26
颜色	水平翻转灰度	—	—	27
	翻转图像颜色	—	—	28
	直方图均衡化	—	—	29
	图像灰度	—	—	30
	随机色彩抖动	亮度与色调	0.6,0.3	31,32,33
	高斯模糊	模糊分布	0.1,2,3	34,35,36
	色调分离	减少位数	1,2,3	37,38,39
自动增强	AugMix[16]	—	—	40,41,42
	TrivialAug	—	—	43,44,45
擦除	涂成灰度	位置与大小	330 与 5,10,30	46,47,48
	空白擦除	位置与大小		49,50,51
	马赛克	位置与大小		52,53,54

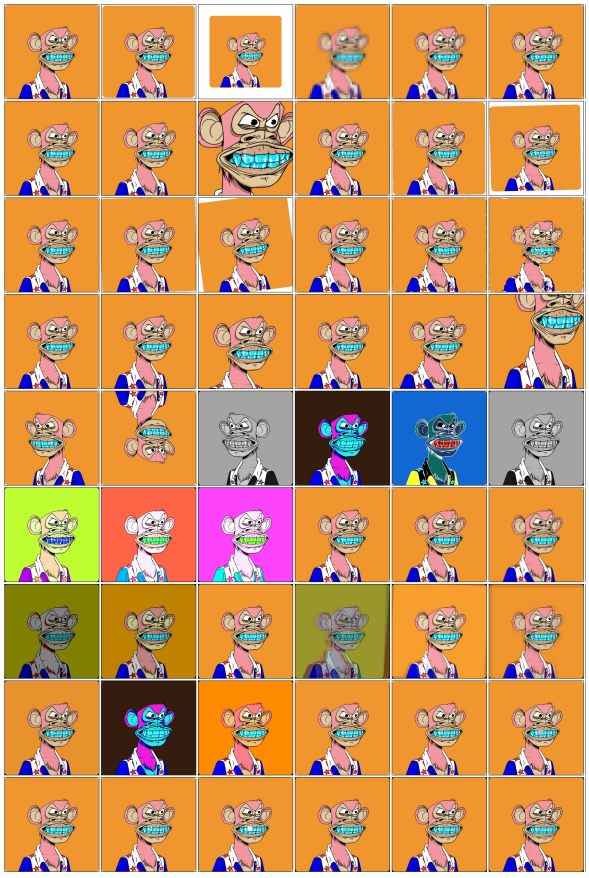

图 8 测试样本示例图组

5.2　对 Fnftf 的黑盒测试

Fnftf 提供了图像相似度值，这有利于进行黑盒测试. 本文先对 Fnftf 进行黑盒测试.

(1) 测试方法. 利用第 4.1 节中所描述的方法，使用 Fnftf 的接口上传测试样本，解析检测结果获取与测试样本相似的 NFT 图像、相似度值和智能合约信息，并记 Fnftf 所返回的相似度值标为 *fnftf_hash*. Fnftf 提供跨公共区块链的仿冒检测服务，当一张图片上传到 Fnftf 时，它将返回全链相似的 NFT 图像、相似度和智能合约信息，也即返回多份相似度数据. 由于本文实验是对 BAYC NFT 进行仿冒检测测试，因此在提取测试样本的相似度 *fnftf_hash* 时，只需要解析提取与原版 BAYC 的智能合约地址(0xbc4ca0eda7647a8ab7c2061c2e118a18a936f13d)相关的相似度值，也即提取测试样本与 BAYC 编号为 3738 的 NFT 原版图像的 *fnftf_hash* 值. 对未被 Fnftf 识别的测试样本 *fnftf_hash* 的值标记为 0.

(2) 有效性评估. 对仿冒攻击者来说，如果构造的 NFT 图像可以欺骗 NFT 购买者的视觉判断，便是一种有效的仿冒. 因此，在构造测试样本时，应尽量满足这一要求. 利用数据增强方法生成的测试样本，加入的扰动越小越好，但又期望加入的扰动能绕过 Fnftf 相似图像的检测. 因此，在如何评估测试样本能否欺骗 NFT 购买者的视觉判断的问题上，本文通过感知哈希算法(Perceptual Hash Algorithm，pHash)[6] 对测试样本进行有效性评估. 感知哈希算法的计算过程分成特征提取和量化压缩两个阶段.

特征提取用于获取图像的整体感知特征，量化压缩用于对感知特征进行压缩编码，从而得到 64 位的哈希摘要，也即图像的“指纹”. 图像之间的相似性可以通过图像的“指纹”之间的汉明距离(表示两个相同长度字符串对应位置的不同字符的数量)进行衡量. 也即，两张图像“指纹”的 64 位哈希值之间不相同的位数. 通常汉明距离小于 10 被判定为两张图像相似[17]，即有大于 54 位相同的哈希值位数，其均值为 54/64 = 0.843 75，映射到 0 到 100 区间为 84.375. 因此，本文选取该值作为评估两张图像相似的阈值. 本文将测试样本的 *fnftf_hash* 的结果评估划分为三类：

(1) 有效仿冒：

$phash \geq 84.375$ 且 $fnftf\_hash = 0.0$；

(2) 无效仿冒：

$phash < 84.375$ 且 $fnftf\_hash = 0.0$；

(3) 有效识别：

$fnftf\_hash > 0.0$.

其中，有效仿冒表示测试样本没有被 Fnftf 识别存在相似图像，但感知哈希算法识别为与原版图像相似，即可以绕过 Fnftf 仿冒检测，但视觉感知为与原版图像相似；无效仿冒表示测试样本没有被 Fnftf 识别存在相似图像，同时感知哈希算法识别为与原版图像不相似；有效识别表示测试样本被 Fnftf 识别存在相似图像.

5.3　对 Fnftf 的结果评估

本文将生成的 5000 个测试样本分成了若干组数据，通过分组测试发现 22 种图像数据增强方法中有 6 种方法可以绕过 Fnftf 仿冒检测系统的检测，以其中一组测试样本组进行说明，其评估结果如表 6 所示.

表 6　评估结果表

样本编号	方法名称	*fnftf_hash*	*phash*	评估结果
1		82.714 843 8	98.4375	有效识别
2	边缘填充	0	95.3125	**有效仿冒**
3		0	57.8125	无效仿冒
4		53.027 343 8	95.3125	有效识别
5	修改像素	90.039 062 5	100	有效识别
6		93.359 375	100	有效识别
7		83.203 125	100	有效识别
8	中心裁剪	0	95.3125	**有效仿冒**
9		0	54.6875	无效仿冒
10		64.550 781 3	95.3125	有效识别
11	透视变换	0	95.3125	**有效仿冒**
12		0	75	无效仿冒
13		83.496 093 8	85.9375	有效识别
14	角度旋转	0	84.375	**有效仿冒**
15		0	75	无效仿冒
16		97.851 562 5	90.625	有效识别
17	弹性变换	51.269 531 3	81.25	有效识别
18		0	75.3125	无效仿冒
19		87.207 031 3	100	有效识别
20	随机裁剪	0	95.3125	**有效仿冒**
21		0	95.3125	**有效仿冒**
22	随机裁剪	96.484 375	100	有效识别
23	修改并像	0	95.3125	**有效仿冒**
24	素大小	0	46.875	无效仿冒
25	水平翻转	100	98.4375	有效识别
26	垂直旋转	0	51.5625	无效仿冒
27	水平翻转灰度	100	75	有效识别
28	翻转图像颜色	0	4.6875	无效仿冒
29	直方图均衡化	0	25	无效仿冒
30	图像灰度	100	75	有效识别
31		69.042 968 8	75	有效识别
32	随机色彩抖动	48.242 187 5	68.75	有效识别
33		45.117 187 5	76.5625	有效识别
34		100	100	有效识别
35	高斯模糊	98.046 875	81.25	有效识别
36		97.265 625	75	有效识别
37		68.359 375	75	有效识别
38	色调分离	81.933 593 8	75	有效识别
39		100	75	有效识别
40		51.855 468 8	75	有效识别
41	AugMix	79.980 468 8	37.5	有效识别
42		86.523 437 5	75	有效识别

（续 表）

样本编号	方法名称	*fnftf_hash*	*phash*	评估结果
43		97.9492188	75	有效识别
44	TrivialAug	0	4.6875	无效仿冒
45		89.0625	75	有效识别
46		99.0234375	100	有效识别
47	涂成灰度	98.2421875	100	有效识别
48		91.015625	100	有效识别
49		95.4101563	100	有效识别
50	空白擦除	90.7226563	100	有效识别
51		83.0078125	100	有效识别
52		95.3125	75	有效识别
53	马赛克	92.4804688	75	有效识别
54		86.6210938	75	有效识别

有效仿冒评估. 该组测试样本中，总共有 7 个样本可以绕过 Fnftf 的仿冒检测，对应的数据增强方法包括边缘填充、中心裁剪、透视变换、角度旋转和随机裁剪等 6 种. 例如，对于透视变换的增强方法，当时失真参数设置为 0.01 时，Fnftf 识别与 BAYC NFT 编号为 3738 的原版图像相似度为 64.66. 当时失真参数设置为 0.05 时，Fnftf 无法从其数据库中匹配到相似的图像，*fnftf_hash* 结果为 0，而用感知哈希算法 *phash* 与 BAYC NFT 编号为 3738 的原版图像相似度为 95.31.

5.4 针对 OpenSea 的黑盒测试

本小节先后使用原版图像和绕过 Fnftf 的仿冒检测的 7 个测试样本，对 OpenSea 的仿冒检测能力进行验证.

2022 年 11 月 OpenSea 推出实时的 NFT 仿冒检测系统，宣称可在几秒钟内识别出完全匹配、翻转和模糊变换的 NFT 数字图像副本（即 OpenSea 同样主要从图像维度进行仿冒检测）[8]. OpenSea 不允许被检测为仿冒数字图像铸造成为 NFT，对于已经铸造的仿冒 NFT 则会被立刻下架[9].

原版图像铸造测试. 向 OpenSea 上传编号 3738 的将 BAYC NFT 原版图像，尝试铸造一个仿冒的 NFT 项目. 虽然 OpenSea 提示铸造成功，但在 3 秒后该仿冒 NFT 就被删除. 本次测试证明 OpenSea 仿冒检测系统工作正常.

测试样本图像铸造测试. 向 OpenSea 上传成功绕过 Fnftf 仿冒检测的 7 个 BAYC NFT 仿冒样本，尝试进行 NFT 铸造. OpenSea 仅拦截了通过透视变换方法生成的测试样本，剩余 6 个测试样本全部成功铸造成 NFT，如图 9 所示.

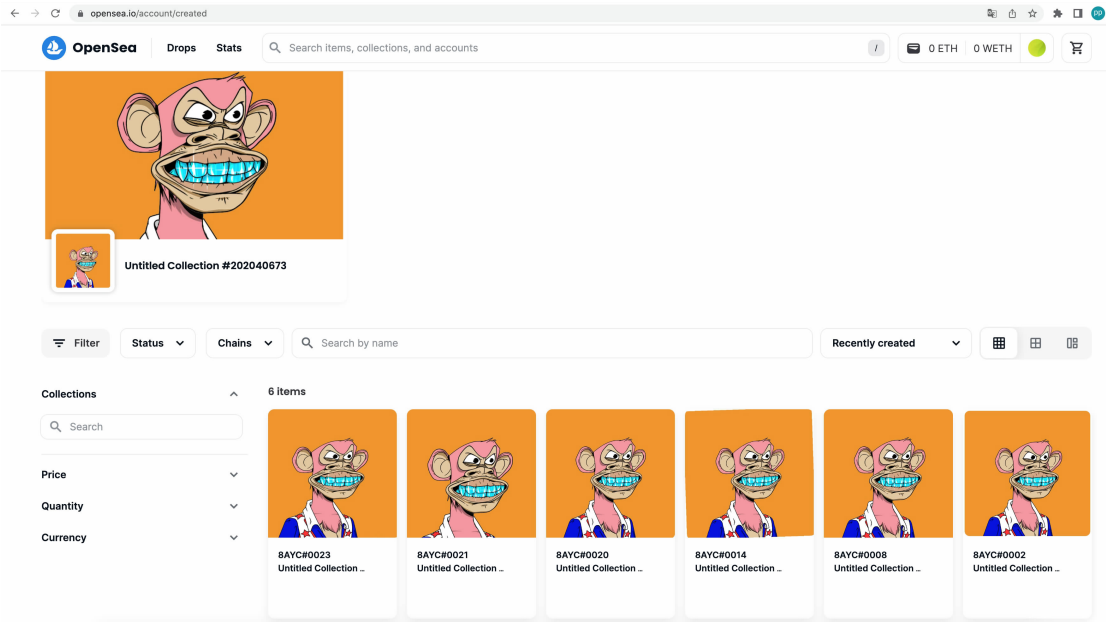

图 9 在 OpenSea 上成功铸造 NFT 的测试样本

另外，测试中还发现 OpenSea 也会对 NFT 藏品名称进行检测，例如不允许用户取名“BAYC”等已经被用户普遍共识的 NFT 名称，只有通过 OpenSea 的 NFT 名称检测后才允许用户铸造 NFT 图像. 因此，测试中还构造了一个仿冒名称“8AYC”完成 NFT 图像上传与铸造.

上述测试表明，虽然 OpenSea 作为行业第一的 NFT 交易平台，成交额最高接近 50 亿美元/月，单月手续费收入超 1.2 亿美元[18]，本文所构造的测试样本仍然可以轻易绕过其仿冒检测系统，足以说明 NFT 行业的仿冒检测还处于起步阶段，还有很大的提升空间.

6　NFT 仿冒检测增强方法

本文在第 5 节论证了现有商业化检测平台在 NFT 仿冒检测方面存在鲁棒性较差和容易被绕过的脆弱性. 本文将提出一种基于深度神经网络的方法来实现 NFT 更具鲁棒性的仿冒检测，本文还将针对基于图像的 NFT 仿冒检测方法的固有缺陷进行讨论.

6.1　检测方法

基于 Lin 等人[19]提出的图像检索思路，利用 ImageNet 数据集预训练一个深度神经网络模型，利用模型的隐藏层特征表示提取图像特征向量，并构建图像特征数据库. 在进行图像匹配时，计算目标图像特征向量与图像特征数据库中的向量之间的距离，作为相似度值.

基于 VGG16[20]深度神经网络，使用 ImageNet 对模型进行预训练，将 10000 张 BAYC 的原版图像作为 VGG16 的输入，提取模型最后一层卷积的特征向量来构建 BAYC 的图像特征数据库. 在对 BAYC 的 NFT 图像进行仿冒检测时，只需要从 VGG16 模型中提取出最后一层卷积的特征向量，随后再与 BAYC 的图像的特征向量进行距离计算，即可得到图像的相似度，本文把训练的 VGG 图像相似度检测模型记为 VGG_SIM.

6.2　效果评估

将上文通过数据增强方法生成的测试样本输入到 VGG_SIM 模型，得到一组测试样本图像与原版图像相似度的值. 本文从有效仿冒维度，评估 VGG_SIM 模型效果，也即测试样本与原图的 $phash \geq 84.375$ 时标记为正样本，测试样本与原图的 $phash < 84.375$ 时标记为负样本，同时与 FNFTF 仿冒检测模型(FNFTF_SIM)进行对比分析.

图 10 展示了两个模型的受试者工作特征曲线

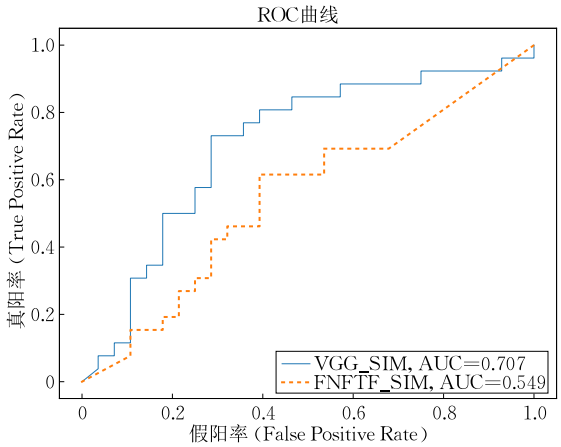

图 10　VGG_SIM 和 FNFTF_SIM 的 ROC 曲线

ROC(Receiver Operating Characteristic Curve)和 ROC曲线下的面积AUC(Area Under the ROC Curve). AUC 可以客观地评价分类器的优劣，AUC 值越大，分类器性能越优[21]. 实验结果表明 VGG_SIM 模型的 AUC 值大于 FNFTF_SIM 的模型的 AUC 值，因此在所构造的攻击测试样本上，VGG_SIM 模型性能优于 FNFTF_SIM，*AUC* 值相较于 FNFTF_SIM 提升了 15.9%.

同时，作为对比参照，在 VGG_SIM 模型上检测了绕过 Fnftf 仿冒检测的 7 个测试样本. 实验结果如图 11 所示，VGG_SIM 模型可全部检出这个 7 个测试样本，且其中最小相似度值为 74.2278. 因此，VGG_SIM 模型在这 7 个测试样本上的性能也优于 FNFTF_SIM 模型.

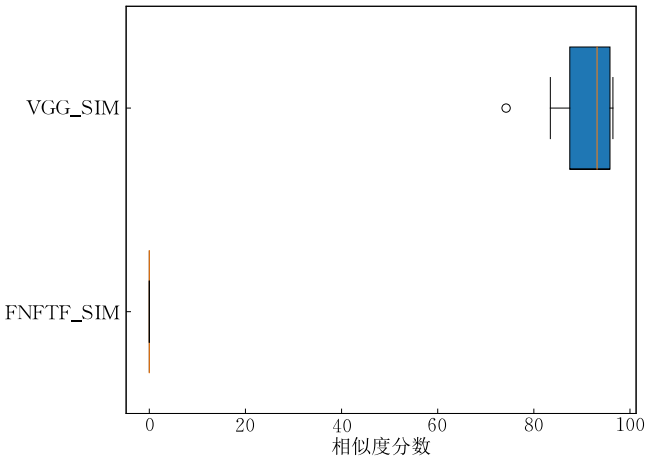

图 11　两个模型相似度分数分布箱型图对比

本文引入主流深度图像识别模型为基础的检测模型，并针对 NFT 仿冒检测场景使用 10 000 张 NFT 原版图像进行模型微调，并且为进一步加强模型的鲁棒性，还采用了数据增强技术. 实验表明，本文所提出的 NFT 仿冒检测模型 VGG_SIM 的表现好于现有的第三方商业 NFT 仿冒检测模型 FNFTF_SIM.

6.3　讨　论

虽然本文所提出的基于深度学习的仿冒 NFT 检测模型能够有效识别那些通过数据增强技术构造的、干扰较小的攻击测试样本，但仍需留意对抗样本攻击[22]的潜在风险. 因此，为了增强模型的鲁棒性，有必要引入一系列防御策略，包括对抗训练[23]、随机性防御[24]等方法，以提升深度学习模型在面对复杂攻击环境时的稳定性和可靠性.

此外，NFT 仿冒检测的方法还可以更加多元化，图像相似度检测只是 NFT 仿冒检测的主要特征之一，并不是全部，例如可以综合考虑链上 NFT 藏品名称的相似性. 同时，区块链外的行为分析也是

重要的一环，如通过分析仿冒者在社交网络平台（如Twitter、Discord 和 Instagram 等）上的行为，来辅助进行仿冒判定.这种多维度的分析方法有助于提高 NFT 仿冒检测的准确性和效率.

同时也必须认识到，所引入的防御策略只能暂时提高对抗样本攻击的难度，在攻防技术不断演进的背景下，对抗样本攻击仍然是一个不可避免的长期挑战.

7 相关工作

随着 NFT 在数字经济领域的地位日益凸显，其面临的安全问题也引起了广泛关注.目前，学术界主要从 NFT 生态安全、NFT 版权保护和 NFT 欺诈研究等角度对 NFT 面临的安全风险展开研究.

在 NFT 生态安全方面，Das 等人[1]首次系统阐述了 NFT 生态系统的运作方式，度量评估了 NFT 生态系统中 NFT 交易市场的计缺陷与安全风险、外部实体所导致的 NFT 脆弱性以及用户的恶意交易行为.该研究将 NFT 仿冒归类为用户的恶意交易行为，并从 NFT 藏品名称和 NFT 图像维度评估了潜在仿冒 NFT 的数量.Gupta 等人[25]对 NFT 安全风险分为三类：第一是所谓的第八层风险，即主要是 OSI 七层模型以外的风险；第二是缺少或失效风险，主要是指黑客攻击或者交易平台错误风险；第三是外部风险，主要是技术体系以外的风险，例如监管和政策风险.Wang 等人[26]重点阐述了 NFT 生态模型，从 NFT 安全的角度出发提出了一种基于 STRIDE 的 NFT 威胁评估模型.Wang 等人[27]对 NFT 数字资产进行了脆弱性分析，从 NFT 的存储、可访问性和可复制性等角度对以太坊网络上的 12353 个 NFT 智能合约进行了测量和评估，并指出：以太坊上 25.24%的 NFT 智能合约不可访问，21.48% 的 NFT 智能合约与重复资产相关.Chen 等人[28]发现当前 NFT 生态系统中存在的严重隐私问题，即每个 NFT 所有者的地址都以明文形式存储在区块链上，这使得攻击者可以轻松获取整个 NFT 资产及其与特定区块链地址相关的价值信息，从而更容易发起针对高价值 NFT 拥有者的特定攻击行为.为此，论文提出了一种基于 OpenSea 市场的新型 NFT 交易方案，以在交易中隐藏 NFT 所有者的地址.上述研究工作主要对 NFT 生态系统的安全问题进行整体概述和分析，但并未对 NFT 仿冒这一问题进行深入研究.

在 NFT 版权保护研究方面，Mochram 等人[29]认为 NFT 作为一种数字资产面临未经授权的副本和不负责任交易的风险，并针对 NFT 和区块链安全生态系统中如何实现 NFT 的产权保护进行研究，遗憾的是，他们的研究结论认为区块链中的 NFT 在版权侵权、数据盗窃和剽窃等方面仍存在无法解决的安全问题，NFT 投资者必须自行规避这些安全问题.García 等人[30]提出了一种基于区块链与语义网结合的 CopyrightLY 解决方案，允许创作者在 NFT 元数据中声明 NFT 的版权，通过使用语义技术对增强的 NFT 元数据进行授权许可，明确 NFT 转移的条件，保持与底层版权的关系，提高了 NFT 的可信度.此外，他们还引入了 CLY 代币和投票机制，以解决虚假声明问题.Pungila 等人[31]提了一种基于非确定性有限自动机的近似模式匹配方法，用于检测剽窃行为，并将该方法用于 NFT 图像剽窃检测，在整个测试过程中吞吐量得到了显著提升，但准确性几乎没有受到任何影响.Prihatno 等人[32]提出了一种基于深度学习的高效 NFT 图像剽窃检测方法，利用 EfficientNet-B0 架构和 Triplet Semi-Hard 损失函数训练了一个深度神经网络模型，并利用损失和准确率等指标对该模型进行了性能评估，结果表明，相较于其他检测模型 Resnet50、DenseNet 和 MobileNetV2，所提出的方法在检测剽窃 NFT 方面表现更好.上述研究工作表明 NFT 侵权、NFT 仿冒剽窃等 NFT 安全问题仅通过区块链自身的安全机制是难以解决的.本文的研究重点是对公共区块链上的 NFT 仿冒欺诈进行测量，以及对现有 NFT 仿冒检测产品的脆弱性进行评估，本文的研究成果进一步揭示了 NFT 仿冒问题的严重性.

在 NFT 欺诈行为研究方面，Song 等人[33]提出了一种基于数据挖掘和机器学习的方法来识别 NFT 市场中的异常交易行为，从 NFT 交易的网络图信息、交易量信息和交易频率等维度提取了 26 个特征，并采用 K-means 聚类算法，将行为相似的钱包地址进行分组，并分析潜在的刷量交易钱包地址，最终得出 NFT 市场中的刷量交易钱包地址占比为 5.38%这一结论.Wen 等人[34]提出了一种用于 NFT 市场中刷量交易行为可视化检测与分析的方法 NFTDisk，并通过两个案例研究和对 14 名真实的 NFT 投资者进行了深入的访谈，以证明 NFT-

Disk 的有效性和可用性. Victor 等人[35]测量了两款主流的以太坊去中心化交易所 IDEX 和 EtherDelta 的刷量交易活动，结果表明超过 30%的代币参与了刷量交易，总金额高达 1.59 亿美元. 他们还从以太坊区块链上选取了 52 种交易量较大的 ERC721 NFT 集合，并获取了 2018 年 1 月 1 日到 2021 年 11 月 21 日的交易数据. 分析结果显示，与非法洗白交易有关的交易占总交易量的 2.04%，涉及的地址数占总地址数的 3.93%，导致交易额增加了 1.495 亿美元. Roy 等人[36]跟踪了 439 个的 Twitter 账号，这些账号通常通过赠品的方式推广欺诈性的 NFT，并涉及 1028 起 NFT 网络钓鱼攻击；研究还表明与这些推广活动互动的账号大多数都是机器人，通过增加点赞、关注和转发数量来迅速提升欺诈 NFT 的流行度. Chan 等人[37]利用半监督学习分析和检测了 NFT 智能合约中的欺诈行为，并将检测结果应用于 NFT 社交平台 DTTD①. 此外，研究还基于 DTTD 的社交数据和分类数据，使用多种统计学习模型评估了检测 NFT 欺诈行为的准确性，并在 LGBM 验证集上达到了 94.38%的最高准确率. Li 等人[38]提出了一种基于时间交易聚合图网络的方法，以提高在以太坊上的钓鱼诈骗检测性能. 他们首先对节点之间的历史交易记录的时序关系进行建模，构建交易特征. 随后将交易特征与图神经网络相结合. 最终显著提升了 TTAGN 在检测钓鱼诈骗地址上的检测效果. 在以太坊网络钓鱼欺诈数据集上，TTAGN 实现了 92.8%的 AUC 和 81.6%的 F1-score. Wu 等人[39]提出了一种通过挖掘交易记录来检测网络钓鱼欺诈的方法，利用标记为网络钓鱼地址的历史交易记录，重建了交易网络，构建了基于交易金额和时间戳的特征，并采用单类支持向量机将节点分为正常地址节点和网络钓鱼地址节点. Kim 等人[40]针对 NFT 钓鱼盗窃问题，提出了基于 NFT 交易行为的盗窃检测系统. 首先，他们从以太坊区块链中提取了 8300 万条 NFT 交易数据和 742 个盗窃者账户，发现盗窃者账户的交易行为和社交行为与普通账户存在显著差异. 随后，使用图神经网络对 NFT 生态系统中这一复杂的关系进行建模. 最终，利用 NFT 持有时间、交易类型、交易价格及用户活跃时间、转入转出比、交易邻接点等特征实现了 NFT 盗窃检测. 由上述工作可知，目前学术界探讨了诸多 NFT 欺诈的不同形式，包括 NFT 刷量交易欺诈、NFT 社交网络欺诈以及 NFT 网络钓鱼欺诈等，但在 NFT 仿冒欺诈领域的研究却相对匮乏. 因此，本文的研究是对现有 NFT 欺诈研究成果的进一步完善.

8　总　结

本文对 NFT 仿冒问题进行了全面而深入的分析. 首先从 OpenSea、以太坊区块链以及第三方商业检测服务平台 Fnftf 采集 NFT 交易数据、资产信息和 NFT 图像数据来测量以太坊生态中 NFT 仿冒的交易规模. 其次利用图像数据增强技术，生成了 5000 个扰动较小的攻击测试样本，用于测试全球最大 NFT 交易平台 OpenSea 和第三方商业检测平台 Fnftf 的 NFT 仿冒检测能力的鲁棒性. 实验表明，本文所构造的攻击测试样本能够成功绕过 OpenSea 和 Fnftf 的仿冒 NFT 检测机制. 最后本文提出并实现了一种 NFT 图像仿冒检测模型，实验结果表明本文提出的模型在 NFT 仿冒检测上的性能优于 Fnftf.

值得注意的是，针对以太坊生态中非图像类 NFT(如 GIF、声音和视频等)的仿冒检测技术，以及新兴的比特币生态中的 NFT 仿冒检测，仍需进一步研究. 这些领域的探索将为 NFT 仿冒检测技术的发展和完善提供更多的可能性和方向.

参　考　文　献

[1] Das D, Bose P, Ruaro N, et al. Understanding security issues in the NFT ecosystem//Proceedings of the ACM Conference on Computer and Communications Security (CCS). Los Angeles, USA, 2022: 667-681

[2] Borri N, Liu Y, Tsyvinski A. The Economics of Non-Fungible Tokens. Available at SSRN 4052045, 2022

[3] Weyl E G, Ohlhaver P, Buterin V. Decentralized society: Finding Web3's Soul. https://ssrn.com/abstract=4105763, 2022

[4] Lian A. 27 stats about NFTs in 2022 — who are the big winners. https://cryptoslate.com/27-stats-about-nfts-in-2022-who-are-the-big-winners, 2022

[5] Li Y, Liu B. A normalized Levenshtein distance metric. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29(6): 1091-1095

[6] Zauner C. Implementation and Benchmarking of Perceptual Image Hash Functions[M. S. dissertation]. Upper Austria University of Applied Sciences, Hagenbe, Austria, 2010

① DTTD. https://www.dttd.io/

[7] Devin Finzer. Our Efforts to Curb Fraud and Plagiarism—and What's Next. https://opensea.io/blog/announcements/our-efforts-to-curb-fraud-and-plagiarism-and-whats-next/, 2023

[8] Anne Fauvre-Willis. Authenticity on OpenSea: Updates to Verification and Copymint Prevention. https://opensea.io/blog/announcements/improving-authenticity-on-opensea-updates-to-verification-and-copymint-prevention/, 2022

[9] OpenSea. Tweet. https://twitter.com/opensea/status/1587194630432497664, 2022

[10] Huang Jun, He Debiao, Obaidat M S, et al. The application of the blockchain technology in voting systems: A review. ACM Computing Surveys, 2021, 54(3): 1-28

[11] Ethereum. ERC-721 non-fungible token standard. https://ethereum.org/en/developers/docs/standards/tokens/erc-721/, 2018

[12] Ethereum. ERC-1155 multi-token standard. https://ethereum.org/en/developers/docs/standards/tokens/erc-1155/, 2021

[13] District Court, C. D. California. Statement of Decision Granting Plaintiff Yuga Labs. https://storage.courtlistener.com/recap/gov.uscourts.cacd.855658/gov.uscourts.cacd.855658.225.0.pdf, 2023

[14] Nansen. NFT Statistics 2023: Sales, Trends, Market Cap and More. https://www.nansen.ai/guides/nft-statistics-2022, 2022

[15] Gandhi A. Data Augmentation | How to use Deep Learning when you have Limited Data — Part 2. https://nanonets.com/blog/data-augmentation-how-to-use-deep-learning-when-you-have-limited-data-part-2/, 2018

[16] Hendrycks D, Mu N, Cubuk E D, et al. AugMix: A simple data processing method to improve robustness and uncertainty //Proceedings of the 8th International Conference on Learning Representations (ICLR 2020). Addis Ababa, Ethiopia, 2020: 1-15

[17] Sarohi H K, Khan F U. Image retrieval using perceptual hashing. IOSR Journal of Computer Engineering, 2013, 9(1): 38-40

[18] Sarah Kearns. OpenSea Sets New Record with $5 Billion USD in Monthly NFT Sales. https://hypebeast.com/2022/2/opensea-new-record-nft-sales-january-2022, 2022

[19] Lin Kevin, Yang Huei-Fang, Hsiao Jen-Hao, Chen Chu-Song. Deep learning of binary hash codes for fast image retrieval//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops 2015). Boston, USA, 2015: 27-35

[20] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition//Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). San Diego, USA, 2015: 1-14

[21] Huang J, Ling C X. Using AUC and accuracy in evaluating learning algorithms. IEEE Transactions on Knowledge and Data Engineering, 2005, 17(3): 299-310

[22] Goodfellow I, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). San Diego, USA, 2015: 1-11

[23] Shaham U, Yamada Y, Negahban S. Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing, 2018, 307: 195-204

[24] Xie Cihang, Wang Yunlong, Zhang Zhishuai, et al. Mitigating adversarial effects through randomization//Proceedings of the 6th International Conference on Learning Representations (ICLR 2018). Vancouver, Canada, 2018: 1-16

[25] Gupta Y, Kumar J, Reifers A. Identifying security risks in NFT platforms. arXiv preprint arXiv:2204.01487, 2022

[26] Wang Qin, Li Rujia, Wang Qi, Chen Shiping. Non-Fungible Token (NFT): Overview, evaluation, opportunities and challenges. arXiv preprint arXiv:2105.07447, 2021

[27] Wang Ziwei, Gao Jiashi, Wei Xuetao. Do NFTs' owners really possess their assets? A first look at the NFT-to-asset connection fragility//Proceedings of the ACM Web Conference 2023 (WWW 2023). Austin, USA, 2023: 2099-2109

[28] Chen Z, Omote K. Toward achieving anonymous NFT trading. IEEE Access, 2022, 10: 130166-130176

[29] Mochram R A A, Makawowor C T, Tanujaya K M, et al. Systematic literature review: Blockchain security in NFT ownership//Proceedings of the 2022 International Conference on Electrical and Information Technology (IEIT). Malang, Indonesia, 2022: 302-306

[30] García R, Cediel A, Teixidó M, et al. Semantics and non-fungible tokens for copyright management on the metaverse and beyond. ACM Transactions on Multimedia Computing, Communications and Applications. to appear

[31] Pungila C, Galis D, Negru V. A new high-performance approach to approximate pattern-matching for plagiarism detection in blockchain-based non-fungible tokens (NFTs). arXiv preprint arXiv: 2205.14492, 2022

[32] Prihatno A T, Suryanto N, Oh S, et al. NFT image plagiarism check using EfficientNet-based deep neural network with triplet semi-hard loss. Applied Sciences, 2023, 13(5): 3072

[33] Song Mingxiao, Liu Yunsong, Shah A, Chava S. Abnormal trading detection in the NFT market. arXiv preprint arXiv: 2306.04643, 2023

[34] Wen Xiaolin, Wang Yong, Yue Xuanwu, et al. NFTDisk: Visual detection of wash trading in NFT markets//Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI 2023). Hamburg, Germany, 2023: 1-15

[35] Victor F, Weintraud A M. Detecting and quantifying wash trading on decentralized cryptocurrency exchanges//Proceedings of the Web Conference 2021 (WWW 2021). Ljubljana, Slovenia, 2021: 23-32

[36] Roy S S, Das D, Bose P, et al. Demystifying NFT promotion and phishing scams. arXiv preprint arXiv:2301.09806, 2023

[37] Chan V, Choi E. NFT Fraud Detection System. https://wp.cs.hku.hk/2022/fyp22012, 2023

[38] Li Sijia, Gou Gaopeng, Liu Chang, et al. TGC: Transaction graph contrast network for Ethereum phishing scam detection //Proceedings of the 39th Annual Computer Security Applications Conference (ACSAC 2023). Austin, USA, 2023: 352-365

[39] Wu J, Yuan Q, Lin D, et al. Who are the phishers? Phishing scam detection on Ethereum via network embedding. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2020, 52(2): 1156-1166

[40] Kim H, Cui J, Jang E, et al. DRAINCLoG: Detecting rogue accounts with illegally-obtained NFTs using classifiers learned on graphs. arXiv preprint arXiv:2301.13577, 2023

LIAO Peng, Ph. D. candidate. His research interests include Web3 security and Malware.

FANG Bin-Xing, Ph. D., professor, Academician of the Chinese Academy of Engineering. His research interests include computer architecture, computer network, information security.

LIU Chao-Ge, Ph. D., associate researcher. His research interests include Web security and Botnet.

WANG Zhi, Ph. D. His research interests include Botnet and AI security.

ZHANG Yun-Tao, Ph. D. His research interests include Binary security and AI security.

CUI Xiang, Ph. D., professor. His research interests include blockchain and malware.

Background

NFT counterfeiting is belongs to the field of blockchain security. With the development of blockchain technology and the popularity of NFTs, the counterfeiting of NFT artworks has become a significant security challenge. In a decentralized environment, anyone can easily create and sell fake NFT artworks, making it difficult for creators to distinguish between genuine and counterfeit items. Furthermore, there is a lack of effective defense mechanisms to protect their rights and interests. Currently, the international community's level of addressing the issue of NFT counterfeiting is still relatively low. This is due to the unique characteristics of NFTs and the anonymity of blockchain, which contribute to the complexity and difficulty of combating counterfeiting. While there have been attempts to tackle NFT counterfeiting, such as third-party platform audits and certification mechanisms, there are still many challenges and deficiencies in these approaches. This paper aims to systematically investigate the issue of NFT counterfeiting. The research team collected NFT transaction data, asset information, and similar NFT image data from multiple sources, measuring the scale of NFT counterfeiting in the Ethereum ecosystem. By employing image data augmentation techniques, they generated a dataset of attack samples with minimal perturbations and evaluated the robustness of counterfeit detection systems on mainstream NFT trading platforms and third-party commercial detection platforms. Through the construction of attack samples and the verification of their ability to bypass detection, this paper has made progress in addressing the problem and revealed the limitations of existing detection systems. In the field of blockchain security, previous research achievements may include studies on blockchain protocol and smart contract security, decentralized finance security, identity verification, and other related areas.