面向Tor网络的网站指纹防御技术研究综述

王康旭[1]，扈红超[1]，程国振[1,2,3]，任权[1,2,3]，周大成[1,2,3]，常慧妍[1]，刘文彦[1,2,3]，杨晓晗[1,2,3]
（1.信息工程大学，河南 郑州 450002；2.河南省网络空间内生安全重点实验室，河南 郑州 450000；3.网络空间安全教育部重点实验室，河南 郑州 450000）

摘要：针对Tor网络加密流量中包时序、流量突发等侧信道特征泄露引发的匿名性威胁，首先，阐述Tor网络架构与网站指纹安全问题；其次，提出网站指纹防御层次化分类架构，对流量特征混淆、攻击模型误导、匿名集增强、跨层协同优化防御进行横向与纵向对比分析；接着，梳理公开评估框架与数据集，结合防御效果、系统开销与部署特性开展评估体系构建与防御选型；最后，凝练概念漂移、对抗鲁棒等核心挑战，展望自适应防御、Tor原生集成等前沿方向。

关键词：Tor网络；网络安全；网站指纹防御；综合评估
文献标志码：A
doi：**********

Survey on website fingerprinting defense technologies for Tor network

WANG Kangxu[1], HU Hongchao[1], CHENG Guozhen[1,2,3], REN Quan[1,2,3], ZHOU Dacheng[1,2,3], CHANG Huiyan[1], LIU Wenyan[1,2,3], YANG Xiaohan[1,2,3]
1. Information Engineering University, Zhengzhou 450002, China
2. Key Laboratory of Cyberspace Endogenous Safety & Security of Henan Province, Zhengzhou 450000, China
3. Key Laboratory of Cyberspace Security, Ministry of Education of China, Zhengzhou 450000, China

Abstract: This study clarified the Tor network architecture and website fingerprinting security risks from side-channel leaks of packet timing and traffic bursts in encrypted traffic. It proposed a hierarchical classification framework for website fingerprinting defenses, and horizontally and vertically compared four defense categories: traffic feature obfuscation, attack model misdirection, anonymity set enhancement, and cross-layer collaborative optimization. It reviewed public evaluation frameworks and datasets, and constructed an evaluation system integrating defense effectiveness, system overhead and deployment characteristics for defense selection. Core challenges including concept drift and adversarial robustness are summarized, and frontier research directions such as adaptive defense and native Tor integration are outlined.

Key words: Tor network, cybersecurity, website fingerprint defence, comprehensive assessment

0 引言

随着信息技术飞速发展，用户对隐私保护的需求日益增强[1]，匿名通信技术成为应对此需求的核心手段，Tor网络[2]则是当前应用最广泛、最具代表性的主流匿名通信系统。Tor网络依托全球分布式中继节点设施，基于洋葱路由协议设计[3]，数据经多层加密后由多跳中继转发[4]，每跳仅解密一层密文，隔离用户身份与通信内容，为高敏感、高匿名性需求的网络活动提供端到端的隐私保护。通

收稿日期：XXXX–XX–XX；修回日期：XXXX–XX–XX
通信作者：程国振，邮箱 chenggz86@163.com
基金项目：国家网安重大专项课题（No.2025ZD1501105）；河南省重大科技专项（No.241110210200）
Foundation Items：National Cyber Security National Science and Technology (No.2025ZD1501105) , Major Project The Major Science and Technology Special Projects of Henan Province (No.241110210200)

常，Tor网络采用定长信元封装数据以抵御包长分析。然而，Tor网络难以隐藏传输过程中产生的时序、突发模式、传输方向等侧信道特征[5-7]，这些特征与目标网站具有强相关性，攻击者通过在入口节点或客户端采集流量，提取流量特征并构建分类模型，即可精准识别用户访问的网站，实现网站指纹的识别[8-12]。目前，依托传统机器学习、深度学习等前沿技术持续迭代赋能的网站指纹识别技术[13]，对Tor网络的匿名性构成了严峻威胁。

为有效应对网站指纹识别引发的安全威胁，网站指纹防御技术通过破坏Tor流量特征的可区分性，避免攻击者利用流量特征识别目标网站[14-15]。随着网站指纹识别技术的演进，防御手段也从早期的静态填充，发展到自适应动态优化、对抗机器学习及信息论优化等多个前沿方向[16-17]。

近年来，网站指纹攻防领域涌现大量综述文献，为领域发展奠定了坚实基础。邹等[18]构建了涵盖网络层、应用层及复合层的防御分类框架，结合填充、混淆与分割等机制完成方案梳理，为后续研究提供基础参照。杨等[10]聚焦Tor网络，从实验方法、特征选择与数据集等维度对攻击方法进行详细对比，并将防御方案划分为随机化防御、正则化防御和对抗性防御三类，提升了特定场景下的实验分析深度。Liu等[19]系统综述了深度学习范式下CNN、LSTM及Transformer等架构的演进脉络。Cui等[6]则创新性地将数据集分析提升至与攻防并列的核心地位，构建更精细化的防御体系。尽管上述工作从不同视角推动了网站指纹防御综述研究的发展，但在以下三个方面仍有进一步深化的空间。

其一，在防御分类体系的完备性方面。现有综述多采用单一维度划分，如部署层次[18]、策略类型[10]、设计原则[6]等，虽能勾勒技术轮廓，但难以同时呈现防御方案的机理属性与实现特征，易丢失方案关键技术细节。此外，数据投毒、跨层协同、生成式AI等近年新兴防御技术也难以被完整纳入现有分类体系。

其二，在综述分析的深度方面。现有综述缺乏对防御机理的深层次解构：宏观上，未能充分提炼各类技术的共性优势与本质缺陷；微观上，缺少对不同防御机理、作用环节和适用边界的横向对比。

其三，在防御方案的覆盖时效与前沿追踪方面。早期的研究工作如文献[18]尚未覆盖扩散模型、后门干扰、互信息最小化等近年代表性防御方案；近年来的相关研究如文献[19]也仅局限于深度学习视角，对传统机器学习、信息论优化及协议层协同等非深度学习方案探讨不足。

为系统呈现本文与已有综述的研究增量，本文遵循定位、分类、覆盖、分析的逻辑链，从场景聚焦、分类维度、技术广度与分析深度四个层面构建对比维度，如表1所示，进一步阐明本文在分类粒度、类内总结提炼与类间横向对比上的差异性。

针对上述不足，本文首先从Tor网络匿名通信架构出发，分析网站指纹面临的四类安全问题；其次，提出一种基于防御机理与技术路径的双维度层次化分类架构，将现有防御方案划分为流量特征混淆、攻击模型误导、匿名集增强与跨层协同优化四大类十一子类，逐一对各类方案的技术原理、优势与固有局限进行了纵向对比分析，并通过横向对比厘清不同防护路线的内在联系与核心区别；然后，梳理评估体系与数据集，综合防御效果、系统开销与部署特性三个维度，剖析了Tor网络网站指纹防御的技术演进规律并给出场景化选型思路；最后，归纳概念漂移、对抗鲁棒性、多目标权衡等固有技术瓶颈，并提出了相应的未来研究展望[20-21]。本文组织框架如图1所示。

表1 本文与已有综述工作对比

对比维度	[6]	[10]	[18]	[19]	本文
聚焦Tor场景	□	□	□	□	□
技术机理分类	□	□	□	□	□
防御分类粒度	5类	3类	3类	2类	4大类11子类
传统ML防御	□	□	□	□	□
生成式AI防御	□	□	□	□	□
部署可行性分析	□	□	□	□	□
多维对比分析	□	□	□	□	□

1 Tor网络与网站指纹攻防概述

1.1 Tor网络

Tor是全球主流的低延迟匿名通信系统，基于第二代洋葱路由架构，依托全球志愿者运营的中继节点构建分布式覆盖网络，可实现用户身份与行为的抗追踪隐私保护，是网站指纹（Website Fingerprinting，WF）攻防研究的标准场景[22]。Tor网络架构主要由用户客户端、中继节点与目录权威节点

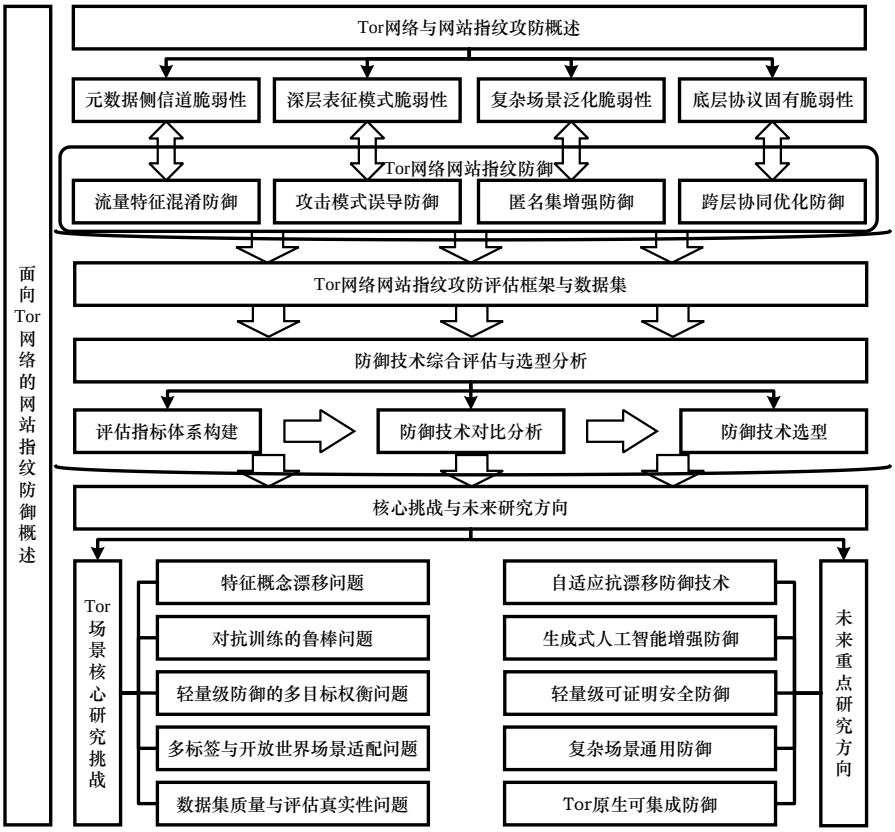

图1　本文组织框架

组成，如图2所示，其通信链路通过三层自主选择的中继节点构建，采用逐跳分层加密机制保障通信匿名性。通信数据被封装成512字节的定长信元（Cell），每个中继仅能解密一层并获知相邻节点地址信息，使得入口节点无法感知通信目标、出口节点无法获取用户真实身份，实现通信双方的身份解耦与匿名隔离。此外，Tor通过配套目录权威节点共识、逐跳Diffie-Hellman密钥协商、电路定期重建等核心机制，形成了完整的匿名通信体系[23]。

Tor的原生设计特性构成了网站指纹攻防博弈的底层约束。一方面，端到端加密完全屏蔽载荷内容，网站指纹攻击仅能依托可被动观测的流量元数据完成站点识别[24]；另一方面，定长信元传输与标准化链路转发逻辑，使得同一网页加载产生的流量元数据具备稳定的可区分性[25]，为网站指纹识别的特征提取与模型训练提供了数据基础[12]。同时，Tor的分布式架构与开源协议，决定了网站指纹防御方案必须兼容其原生通信逻辑，这一约束限制了防御方案的设计空间，任何防御方案均需在匿名保护强度、额外开销以及实际部署可行性三者之间进行权衡[26]，这也成为当前网站指纹防御技术应用的核心挑战[10]。

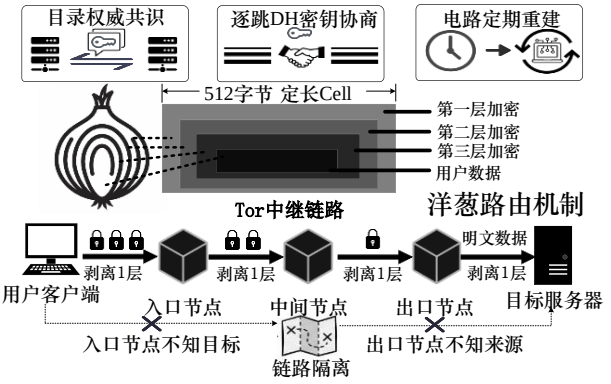

图2　Tor匿名通信系统架构

1.2　网站指纹安全问题

网站指纹识别是针对Tor网络的被动流量分析攻击[29]，其核心是通过流量元数据构成的独特“指纹”来识别用户访问的网站[11]。自2009年Herrmann等首次在Tor环境下验证其可行性以来[10]，该领域在网站指纹识别方法与威胁认知上持续拓展。

典型识别流程包括流量监听捕获、特征提取、

模型训练以及实时识别四个环节[29-30]（如图3）。攻击者在Tor客户端与入口中继间的链路上实施被动监听，捕获原始加密流量[27]，随后从流量中提取区分特征。传统攻击依赖于专家设计的统计特征[31]，例如包总数、上下行包数量比例、包序等，而基于深度学习的攻击则实现了从原始流量中端到端地自动学习隐式高阶特征[32]。模型在标注数据集上完成训练后，即可对未知流量实施在线分类[27]。

随着机器学习技术的发展，网站指纹识别逐步揭示出Tor匿名性在人工统计特征、深度学习表征、大流量混淆特征以及底层协议栈等不同层面的固有脆弱性：

（1）在早期机器学习阶段，基于kNN、SVM等算法的攻击证实，加密流量在传输时序、突发间隔及上下行包序列等统计特征中，存在可利用的侧信道信息。代表性模型如基于累计包大小特征的CUMUL以及采用随机森林特征选择与kNN分类的k-FP等，通过专家设计的统计特征建模流量模式与目标网站间的映射关系，在封闭世界中将识别准确率提升至90%以上[6]。

（2）基于深度学习的网站指纹识别采用训练好的模型对捕获的流量实时分类，通过捕捉特定包方向序列模式、微小时延抖动规律等细粒度侧信道特征推断目标网站[27]。代表性的模型如端到端CNN的DF、多通道残差网络的Var-CNN、融合时序特征的Tik-Tok等[10]，核心是利用深度神经网络自适应表征学习能力，从原始流量序列中挖掘高判别性深层特征。其揭示了攻击方已演变为具备持续学习能力的智能代理，可破解基于固定规则的防御模式。

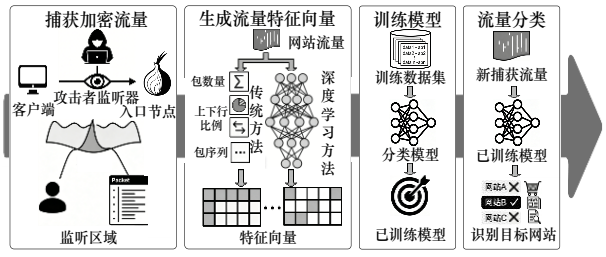

图3　Tor网络网站指纹识别流程

（3）在大流量混淆的复杂真实场景下，针对Tor流量的长时序依赖、多流交织、细粒度突发等特征，攻击者围绕开放世界、小样本学习、泛化能力提升[33-34]、Transformer架构应用[35]、多标签浏览场景适配[36]等方向建模，提出LASERBEAK等方法，通过多通道Transformer注意力机制来捕捉细粒度与时序特征，可从海量背景流量中高效低误报识别目标。此类方法暴露了防御策略在应对混合流量与低信噪比环境时的不足，表明仅依靠流量特征局部优化无法从根本上降低目标被识别的风险。

（4）在Tor网络的底层协议栈设计层面，电路指纹攻击、信息论不变量攻击与端到端时序攻击等方法，直接利用Tor协议栈或系统架构中固有的侧信道信息[10]，如固定信元处理、电路复用调度以及内核操作引入的微小时序偏差等，绕开混淆防御策略实现精准识别。其揭示了仅依靠流量特征混淆难以从根源上遏制侧信道信息泄露，Tor在面对持续演进的自适应攻击时，长期匿名性存在根本性隐患。

综上，以深度学习为代表的先进网站指纹识别技术，对Tor网络造成了极大的威胁，在封闭世界下顶尖模型的识别准确率超98%[37]，开放世界中模型也能在保持高召回率的同时将误报率控制在极低水平[38-39]。同时，Tor的多层级脆弱性构成攻击技术持续演进的基础，攻击的每次革新都不断重塑防御安全边界，推动着防御机制突破单纯特征混淆局限，向更具对抗性、智能性与系统性方向演进。

1.3　网站指纹防御

网站指纹防御技术旨在保护Tor网络免受侧信道攻击，核心原理是通过流量改造、模型干扰、匿名集扩容、协议优化等手段，破坏加密流量与目标网站的关联关系，降低网站指纹攻击模型的识别准确率，同时最小化带宽、延迟与计算开销[29]。早期研究主要依赖基于固定规则的静态混淆方案，以消除流量特征差异[40-41]。为缓解安全与开销的矛盾，后续研究转向自适应优化与匿名集增强[15,42-44]。目前，研究前沿已经拓展至对抗人工智能驱动的自适应伪装、基于信息论探索的可证明安全边界，以及追求防御轻量化与协议的原生适配[45-49]。

为增强网站指纹的防御能力，本小节从上述各类网站指纹安全问题出发，将当前主流的网站指纹防御技术归纳为四大类：

（1）流量特征混淆。该类技术旨在通过改变流量的外在表现，切断统计特征与特定网站之间的关

联。主要手段是对数据包的时序和长度分布进行重塑，构建一个特征重叠的混淆空间，从而增加指纹识别的难度。由于其对现有协议兼容性好、部署门槛低，且可有效提升侧信道的观测不确定性，已成为当前应用最广泛的主流防御方案之一。

（2）攻击模型误导防御。该类技术将防御重心从被动隐藏转向主动对抗，引入了对抗博弈的思想。其核心是主动构造误导性流量或对抗样本，干扰攻击者的特征提取链路，通过模糊分类边界诱导模型产生误判。其不仅有流量层面的伪装，更有认知层面的对抗，旨在大幅增加攻击者的训练成本与网站指纹识别难度。

（3）匿名集增强防御。该类技术通过特征归一化或引入诱饵流量，构建特征高度相似的候选集群来稀释个体的可辨识度。该机制实质上是将防御转化为集合可区分性难题，使得攻击者的搜索空间呈指数级扩张，从数学上降低单点暴露风险。在这种策略下攻击者即便获取指纹特征，也难以在大规模匿名集中锁定目标，从而有效提升系统整体鲁棒性。

（4）跨层协同优化防御。该类技术突破单一流量层局限，从协议栈各层级交互中寻求系统性增益。其通过融合信息论与协议设计原理，利用跨层资源联合调度来探索设计防御方案。这种多维协同机制能有效弥补单层防御盲区，在系统层面实现纵深防御效应，为平衡安全性与实用性提供新的解决路径。

上述四类主流防御技术从不同维度阐述了网

站指纹防御的核心机理，能够在不同应用场景下削弱流量特征泄露风险、抵御指纹识别攻击。但当前防御方案在实际落地中仍面临仿真环境与真实场景差距大、防御开销高、长期鲁棒性缺失[50]以及工程部署困难[51]等多重挑战。

从本质上看，Tor网络网站指纹攻防是持续动态博弈，其演进受限于协议设计、网络环境与机器学习能力等的多重约束。攻击方持续提升识别精度，对用户隐私构成严峻威胁；防御方则是在兼容Tor机制与控制开销的前提下，通过多维技术手段削弱甚至压制这种识别能力。面对这一复杂攻防博弈态势，亟需对现有防御手段进行体系化的梳理与解构，

以厘清不同技术路径的演进规律与适用边界，从而

为后续构建高效的防御体系提供理念支撑。

2　防御技术分类与分析

本节立足Tor网络网站指纹防御的机理与技术演进逻辑，提出了一种层次化的网站指纹防御分类架构。该架构一方面依据Tor网络的多层级脆弱性与防御机理，将现有防御技术划分为流量特征混淆、攻击模型误导、匿名集增强及跨层协同优化四大类；另一方面，从技术实现路径角度进行二级细分，实现对防御技术精细化覆盖与多维度解构。该框架较为完整覆盖Tor场景下的经典方案与前沿研究成果，并能灵活适应未来新型技术的发展。本节将遵循此分类体系（图4所示），对防御技术逐一展开分析。

2.1　流量特征混淆防御

流量特征混淆防御是Tor较成熟的防御手段，通过修改信元序列、时序及突发模式等关键属性，解耦流量与网站的关联，干扰或破坏攻击模型的特征提取过程。本文将该类方法分为静态填充、自适应动态、多路径分散三种实现方案。由于这类技术通常适配Tor固定信元传输机制，且无需修改Tor核心协议，因此其部署与实现难度相对较低。

2.1.1　静态填充型

静态填充型是流量特征混淆防御中最早被探索的分支，其核心是通过预定义的固定规则对数据包进行无差别填充，弱化不同网站流量的特征差异来抵抗网站指纹攻击，原理图如图5所示。Liberatore等[52]搭建了基础填充框架，验证了规则填充的可行性，这一思想也被后续早期方案所沿用[53]。

早期BuFLO[41]通过固定包长与间隔掩盖特征，明确了安全开销制衡关系，但均匀的流量特征易被识别，且开销较高。为改善工程局限，Cai等[40,54]率先引入拥塞感知与自适应传输机制，提出拥塞敏感

型缓冲定长混淆方案CS-BuFLO以提升复杂网络环境下的流量适配性能；继而其采用上下行流量差异化处理与固定倍数填充策略，构建具备可证明安全属性的Tamaraw，实现对多维流量特征的全面混淆。

该方向随后向理论建模与轻量化深化。Wang等[11]证明匿名集最短公共超序列为带宽最优解，

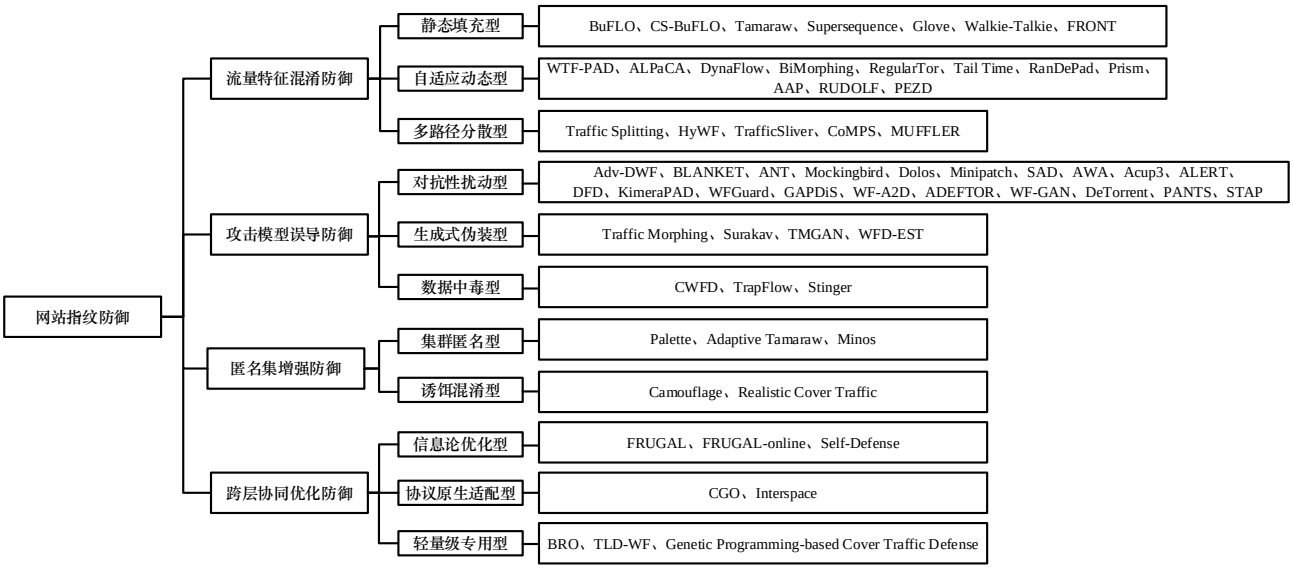

图4　Tor网络网站指纹防御技术分类架构

通过预聚类匿名集与预定义超序列填充规则，实现了可证明安全。Glove[55]沿此思路借助预定义超序列实现轻量提效，但其维护成本较高。在此基础上，后续研究突破单纯填充局限。Walkie-Talkie[43]引入半双工通信重塑流量形态；FRONT[56]基于瑞利分布规则聚焦混淆流量的前端特征，无额外主动延迟，在较低带宽开销下超越了同期最优轻量化防御方法。

图5　静态填充型防御原理图

表2对比了典型静态填充型防御技术。该类防御经历从规则固化到多维优化、高开销到轻量化的演进，且逐步具备可证明安全性，其机制简明易实现，但静态规则导致开销较高并难以适应动态环境。

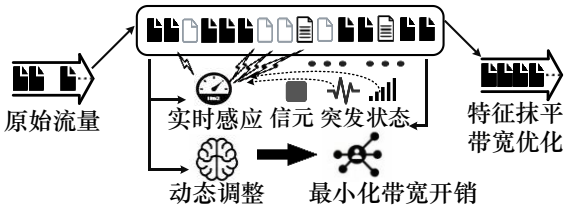

图6　自适应动态型防御原理图

2.1.2　自适应动态型

自适应动态型防御通过实时感知流量特征与网络状态，动态调整填充策略，以更低的带宽开销解决静态填充规则固化、易被识别的局限，原理如图6。WTF-PAD[42]作为该方向开创性研究，搭建突发与间隙双模式有限状态机，实现无主动延迟的动态填充。在此基础上，ALPaCA[57]在应用层通过概率采样生成平均化资源特征，按参数动态补齐资源维

表2　　**典型静态填充型防御技术对比**

防御方案	攻击模型	核心优势
BuFLO[41]	SVM、朴素贝叶斯、VNG++	固定包长与时序，强混淆粗粒度特征
CS-BuFLO[54]	VNG++、SVM、DLSVM	拥塞感知自适应
Tamaraw[40]	理想攻击者、SVM、VNG++	上下行差异化填充，可证明安全
Supersequence[11]	kNN、SVM、CUMUL	理论最优开销混淆
Glove[55]	kNN、SVM、CUMUL	轻量化模板混淆
Walkie-Talkie[43]	主流机器学习、传统指纹识别模型	半双工突发成型，阻断双向时序泄露
FRONT[56]	kNN、CUMUL、k-FP、DF	极低延迟轻量防御

度。后续研究主要围绕以下不同层面开展优化。

为削弱流量突发特征，DynaFlow[58]通过规整双向流量动态更新发送间隔，摆脱模板数据库依赖，适配动态网页场景；BiMorphing[44]聚焦双向流量关联特征，以差异化时序混淆降低带宽开销 。

在优化混淆传输性能，RegulaTor[51]利用 Tor 流量的突发衰减规律与上下行关联特征，实现双向特征正则化处理，弱化流量关联指纹；Tail Time[59]利用自适应尾部超时调度策略，有效缓解半双工通信引发的页面加载阻塞。

面向复杂攻击场景，RanDePad[60]通过滑动窗口实时评估，扰动流量时空分布；Prism[61]利用时序学习模型挖掘流量依赖关系，兼具双防御模式与抗重训练攻击能力；AAP[62]与 PEZD[63]均采用客户端单侧部署模式，在无额外主动延迟条件下完成流量伪造与特征隐匿。

随着智能算法的融入，RUDOLF[16]引入强化学习机制生成动态填充策略，提升模型对流量概念漂移的环境适应性。综上，自适应动态型防御实现了从单一感知到多元融合的技术演进，兼具灵活性与可控开销，但其效能依赖流量感知精度，且算法复杂度偏高。表 3 总结了典型自适应动态型防御技术。

2.1.3　多路径分散型

多路径分散型防御（图 7）从传输路径维度拆解流量，阻断攻击者完整指纹采集，避免了传统填充方法的高开销问题，具备效率优势。该方向早期源于 Tor 性能优化，Conflux[64]通过信元级流量拆分优化吞吐量，客观实现特征混淆效果。随后，De la Cadena 等[65]基于 Tor 多入口中继架构，设计多种流量拆分策略，验证了多路径混淆防御的可行性。

在此基础上，相关研究向客户端延伸，HyWF[66]利用终端多宿主接入特性，借助 MPTCP 协议完成端侧流量拆分；TrafficSliver[67-68]设计了网络层与应用层互补的异构拆分架构，提升部署通用性。为降低协议依赖，研究转向轻量化演进，CoMPS[69]依托 QUIC 等协议的连接迁移与 IP 漫游能力实现异构路径动态调度，无需修改服务端；MUFFLER[70]通过动态连接洗牌弱化流量关联，适配多种场景。

多路径分散型防御推动了防御理念从被动流量填充向主动特征消除的转变，其开销低且隐蔽性强，但也因多链路而面临调度延迟高、部署复杂等挑战。典型多路径分散型防御技术对比如表 4 所示。

图 7　多路径分散型防御原理图

2.1.4　分析总结

静态填充型、自适应动态型与多路径型防御均通过重塑流量统计分布，扩大特征重叠域来弱化攻

表 3　典型自适应动态型防御技术对比

防御方案	攻击模型	核心优势
WTF-PAD[42]	kNN、SVM、VNG++、DL	自适应间隙填充，极低延迟，轻量化高鲁棒
ALPaCA[57]	kNN、k-FP、CUMUL	易部署且开销可控
DynaFlow[58]	kNN、k-FP	开销低无需预存模板
BiMorphing[44]	BIND、CUMUL、kNN	双向突发混淆，极低延迟低开销强防御
RegulaTor[51]	Tik-Tok、DF、CUMUL	上下行流量规整，防御强且开销更低
Tail Time[59]	kNN、DF	半双工尾部阻塞降时延
RanDePad[60]	CUMUL、k-FP、DF	自适应延迟可控填充
Prism[61]	FS-Net、DeepCorr、WF-LSTM	非对称实时混淆，鲁棒性强，开销低
AAP[62]	CUMUL、k-FP、DF	极低延迟混淆突发特征
RUDOLF[16]	Tik-Tok、RF、TF、DF、CUMUL、k-FP	SAC 实时自适应，低带宽极低延迟抗漂移
PEZD[63]	Var-CNN、DF、CUMUL、Tik-Tok	多分布极低延迟，网站无关低耗高效防御

表4　典型多路径分散型防御技术对比

防御方案	防御机制	核心优势
Traffic Splitting[65]	多入口随机加权分流	无虚拟包极低延迟，强效破坏流量指纹
HyWF[66]	多宿主随机集合分流与MPTCP调度	零开销多路径混淆兼容Tor强防御
TrafficSliver[67]	网络层与应用层多入口批量加权分流	易兼容Tor部署，强效破坏指纹
CoMPS[69]	连接迁移驱动，会话内跨多路径分流	链路迁移混淆，轻量高效强抗指纹
MUFFLER[70]	动态连接洗牌拆分、eBPF隧道优化	虚实连接动态混拆，极低开销抗流关联

击模型的特征区分能力。但在解决流量特征如何改造的核心问题时，三类方案采取了截然不同的技术路径，进而在防护思路、适用场景与固有局限等方面呈现出显著分化，具体对比如表5所示。

总体而言，流量特征混淆技术演进沿双重逻辑推进：作用机制由被动掩蔽转向主动解耦，实现路径由静态规则迭代为智能适配及架构级拆分。该类方案依托对Tor传输机制的深度适配形成独有优势，混淆操作仅在客户端或应用层执行，无需修改Tor协议栈，部署门槛较低。填充、延迟与分流等操作均为轻量级原语，不涉及模型训练或高维优化，实时性有保障。防御逻辑与攻击模型无关，面对未知的新型攻击具备一定的通用防护能力。然而，该类方案受底层机理制约，仅能弱化流量区分特征，无法根除侧信道信息残留，持续迭代的高阶攻击可逐层剥离混淆增益，且防御方在安全强度与传输开销的二元博弈中处于被动，难以抵御攻击模型迭代与流量概念漂移的长期挑战。

2.2　攻击模型误导防御

攻击模型误导防御是Tor网络网站指纹智能防御的主流研究方向，其核心并非直接修改流量特征，而是通过干扰攻击模型的特征学习与分类决策过程，诱导其产生误判，导致无法正确识别目标网站。该方法通常无需大幅增加带宽或延迟开销，防御效果常优于传统特征混淆型方案，并能与现有填充、多路径等防御机制兼容叠加，尤其适合Tor客户端等算力受限环境。根据攻击链路的干预环节及实现原理可将其分为对抗性扰动型、生成式伪装型与数据中毒型防御三类。

2.2.1　对抗性扰动型

对抗性扰动型防御基于对抗机器学习思想，向流量中注入定制化扰动干扰指纹识别模型，并结合Tor信元传输特性调整时序与发包间隔，兼顾低开销与高隐蔽性，其基本原理如图8所示。该思路源自经典对抗样本理论[71]，Imani等[72]率先将其引入网站指纹防御，提出利用梯度迭代生成对抗轨迹的Adv-DWF方案，奠定领域研究基础。

为实现实时防御，后续研究突破了全量流量预知约束。BLANKET[73]实现无流量先验的盲对抗扰动生成，后续方法采用网站级离线预生成通用扰动模式，显著提升防御响应效率[74-75]。扰动生成机制主要围绕策略优化设计与实现简化展开。一方面，动态切换扰动目标以抵御对抗训练[26]；另一方面，则探索利用预计算对抗补丁进行干扰[76]，并进一步摆脱对白盒梯度的依赖[77]以及利用分块处理降低开销[78]。

为应对对抗训练攻击，提升防御鲁棒性能成为研究重点。现有方案借助多模型变换、特征模仿、生成式扰动等手段，适配高阶对抗攻击与开放世界场景[79-81]。为平衡防御性能与部署成本，研究采用精准注入、强化学习等智能方法优化扰动策略，同时完善梯度适配、客户端轻量化部署等技术[46,82-86]。此外，该方向不断拓展应用边界，利用生成式模型优化离散扰动、抵御关联攻击，衍生出

表5　流量特征混淆防御子类对比

子类	解决思路	适用场景	根本局限
静态填充	预设固定规则统一流量形态	高安全容忍、不计开销	固定模板与动态环境持续失配
自适应动态	实时感知流量，按需注入混淆	通用浏览，平衡隐私与体验	策略可学习，攻防陷入循环
多路径分散	拆分流量，阻断指纹采集	多接入点终端	路径独立性在真实Tor中难以保证

低延迟工程化部署、服务端单侧部署等多元方案[87-91]。

总体来说，对抗性扰动型防御方案开销低、隐蔽性强，但整体鲁棒性易受对抗训练制约。其技术体系较为完备，但真实网络中的落地效果与开放世界适配能力仍有待验证。典型对抗性扰动防御技术对比如表6所示。

2.2.2　生成式伪装型

生成式伪装基于生成式人工智能（如GAN、扩散模型）学习Tor正常流量的分布特征，生成与真实Tor流量高度相似的伪装信元序列，与原始流量融合传输，使攻击者无法区分真实流量与伪装流量，其基本原理如图9所示。

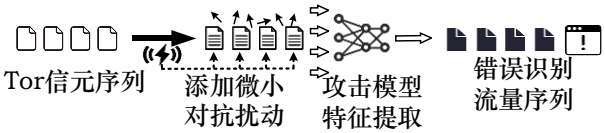

图8　对抗性扰动型防御原理图

早期奠基性工作Traffic Morphing[53]首次将凸优化技术引入该领域，通过离线构建映射矩阵拟合流量分布，误导攻击模型的分类决策。随着生成式模型发展，Surakav[45]基于改进的WGAN-div

表6　典型对抗性扰动型防御技术对比

防御方案	核心对抗机制	可防御攻击模型
Adv-DWF[72]	梯度迭代匹配目标轨迹	DF、CUMUL、CNN SDAE、
BLANKET[73]	离线生成器与在线三维盲扰动	DF、Var-CNN、 DeepCorr
ANT[75]	三维通用对抗扰动注入	CNN、SDAE、DNN
Mockingbird[26]	随机目标轨迹与梯度逼近扰动	DF、Var-CNN、AWF、CUMUL
Dolos[76]	密钥参数化对抗补丁注入	DF、Var-CNNkNN、k-FP、CUMUL
Minipatch[77]	黑盒对偶退火生成补丁	AWF、DF、Var-CNN
SAD[78]	分块梯度扰动	DF、Var-CNN、AWF
AWA[79]	密钥随机变换对抗扰动	DF、Var-CNN、AWF
Acup3[80]	多对一网站模仿与概率扰动变异	DF、Var-CNN、AWF
ALERT[81]	GAN无迹对抗扰动	DF、Var-CNN、AWF
DFD[82]	逐突发内按比例注入包	DNN、DF
KimeraPAD[83]	DRL智能体实时调度注入	DF、TF、CUMUL、k-FP、Tik-Tok、RF
GAPDis[84]	梯度序列编辑禁忌搜索	DF、Var-CNN、AWF
WF-A2D[85]	两阶段级联对抗学习与填充拆分	FS-Net、Var-CNN、NeuTic、DeepCorr、AppSniffer、TAM-CNN、Tik-Tok
WFGuard[86]	模糊测试驱动梯度注入	DF、Var-CNN
ADEFTOR[46]	增量距离缩减与通用失真MIM梯度优化	DF、Var-CNN、CUMUL、k-FP、kNN
WF-GAN[87]	GAN突发特征转换	DF
DeTorrent[89]	LSTM实时生成dummy填充	Tik-Tok、DF、DeepCoFFEA
STAP[90]	潜态转移对抗扰动QUIC填充与插入	NeuTic、AppSniffer
PANTS[91]	AML与SMT结合的延迟填充与注入	MLP、RF、CNN 、Transformer

表7　典型生成式伪装型防御技术对比

防御方案	伪装机制	可防御攻击模型
Traffic Morphing[91]	凸优化求解最优映射，实时包长分布匹配	早期分类器
Surakav[45]	WGAN-div生成真实轨迹，实时动态调节发送	k-FP、CUMUL、DF、Tik-Tok
TMGAN[92]	定向GAN扰动流量特征变形	DF、AWF、Var-CNN
WFD-EST[14]	扩散生成时空特征模板，突发感知流量对齐	FS-Net、NeuTic

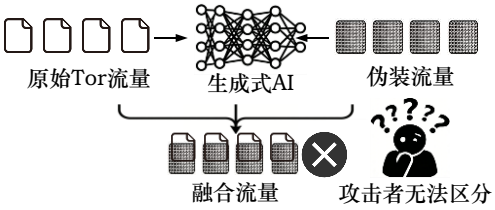

图9　生成式伪装型防御原理图

架构生成高逼真度流量突发轨迹，以可插拔模块形式完成工程化部署，无需修改 Tor 协议。为进一步提升对抗的针对性，TMGAN[92]方案实现了定向流量伪装，能够将特定源网站流量刻意伪装为目标网站流量，但离线处理模式与较高开销限制了其实际应用。

近年来该类技术向高阶生成模型迭代，WFD-EST[14]方案引入扩散模型，构建了以时序判别器为指导的三级防御架构，能合成高质量流量模板，但其尚未在真实的 Tor 网络环境中完成专项验证。生成式伪装防御可实现深层次特征混淆，但其计算开销与实时性仍是制约实际部署的关键问题。典型生成式伪装型防御技术对比如表 7 所示。

2.2.3　数据中毒型

数据中毒型防御旨在攻击模型训练阶段实时干扰，通过污染其训练数据分布，以破坏攻击模型的特征学习与泛化能力,如图 10。该思路源于对抗机器学习中的后门攻击，TrojanFlow[93]方案首次将后门攻击引入流量分类，但其“标签翻转”策略难以适配网站指纹防御场景。为此，Seceri 等[94]提出了干净标签投毒机制，证明仅修改训练样本即可毒化模型，但其高度依赖攻击者模型先验知识。

为摆脱先验约束，CWFD 方案[95]构建触发关联机制主动控制策略，完成 Tor 流量数据中毒设计，但其开销偏高且触发模式易被过滤。TrapFlow 方案

在此基础上优化触发器生成逻辑，引入统计对齐机制，实现了防御强度与通信开销的均衡优化[96]。

区别上述后门触发路线，Stinger 方案[97]基于双

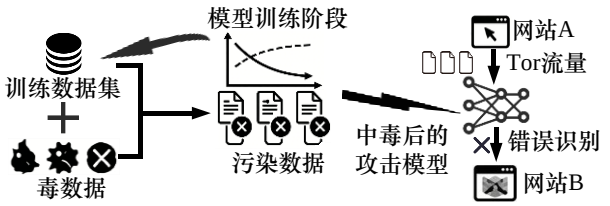

图10　数据中毒型防御原理图

阶段欺骗架构，生成符合 Tor 特征的中毒序列，并利用差异化密钥诱导模型过拟合，但效果仍依赖于中毒数据比例，且存在密钥泄露风险。数据中毒型防御为实现对攻击模型的根源性干扰提供了新路径，但现有方法在毒化机制的隐蔽性、抗过滤及动态鲁棒性等方面仍存在不足。典型数据中毒型防御技术对比如表 8 所示。

2.2.4　分析总结

对抗性扰动、生成式伪装与数据中毒三类防御方案共享核心防护逻辑，将防御重心由流量特征改造转向攻击模型干扰，依托机器学习模型固有脆弱性干扰特征表征或污染训练分布诱导误判，可在较低带宽开销下取得显著防御增益。但三者在攻击链中的干扰介入时机与技术路径差异显著，具体对比如表 9 所示。

总体而言，攻击模型误导防御实现了从流量空间被动混淆向认知空间主动博弈的范式跃迁，其技术演进沿静态对抗向自适应博弈推进。其核心优势在于实现了攻击面的主动前移，即不再被动等待攻击者提取特征，而是主动向攻击模型的推理或训练过程注入干扰，以极低的带宽开销即可显著压制深度学习攻击的识别精度。对抗扰动支持定制化设计，既能面向特定攻击网络架构生成专用干扰，也可依托黑盒查询反馈适配未知模型，差异防护灵活性更强。该类方案将防御窗口从“部署后”前移至

表 8　　典型数据中毒型防御技术对比

防御方案	毒化机制	可防御攻击模型
CWFD[95]	标签已知后门，红蓝丸可控防御	RF、DF、Tik-Tok、Var-CNN、TMWF、ARES
Trap-Flow[96]	双机制中毒，目标标签统计对齐	RF、DF、Tik-Tok、Var-CNN、TMWF、ARES
Stinger[97]	诱导分类器通过拟合序列、用户密钥双向混淆	DF、AWF、Var-CNN、RF、CUMUL、k-FP、kNN、GANDaLF、Tik-Tok、Matures c、Laserbeak

表9　攻击模型误导防御子类对比

子类	解决思路	适用场景	根本局限
对抗性扰动	推理阶段注入干扰，偏移分类边界	需快速轻量部署	攻防非对称，重训练即可恢复
生成式伪装	合成高仿真流量，消除可区分特征	需深度伪装的高对抗场景	生成器可被逆向拟合
数据中毒	训练阶段污染数据，破坏模型学习	可接触攻击者训练管道	投毒可检测，跨模型迁移未验证

“训练前”，从源头抑制攻击模型的特征学习能力，攻击者即使收集全部防御后流量也无法恢复模型性能。然而，该类技术的根本困境在于攻防双方的非对称博弈格局，防御方需适配高维模型空间中任意可能的攻击策略，而攻击方仅需针对已知防御重训练即可恢复识别能力，攻防双方持续迭代优化，长期无法达到稳定均衡状态。此外，现有方案多基于封闭世界强假设与单轮对抗设定，在开放世界场景和多轮自适应攻击下的效能缺乏可证明安全保证。

2.3　匿名集增强防御

匿名集增强防御是Tor轻量级终端适配方案，其核心是扩大目标网站的匿名集规模，降低单个网站的可识别概率，使攻击者无法从海量相似流量中精准识别目标网站。依据技术路径与作用机制的不同，主要分为集群匿名型与诱饵混淆型两大类。

2.3.1　集群匿名型

集群匿名型防御是将Tor网络中流量特征相似的网站划分为同一集群，通过轻微特征调整使同类网站流量特征趋于一致，依托集群扩大匿名集规模来降低攻击模型的识别准确率，如图11所示。该技术率先完成可行性验证，Chan-Tin等[98]结合元数据聚类与基准填充策略实现k-匿名保护，在无额外时延的前提下验证了聚类防御的有效性，但基准填充方式存在高开销的缺陷。

为改善流量表征能力，精细化模板设计成为研究重点。Palette方案[15]引入流量聚合矩阵，依托k-匿名约束的聚类算法划分匿名集，并构建超矩阵模板适配Tor稀疏流量。其综合采用模板收缩、概率采样与动态填充机制，具备低泄露、强鲁棒性的优势，但其泛化能力有限且高隐私配置下开销仍偏高。

为平衡安全与开销。Adaptive Tamaraw[47]方案融合正则化防御与超序列集群匿名思想，采用分阶填充策略缩减通信成本，却难以适配多标签复杂场景。为追求高性能部署，Minos[99]借助可编程数据平面完成多流量实时交织聚合，达成网关高速部署。

综上，集群匿名型防御实现了从简单聚类填充到硬件加速优化的多元演进。但在集群智能构建、复杂网络环境下的长效防御等方面仍有待深入探索与突破。

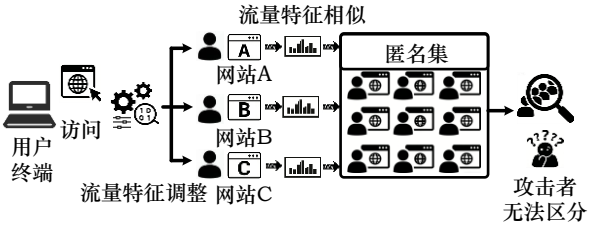

图11　集群匿名型防御

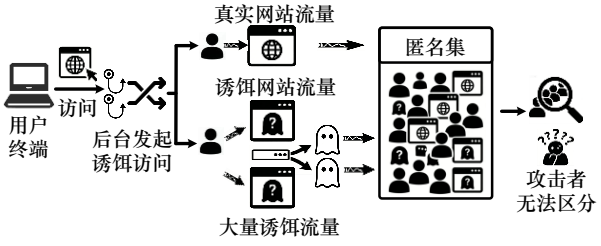

图12　诱饵混淆型防御

2.3.2　诱饵混淆型

诱饵混淆型防御通过终端并行生成诱饵流量，扩充流量匿名集以混淆真实访问行为，让攻击者无法区分真实访问与诱饵访问，如图12所示。

Camouflage[100]作为该方向奠基方案，以同步加载随机网页的方式构造诱饵流量，验证了该类方案的可行性，但随机诱饵存在特征适配性差、通信开销偏高的缺陷。为解决诱饵特征失真问题，Cui等[101]提出了真实背景流量防御，该方案依托用户历史浏览元数据，复刻真实流量的数据包序列、时序间隔及突发特征，生成高相似度诱饵流量，并通过自建噪声服务器完成通信，无需篡改原始流量特征，在压制攻击准确率的同时，实现了低带宽开销并无额外传输延迟。

总体而言，诱饵混淆型防御效果显著，技术由

高开销、易甄别逐步向轻量化、高逼真迭代优化。但该类方法高度依赖历史数据质量，且生成的诱饵流量在微观特征上仍存在被高级攻击者甄别风险。

2.3.3 分析总结

集群匿名与诱饵混淆两类方案底层防御逻辑相似，将单点可识别问题转化为集合可区分难题，通过扩增特征同构的候选流量集合，从统计层面稀释攻击模型的识别置信度。但二者在匿名集构建的核心路径上存在本质区别，前者通过被动聚类归并已有网站，后者通过主动生成仿真流量扩充候选集，具体对比如表10所示。

总体而言，匿名集增强防御的演进沿被动聚类归并向主动诱饵构造推进。该类技术的独特优势源于其统计防护的通用性，防御机制不依赖攻击模型的具体架构、参数或训练算法，具有攻击无关的天然鲁棒性，即无论攻击者采用传统机器学习还是深度学习模型，只要匿名集足够大且内部不可区分，识别精度存在理论上界。集群匿名型方案对原始流量的修改幅度极小，仅在必要时进行轻量填充以收敛群内差异，保留了流量的自然性，降低了因异常特征模式被反向定位的风险。但该类技术局限性在于防护效能受匿名集规模约束，要获得高安全保证就需要大匿名集，而大匿名集意味着高带宽开销或高诱饵构造成本，其本质为概率层面的风险分摊，无法从根源消解侧信道信息的固有可区分性。此外，匿名集防护性能取决于集群稳定性与诱饵仿真程度。网站迭代将引发集群边界漂移，高阶攻击者还能借助细粒度时序特征区分诱饵，持续削弱理论层面的安全防护上限。

2.4 跨层协同优化防御

跨层协同优化防御突破了应用层限制，从协议层、信息论及网络架构等进行协同优化，在不破坏Tor匿名性前提下，实现理论可证明安全、低资源开销与强鲁棒性。依据优化维度与介入层次不同，主要分为信息论优化型、协议原生适配型与轻量级专用型，其防御原理如图13所示。

2.4.1 信息论优化型

信息论优化型防御是Tor理论安全前沿，通过互信息最小化、信息熵最大化等弱化流量特征与网站间的关联，从本质上消除流量可区分性。

早期研究多将互信息作为防御评估指标，并未纳入优化范畴。Li等[102]率先量化Tor流量泄露机制，证实残余互信息是防御被攻破的关键，但受限于高维计算复杂度，未能构建通用优化防御模型。近期研究将理论目标转化为可实现的优化算法，在策略优化上，FRUGAL方案[103]以互信息最小化为核心优化目标，结合强化学习优化扰动策略，具备攻击无关性，开销可调控及抗流量概念漂移等优势，但离线方案依赖完整流量序列，且实验评估未针对Tor的多标签场景。在部署优化上，Self-Defense方案[48]基于信息流理论构建泄露抑制模型，通过线性规划求解最优扰动。其可在恶劣条件下有效降低攻击准确率，但目前仅验证页面尺寸这一特征维度。

综上，该类型防御理论安全、开销低且鲁棒性强，已从理论构建到算法实现并走向实际部署，但仍存在高计算开销与复杂场景适配不足等局限。

2.4.2 协议原生适配型

协议原生适配型防御是以Tor协议层级优化为核心，对Tor信元传输、电路建立、流量调度等关键流程进行原生优化，从协议根源弱化网站指纹的可区分性。该类技术主要分为以下两大分支：

在密码学优化层面，CGO方案[104]针对协议延展性缺陷与标记攻击漏洞，构建了可更新可调谐密码架构与密文链式处理机制，在零通信开销的前提下提升加密等级，增强流量随机性，为跨层防御提供了底层支撑。

在协议改良层面，Interspace方案[50]基于Tor原生填充框架设计概率化混淆机制，通过动态注入填充信元、随机扰动电路参数，解决固定混淆规则易被逆向识别的问题。该方案兼容性强、部署成本低，但易受海量训练样本拟合攻击，防御稳定性

表10 匿名集增强防御子类对比

子类	解决思路	适用场景	根本局限
集群匿名	聚类网站为群组，群内不可分	网站集稳定，可容忍填充开销	规模与开销线性绑定，边界漂移
诱饵混淆	生成仿真流量，扩充候选集	历史记录丰富，追求低延迟	逼真度依赖数据质量，可被甄别

受限。

综上，该类型防御适配性强、易部署且底层安全，但现有方案优化维度单一、综合防护能力弱，后续可融合多类优化思路构建复合型防御方案。

2.4.3　轻量级专用型

轻量级专用型防御面向移动端、物联网等资源受限终端，以跨层协同设计为优化思路，在严控资源开销的前提下实现防御。

早期方案主要以随机化策略实现轻量化。BRO方案[105]基于Beta分布进行虚拟信元的时序调度，通过参数随机初始化与冗余填充包剔除机制，以极低带宽开销实现在弱资源终端上的基础防御，为轻量化设计奠定设计基准。

为优化混淆精度，TLD-WF方案[49]依托可插拔传输架构，精准挖掘流量空窗区间并完成定向填充，

无需改动Tor核心协议且无额外时延，实现了从盲目随机混淆到感知式精准处理的演进；为适应复杂开放世界场景，轻量化防御融合智能优化算法，Ling等[106]基于遗传规划构建信元注入策略，以特征距离为约束，自适应生成轻量级扰动模式，在开放世界场景下以较低带宽开销有效抵御深度学习攻击。

综上，该类型防御开销极低、无额外延迟、易部署，适配资源受限场景，但其复杂场景下鲁棒性不足，需在资源约束与防御性能之间动态权衡。

2.4.4　分析总结

信息论优化、协议原生适配与轻量级专用三类防御方案，均突破单一流量层的防御局限，从系统层面重构防御边界，根源弱化侧信道信息的固有可区分性。然而，在从系统栈的切入层面这一核心问题上，三者遵循了截然不同的技术逻辑。信息论优化方案是从数学层面量化并最小化信息泄露，协议原生适配方案从架构层面改造Tor关键流程以消除泄漏源，轻量级专用方案则从工程层面将跨层设计适配至资源受限的真实部署环境，具体对比如表11所示。

总体而言，跨层协同优化防御技术的演进沿理论安全探索向多维度系统协同推进。该类技术的核心优势源于系统级防御视角，直接作用于侧信道信息产生的底层机制，从源头而非表象解决问题，具备可证明安全潜力。同时防御机制不依赖攻击模型具体实现，具有攻击无关的通用性，不受攻击者分类器选择的影响。此外防御逻辑深度嵌入Tor传输与加密架构，不引入额外流量层可识别模式，理论

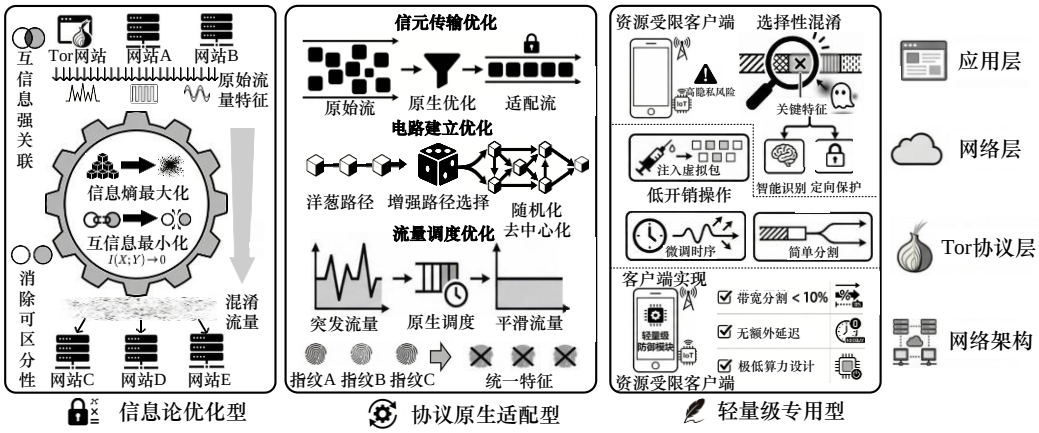

图13　跨层协同优化防御原理图

表11　　**跨层协同优化防御子类对比**

子类	解决思路	适用场景	根本局限
信息论优化	互信息最小化，量化并削减泄露	理论安全分析与参数标定	理想假设与真实Tor存在偏差
协议原生	改造Tor协议栈，根源消除可区分性	协议升级与长期架构改进	社区采纳壁垒，向后兼容约束
轻量专用	严控开销，适配弱资源终端	移动端、物联网等受限环境	复杂场景鲁棒性不足

上不存在防御痕迹被学习的风险，从根本上规避了混淆与误导类方案的共同困境。但其根本局限同样源自系统级定位。理论模型多基于理想化信道假设，与 Tor 网络实际的电路拥塞、流水线调度等复杂特性存在适配偏差，导致理论最优解在实际部署中效能折损；协议级改造需经 Tor 社区严格安全审计并保持向后兼容，落地周期漫长；此外信息论、密码学与协议调度的多维度联合优化涉及复杂跨层协调，目前尚无成熟的协同设计方法论。

2.5　防御综合分析总结

前文已就各类技术的内部机理与子类分化进行了纵向梳理，本小节从横向视角出发，对四类防御技术进行综合对比分析，系统揭示四类技术在防御逻辑、作用深度、安全属性等方面的本质区别与内在联系。表 12 总结了网站指纹防御分类体系。

各类防御技术横向对比差异性主要表现为：（1）流量特征混淆与攻击模型误导核心差异为防御逻辑。前者消除流量特征可区分性，具有攻击无关性但带宽开销高；后者干扰模型决策，低开销下防护即时性强，但易被重训练攻击突破。（2）流量特征混淆与匿名集增强核心差异为防护机理。前者消解类间特征差异，适配高安全需求场景；后者扩增候选集合稀释识别唯一性，适配低带宽约束场景。（3）流量特征混淆与跨层协同优化差异为作用深度。前者仅在流量表征层调整，无法根除侧信道泄露；后者深入协议底层，从源头消除可区分性，安全上限更高但工程代价更大。（4）攻击模型误导与匿名集增强核心差异为策略范式。二者分属主动欺骗与被动稀释两类互补策略。前者防护增益高但稳定性不足；后者性能稳定但受集合规模约束，二者协同可双向提升防护效能。（5）攻击模型误导与跨层协同优化核心差异为攻击依赖度。前者高度依赖攻击先验知识，防御效能随攻击者策略迭代持续衰减；后者完全攻击无关，安全性基于理论支撑，但存在理想假设与实际环境的适配鸿沟。（6）匿名集增强与跨层协同优化核心差异为安全保证路径。二者均可提供形式化安全保证，但理论路径不同。前者依托 k-匿名约束，属统计意义安全，效果受集群边界漂移影响；后者依托信息论，属理论意义安全，性能受理想信道与真实传输特性的偏差制约。两类路径互补，为构建高可行性防御体系提供双重理论支撑。

上述对比从多个维度厘清了四类技术的核心边界，各类技术差异可归纳为防御逻辑、作用深度的表层区分，以及攻击依赖程度、安全保证路径的

本质区别。在明确各类方案的特性边界基础上，可进一步梳理四类技术的演进脉络与内在关联，厘清其在整体防御体系中的定位与协同逻辑。

从整体来看，四类技术并非线性迭代替代关系，而是存在清晰的技术继承与分支演进脉络，沿被动遮蔽、主动对抗、概率稀释到系统优化路径逐步深化。流量特征混淆作为领域最早成熟的基础范

表 12　面向 Tor 网络的网站指纹防御技术分类、机制与演进

研究问题	解决思路	防御技术	方法	技术演进	优缺点
固定特征致网站指纹易识别问题	破坏流量与标识的统计关联	流量特征混淆防御	流量填充、时序混淆、路径分散	被动掩盖到主动切断关联，规则固定到动态智能	□技术成熟易部署，兼容性好 □对深度学习攻击鲁棒性有限
防御难应对对抗自适应 DL 攻击	误导攻击模型，构造对抗样本	攻击模型误导防御	对抗扰动、梯度误导、特征投毒	从单样本对抗向通用、自适应及轻量化部署演进	□压制高精度深度学习攻击 □依赖先验，开销大，稳定性弱
细粒度建模高精度攻击	扩大匿名集，降低流量可区分度	匿名集增强防御	相似流量聚类、生成高逼真诱饵	从基础聚类向精细模板与高逼真诱饵生成演进	□终端适配，降低识别置信度 □泛化能力弱，安全开销难兼顾
现有防御缺乏理论根基，难从根源阻断信息泄露	从协议层、信息论、网络架构等维度跨层协同优化	跨层协同优化防御	理论安全约束与协议原生机制，根源削弱目标可区分度	从理论安全分析向协议适配演进，升级为主动跨层协同的系统性防御	□理论安全性强，可从根源削弱侧信道泄露 □建模难，泛化能力弱

式，奠定了解耦流量特征与站点身份的核心防护目标，是后续三类技术的共同技术源头。其中，攻击模型误导继承其干扰特征可分性的核心逻辑，将作用对象从流量特征延伸至攻击模型决策环节，以主动欺骗思路应对高对抗场景下的深度学习攻击；匿名集增强继承其攻击无关的核心属性，将消除单流特征差异的思路拓展为扩增同构候选集合稀释识别唯一性，以更低流量改动成本适配轻量防护场景；跨层协同优化属于底层范式层级演进，继承全场景匿名防护的核心目标，深入协议底层从根源消解侧信道泄露，同时可为上层技术的安全边界标定、参数优化提供理论支撑。四类技术共同构成覆盖流量表征、模型决策、候选集合、协议底层的纵深防护体系，呈现出从单点特征修补到全链路体系化防御的发展趋势，功能上形成显著互补。整体而言，四类技术均受防护强度、系统开销、部署可行性构成的三元困境约束，其技术演进本质是多重约束条件下的持续寻优过程，不同技术仅权衡侧重不同，分别适配差异化场景需求，不存在适用于所有场景的普适最优方案。

3　Tor攻防评估框架与数据集

本节系统梳理网站指纹领域主流评估框架、关键评测指标与典型公开数据集，为攻防技术的标准化对比分析提供依据。

3.1　Tor场景评估框架与常用指标

3.1.1　Tor场景评估框架

Tor网络网站指纹评估体系持续迭代，逐步弥补传统评估与真实网络环境脱节的固有缺陷。早期工作依赖封闭世界假设与合成数据[30]，存在现实偏差。Juarez等[107]针对该评估前提指出了开放世界场景下基础率谬误问题。相关研究沿着四个维度展开：

一是提升评估环境的真实性。Cherubin等[38]部署出口节点探针构建真实流量数据集，确立了逼近现实威胁的评估基准。Deng等[108]构建涵盖防御机制、流量漂移、多标签浏览、早期检测、开放世界及少样本学习六重现实约束的多维评估矩阵，系统性纠正了传统评估高估攻击能力的偏差[109]。

二是构建理论与形式化评估的坚实基础。现有研究依托贝叶斯决策、信息论等工具，构建模型无关的可证明安全边界。其中WFES[110]与DeepSE-WF[111]分别从概率误差与高位特征空间角度量化隐私边界；Adaptive Tamaraw与Hydra等通过数学约束优化防御性能，兼顾理论严谨与工程实用[47,112]。

三是为支持上述评估工程实现与算法开发，系列支撑性平台逐步标准化。Experiment Tor[113]提供了全网仿真环境。Maybenot框架[114]标准化防御逻辑，支撑防御算法开发与迭代。在数据集层面，WFP-Collector[115]与GTT23[116]分别依托合成数据与脱敏真实流量构建，推进领域数据基础设施的标准化。

四是针对网络动态变化引发的概念漂移问题，Proteus框架[117]引入自适应更新机制，依托无标签流量完成模型迭代，表明未来的评估范式中需纳入对持续自适应博弈能力的考量[118-119]。

Tor网络网站指纹评估体系已完成从单一实验室测试到实测验证、理论量化、工程支撑、自适应博弈的多维体系演进，持续推动改领域研究向着更贴合实际的方向发展，各框架对比如表13所示。

3.1.2　Tor场景评估常用指标

在封闭世界假设下，总体攻击性能通常以准确率（Attack Success Rate,ASR）衡量[30]。

$$\mathrm{ASR} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{1}$$

其中，TP（true positive），TN（true negative），FP（false positive）与FN（false negative）分别表示真阳性、真阴性、假阳性与假阴性样本数。在开放世界场景的评估中[11,107]，通常采用更为精细的指标。其核心指标包括真阳性率（TPR,true positive rate，即召回率）与假阳性率（FPR,false positive rate）：

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \tag{2}$$

为综合权衡召回率与精确率$\mathrm{P} = \mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$，常采用F1分数[14]，其对类别不平衡的数据集更稳健，计算公式为：

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{P} \cdot \mathrm{TPR}}{\mathrm{P} + \mathrm{TPR}} \tag{3}$$

针对开放世界特性，*r - precision*，（recall-based precision）指标[120]通过固定召回率r来克服基础率的影响。其定义为：

表 13　　Tor 网络网站指纹防御评估框架

框架类别	目标贡献	关键特性	代表成果	评估指标
简化评估框架	在封闭世界假设下进行攻防原理的初步验证	基于合成数据的实验室评估	[30]	基础分类准确率、可行性验证
真实化评估框架	逼近真实威胁，提升评估实用性，纠正传统评估的偏差	在线真实世界监测、构建开放世界、概念漂移等专用数据集，多场景验证	[38,107-109]	准确率、F1、AUC、监控规模影响分析
理论与形式化评估框架	建立可证明安全边界，提供理论评估标准	基于贝叶斯错误、信息论、隐私参数、可证明安全设计，用于安全边界评估，提供防御设计范式	[47,110-112]	贝叶斯错误率、互信息、隐私参数、MAA
支撑性评估框架	支撑工程实现与数据标准化	全网络仿真、通用防御实现平台、标准化真实与合成数据采集、保障研究可复现	[113-116]	数据集质量、带宽开销、跟踪相似性、可复现性
自适应评估框架	应对概念漂移，适配动态威胁	特征分布对齐、自适应伪标签生成等，可用于概念漂移适应、无监督/半监督学习、动态威胁评估，推动评估范式向动态自适应演进	[117-119]	F1 分数、能量距离、消融实验性能、延迟

$$r\text{-}precision = \frac{\mathrm{TPR}}{\mathrm{TPR} + r \cdot \mathrm{FPR}} \tag{4}$$

其中 r 表示非监控网站与监控网站的数量比值。

防御强度可从实证与理论两个层面评估。防御成功率（Defense Success Rate, DSR）[11]是衡量方案实际效果的直观指标，计算公式为：

$$\mathrm{DSR} = 1 - \mathrm{ASR} \tag{5}$$

其表示在防御生效后，攻击模型能达到的分类准确率。

从信息论出发，互信息 $I(W;\mathrm{X})$ 可直接量化防御后观测特征 X 与原始网站标签 W 之间残余的统计依赖性，是衡量信息泄露度的指标[102]。

$$I(W;\mathrm{X}) = H(W) - H(W|\mathrm{X}) \tag{6}$$

其中，$H(W)$ 为网站标签 W 的熵，$H(W|\mathrm{X})$ 为给定 X 条件下 W 的条件熵。$I(W;\mathrm{X})$ 的值越低，表明信息泄露的越少，理论隐私保护越强。

性能开销直接关系到方案的实用性，带宽开销是核心度量指标[40]：

$$\text{带宽开销} = \frac{\mathrm{Size}_{\mathrm{def}} - \mathrm{Size}_{\mathrm{orig}}}{\mathrm{Size}_{\mathrm{orig}}} \times 100\% \tag{7}$$

其中 $\mathrm{Size}_{\mathrm{def}}$ 和 $\mathrm{Size}_{\mathrm{orig}}$ 分别为防御前后的 Tor 流量字节数。全面的开销评估必须超越单点测量，考虑由网络排队等效应引发的网络级延迟影响。

3.2　Tor 场景公开数据集

在网站指纹攻防研究中，高质量公开数据集是支撑领域迭代的基础。本小节系统梳理了 2008-2026 年 Tor 场景主流公开数据集。规模上，数据量从 2008 年的 8.5×10^3 条痕迹发展到 2023 年 GTT23 的 1.4×10^7 条痕迹，增长超过三个数量级[116]；性质上，由早期小规模封闭世界合成数据，逐步演进为大规模、多场景、高真实度的复杂数据集，催生出针对特定防御和新兴应用场景的专用数据集。

早期的数据集主要服务于封闭世界环境，主要应用于传统机器学习攻击的性能验证[30]，随着深度学习技术普及，AWF 等百万级大数据集成为模型评估的新标准[32]。近年来数据集呈现规模化与专用化双方向发展趋势：一方面，以 GTT23 为代表的数据集依托真实 Tor 出口节点被动采集，大幅提升流量真实性；另一方面，面向防御评测、多标签浏览、

概念漂移及暗网识别的细分专用数据集不断完善。此外，WFlib 等[121]开源工具集成多类流量样本，为实验复现与横向对比提供标准化数据支撑。

表 14 汇总了各类数据集的发布年份、规模、特点与贡献，完备的数据体系能够为研究者提供可靠实验支撑，保障评测过程中的严谨性与结论可信度。

4　防御技术综合评估与选型分析

本节旨在界定与分析多维评估体系，对主流防御方案开展综合性能对比。首先明确以防御效果、系统开销及部署特性为核心的三维评估框架；其次基于此框架对四类防御技术的代表性方案展开分

表 14　　Tor 网络网站指纹研究常用公开数据集

数据集名称	年份	监控集	开放世界集	核心特点与贡献
Herrmann et al.[30]	2008	775×10	NA[2]	首个 Tor WF 数据集，封闭世界攻击验证
Cai et al.[122]	2012	800×40	NA[2]	早期大规模封闭世界数据集，用于封闭世界攻击评估
Levdata2[123]	2013	100×40	NA[2]	中等规模封闭世界算法快速验证
Juarez et al.[107]	2014	200×40	35000×1	早期开放世界基准，支持基础率评估
Wang et al.[124]	2014	100×40	5000×1	明确区分封闭/开放世界的基础评估框架
RND-WWW[125]	2016	1125×40	120000×1	大规模开放世界攻击评估，提供了丰富的背景流量
DSTor[31]	2016	85×90	100000×1	针对 .onion 暗网站点指纹识别
AWF[32]	2017	95×2500	200×2000+400000×1	高实例数封闭+大规模开放，深度攻击模型综合评估
Walkie-Talkie[43]	2017	100×100	9900×1	含防御流量，支撑流量变形测试
DF[27]	2018	95×1000	40700×1	深度学习基准，支持攻击对比
WTT-time[126]	2018	100×300	50000×1	面向概念漂移，评估时间鲁棒性
Wang et al.(2019)[120]	2019	100×200	80000×1	含子页面数据，支持内部页识别
Good Enough[50]	2020	500×20	10000×1	填充防御专用，评估抗干扰性能
BigEnough[29]	2021	950×20	19000×1	结构化中等规模，支撑模型训练评估
TMWF[12]	2023	50×100	5000×1	去先验多标签，用于攻防模型评估
Multi-tab[25]	2023		570000×1	大规模多标签，支持混合流量攻防研究
Drift[127]	2023	90×100	90×20+5000×1	长时序采集，支持概念漂移与自适应研究
GTT23[116]	2023	NA[2]	14000000×1	基于大规模 Tor 真实流量，用于开放世界与背景流量评估
D1-D7[118]	2024	100×700	4000×1	依托排名网站构建，适用与结构化场景算法评测对比
CW/OW[128]	2024	1000×10	9300×1	区分双世界结构清晰，适配两类场景攻防精细评测
ALEXA-WSC-FG/BG[129]	2024	9000×90	45000×1	依托权威榜单构建，适配开放场景大规模模型训练评测
Defence Dataset[130]	2025	95×1000	NA[2]	集成混合防御流量，用于防御鲁棒性与抗防御评测
Reality Check[131]	2026	跨国监控流量	＞80000×1	异构网络实测，面向开放世界探究路径延迟影响

析，揭示不同技术路径的内在权衡。最后结合分析结论，针对不同 Tor 应用场景给出防御技术选型建议，为理论研究与工程部署提供参考。

4.1　评估指标界定与分析

为实现网站指纹防御技术的规范化横向对比，本节界定了兼具理论严谨性与工程适配性的标准

化评估体系，包含防御效果、系统开销与部署特性。

防御效果作为核心指标，需兼顾受控环境下的

识别抑制能力与开放场景中的综合防护效能；同时从环境鲁棒性与对抗鲁棒性层面，考量防御方案抵抗流量漂移、防范对抗重训练攻击的长期稳定性。

系统开销决定方案实用性与用户体验，主要涵盖三类指标：带宽开销以流量增幅占比进行量化；时延开销体现为防御机制带来的额外通信延迟，对 Tor 低时延传输场景影响显著；计算与内存开销则是轻量化终端及高负载中继部署的关键约束。

部署特性衡量防御技术嵌入 Tor 生态的落地能力，涵盖三个方面：部署条件主要包括部署位置与对 Tor 协议的修改幅度；策略隐蔽性要求防御行为不引入可识别的新侧信道特征；方案通用性是评估技术对先验知识的依赖度及对场景的适配能力。

4.2　基于三维评估体系的防御技术对比分析

本小节依托上述评估体系对各类防御方案开展对比分析。为保障公平性与科学性，性能数据均取自原始文献典型配置下的公开实验。防御效果以封闭世界中深度指模型攻击成功率降幅为主基准，辅以原文 kNN 等传统攻击模型与开放世界 TPR、

FPR 指标，并综合评估流量漂移与对抗重训练

鲁棒

性。结合原始文献的实验设定，本文以带宽增幅、页面加载时延增量量化系统开销，并从部署位置、协议修改幅度等维度定性评判部署特性。基于以上评测原则，本文筛选四大类别、十一个技术子类的典型防御方案，整理形成综合性能对比表（表15）。

经对各类防御范式开展横向对比分析。从防御

表15　　Tor网络典型网站指纹防御技术方案指标对比

技术大类	技术子类	代表方案	防御指标				轻量化	鲁棒性	抗概念漂移
			防御效果	带宽开销	延迟增量	部署位置			
流量特征混淆	静态填充	Tamaraw	闭世界（理想攻击）:100%→0.4%	199%～687%	320%±70%	客户端	□	★	□
		Walkie-Talkie	闭世界（kNN）:95%→28%	31%±16%	34%±5%	客户端	□	★★	□
		FRONT	开世界（DF）：TPR 96.40%→58.95%	33%～49%	0%	客户端	□	★★	□
	自适应动态	WTF-PAD	闭世界（kNN）:91%→20%	30%～60%	0%	客户端	□	★★	□
		RegulaTor	闭世界（DF）:98.4%→19.6%	+48.3%轻量模式	+8.9%轻量模式	客户端	□	★★	□
		RUDOLF	闭世界（DF）:98.4%→19.6%	10.83%～25.21%	0%	客户端	□	★★★	□
	多路径分散	TrafficSliver	闭世界（DF）应用层:98.75%→8.07%	0%	0%	客户端/中继	□	★★	□
攻击模型误导	对抗扰动	ALER	闭世界（DF）:97.05%→12.68% 开世界（DF）:TPR 91.80%→37.46%	约20%	0%	客户端	□	★★★	□
		WFGuard	闭世界（DF）:98.35%→14.18%	轻量版：11.04%	0%	客户端	□	★★	□
	生成伪装	WFD-EST	闭世界（FS-Net）:F1 95%→70%	8.1%～26.7%	8.1%～17.4%	客户端	□	★★	□
	数据中毒	TrapFlow	闭世界（DF）:98.8%→2.8%	轻量版：14.7%	20%～36%	出口节点	□	★★★	□
匿名集增强	集群匿名	Palette	闭世界（DF）:98.4%→36.43%	84%	9%	客户端	△	★★	□
	诱饵混淆	Realistic Cover Traffic	闭世界（kNN）:91%→14%	约20%	0%	客户端＋服务端	□	★	□
跨层协同优化	信息论优化	FRUGAL	闭世界（DF）:98.27%→2.68%	可调20%/30%/80%	约0%	客户端	□	★★★	□
	协议原生适配	Interspace	闭世界（DF）:88%→35%	可调，无膨胀	0%	客户端	△	★★	□
	轻量级专用	BRO	闭世界（DF）:72.2%→43.0%	约24.2%	0%	客户端/入口节点	□	★	□

轻量化设计中□表示明确支持轻量、低延迟架构，△表示仅部分优化开销，□为高开销无优化；鲁棒性中★★★表示对抗训练、模型迁移、多标签场景均强适配，★★表示单一攻击鲁棒性中等，★表示仅对早期攻击有效；抗概念漂移中□表示内置自适应机制，□表示静态规则无适配。

效能看，各类技术防护阈值差异源于底层作用机理：

跨层协同优化依托信息论约束达到最优防护上限，

FRUGAL 与 TrapFlow 分别通过信元扰动优化、训练集投毒机制，将攻击识别率控制在 3% 以内；攻击模型误导可有效抑制深度指纹攻击；流量混淆技术性能分化显著，传统静态填充代价高昂 RUDOLF、TrafficSliver 分别通过强化学习自适应、多路径拆分实现低损耗防护；匿名集增强仅能压缩样本区分度，防护增益有限。鲁棒性层面，信息论优化、对抗扰动与数据中毒类方案具备更强的场景适配能力，现有研究中针对流量概念漂移的自适应设计仍较为缺乏，动态环境适配能力整体偏弱。

从系统开销看，静态填充资源损耗最为突出 Tamaraw 带宽膨胀率最高可达 687%；自适应混淆、跨层优化及对抗防御的带宽开销普遍低于 20%；多路径分散实现零带宽开销。时延上，主流防御均规避主动人为延迟，仅早期方案存在固有滞后，但规模化部署的链路排队时延仍不可忽略。算力层面，智能对抗类方案算力门槛较高，轻量化混淆与路径拆分技术更适配异构受限终端。随着防御强度提升，资源开销呈加速增长趋势，实用化部署需在隐私保护与系统效率间进行合理权衡。

从部署特性来看，客户端部署无需修改 Tor 核心协议，兼容性最优、落地简便，为当前主流形态；节点侧部署防护性能优异，但受限于分布式架构，规模化推广难度较大；多路径协同方案现阶段仅适用于实验验证；协议原生优化安全增益突出，但审核迭代周期冗长。

综合三维评估，防御范式呈现清晰演化脉络：防御逻辑由被动特征遮蔽迈向主动模型对抗，设计范式由经验规则构造演进为理论约束求解，防护维

度由单层流量整形拓展至跨层协同优化。防护鲁棒性、通信开销与部署难易度构成不可调和的三元约束，尚无普适性防御方案。

4.3　Tor 多场景防御技术选型

本小节基于上述分析，提出面向 5 类典型场景的适配技术路径：

（1）匿名优先场景以抵御高级持续性威胁为目标，可容忍较高通信开销。其中 FRUGAL 等信息论优化方案安全边界完备、开销可调，适用性较强；可证明安全的静态填充方案 Tamaraw 等提供严格的理论安全基准，适用于高保密场景；高效对抗扰动方案如 Dolos 与 BLANKET 等可在可控开销下实现高强度防护。数据中毒防御虽能提供源头级干扰，但受限于出口节点部署要求，适用范围较窄。

（2）延迟敏感的通用浏览场景需维持 Tor 网络低延迟特性。纯客户端自适应混淆方案 WTF-PAD、RegulaTor 部署门槛低，可实现无额外主动延迟或低延迟动态调整，是工程实用的优先选项；轻量级对抗扰动方案 WFGuard、ALERT 及专用轻量专用方案 BRO 可作为重要补充。多路径分散技术虽理论上具备零开销优势，但受限于 Tor 协议栈修改要求，工程落地难度较大。

（3）终端资源严格受限的移动与物联网场景以极轻量化为核心，首选 BRO 等专用轻量混淆方案；计算资源充足时，可采用轻量版 WFGuard 以获得更优防护效果。

（4）追求长期鲁棒性的场景需重点考量抗对抗训练能力。ALERT 等对抗扰动方案、从信息源头最小化特征相关性的信息论优化方案，以及从网络层破坏特征完整性的多路径分散机制，均对自适应重训练攻击具备较强抵抗能力。

（5）工程部署层面，与 Tor 协议原生框架集成的方案兼容性最佳，成熟的纯客户端自适应防御方案如 WTF-PAD 部署简便，落地可行性最高；需服务端或网络侧部署的方案如部署在出口节点的 TrapFlow 等方案虽防护性能优异，但受限于特殊部署位置或硬件依赖，规模化推广难度较大。

5　核心挑战与未来研究方向

前文剖析了四类防御技术，尽管范式各异，但共同面临五类共性约束：特征层面，静态模板与动态流量持续失配；模型层面，确定性策略易被攻击者自适应破解；系统层面，安全强度、带宽开销与部署可行性难以兼得；场景层面，封闭世界假设与开放环境真实需求存在显著偏差；评估层面，合成数据与简化设定系统性高估防御效能。上述约束相互耦合——特征漂移加剧模型脆弱性，开销约束限制场景适配，评估失真则使前四类瓶颈的真实严重程度难以准确度量。本节将上述约束凝练为五个核心挑战，并据此推演未来研究方向。

5.1 Tor场景核心研究挑战

本小节遵循特征、模型、系统、场景与评估的层级逻辑，系统解构制约Tor网站指纹防御的五大核心瓶颈：其一，分析特征漂移对防御长效性的负面影响；其二，阐释持续对抗博弈诱发的模型鲁棒性局限；其三，讨论防御效能、系统开销、部署可行性的多目标权衡悖论；其四，剖析多标签浏览及开放世界环境对传统封闭世界假设的颠覆；其五，分析数据集与实验偏差带来的评估失真问题。五类挑战逐层递进，共同构成了对动态对抗环境下Tor长效隐私保护核心命题的系统性拆解。

5.1.1 特征概念漂移问题

特征概念漂移是指网站流量特征与其标签间的映射关系，因多种动态因素持续变化，导致基于历史数据训练的模型在实际动态环境中性能显著衰退[10]，Juarez等证实了这一现象[107]。Deng等开展的研究则通过270天真实时间漂移实验进一步证实，即使是最先进的鲁棒攻击模型，若未针对分布变化进行适配，其F1值也会随时间推移显著下降[117]。其本质是训练阶段与部署阶段的数据联合概率分布存在差异，这使模型学到的特征表示与决策边界在新分布下失效。此外，漂移具有多模态特性，不同网站、网站路径或时间阶段的漂移模式各异，单一策略难以实现全场景适配[132]。绝大多数现有防御方案依赖固定的流量混淆模板[10]，在动态环境中极易因模板与实际分布失配而失效。而现有应对策略，如定期重训练或数据增强，普遍受限于标记数据收集成本高、效率低，或难以适应复杂多变的真实漂移模式[133]。因此，如何在有限或无标注数据条件下实现对概念漂移的高效、实时适配，成为网站防御领域具有挑战性的难题。

5.1.2 对抗训练的鲁棒问题

对抗训练的鲁棒性问题表明，防御效能不仅取决于当前混淆效果，更依赖抵御攻击者自适应学习的能力。多数现有轻量级防御在对抗训练攻击下极为脆弱，其混淆规则易形成可被学习的新模式，攻击者仅需收集防御后流量重训练即可快速破解[18]。Shen等[24]发现，即便未知防御参数，RF攻击对WTF-PAD防御的准确率仍超96%，揭示传统防御时效性有限，易因攻击者自适应迅速失效，持续更新策略将导致部署成本激增。虽有研究尝试通过后门攻击[96]、异构策略组合[56]、动态自适应特征空间[51]或理论安全边界[103]提升鲁棒性，但现有评估多基于简化单轮对抗假设，与攻击者多轮迭代的真实威胁差距显著。Grini等[134]指出，大量对抗样本因违反数值与类别依赖规则缺乏物理可行性，导致鲁棒性评估被高估；攻击者仍可通过大规模数据收集与模型优化恢复判别信息。Deng等[117]在Proteus框架中尝试持续对抗训练，却受限于算力。因此，构建贴近真实威胁的持续对抗训练评估框架，是验证防御长期有效性的关键。

5.1.3 轻量级防御的多目标权衡问题

Tor网站指纹防御面临隐私强度、系统开销与部署可行性的固有权衡[107]。现有轻量级防御宣称的“零延迟”存在评估偏差，Witwer等[135]基于Tor全网仿真证实，纯填充机制在规模化部署时将加剧网络拥塞并引致额外时延。多目标权衡的复杂性体现为开销维度的非线性关联，如RegulaTor揭示的延迟与带宽近似反比关系，并通过TPE超参数优化在权衡曲面中锁定可行配置[51]。前沿研究聚焦于可证明安全，FRUGAL则通过互信息最小化抵御攻击者适应，却牺牲了对动态网络的适应能力[103]。这表明初始防护强度与长期开销效率间存在固有取舍。突破传统范式需另辟蹊径，如TrafficSliver采用多路径分发阻断攻击者获取完整视图，实现无额外延迟防御，但其有效性依赖路径的随机独立性，面对节点共谋攻击时存在局限[67]，进一步印证了单路径观测模型下强隐私与低开销难以兼得的约束。

5.1.4 多标签与开放世界场景适配问题

现实Tor应用中的多标签浏览与开放世界访问，虽然突破了传统防御的封闭世界假设，但也引入了流量混合、背景噪声及极端类别失衡等复合挑战[38]。多标签场景的核心在于流量解混，攻击目标由单一网站转向组合识别[25]；现有研究已提出多标签分类等攻击建模手段，但因需兼顾标签一致性与可控开销，尚无成熟防御方案[12]。开放世界场景[137]则面临假阳性失控难题，非监控流量占绝对多数，即使极低的假阳性率也会产生海量误报，严重削弱攻防实用性[120]。此外，流量拆分机制虽能增强抗分析能力，但低延迟调度特性也可能被攻击者利用[64]。而多标签与开放世界的叠加效应研究尚不充分。上述研究表明，多标签流量虽能增加分析难度，但保护效力高度依赖标签组合的随机

性，无法替代主动防御[130]。针对此类复杂场景，上下文感知的动态适配防御已成前沿趋势，但普遍受限于真实、细粒度评估数据的匮乏，严重制约了防御方案的验证与发展。

5.1.5　数据集质量与评估真实性问题

现有网站指纹攻防研究的可靠性因系统性评估偏差而受到质疑。这种偏差主要体现在两个方面，一是过于简化的数据集与合成数据集在电路长度等关键特征上与真实Tor流量存在显著差异[38,116]；二是评估框架本身存在位置偏差与时间偏差，即训练与测试位置的不一致会显著高估模型性能，而短期评估则无法反映概念漂移导致的防御效果随时间衰减的问题，这使得现有文献中报告的防御效果在动态的真实Tor网络中可能被严重高估[107,134]。威胁模型与现实场景脱节进一步加剧了这一困境，研究发现，即使在更现实的设置下，深度指纹攻击仍保持高精度与召回率，但其性能特征与实验室评估存在显著差异[131]。尽管采用真实流量、位置转换、长期监测等方法可提升评估真实性，但这些方法普遍面临隐私风险、成本增加与可行性之间的严格制约，构成了该领域在提升评估真实性时必须应对的方法论挑战[124]。

5.2　未来重点研究方向

针对5.1节揭示的Tor网站指纹防御在概念漂移、对抗鲁棒性、多目标权衡、场景适配及评估真实性等方面面临的核心瓶颈，本节结合第2章梳理的四类防御技术演进规律，聚焦五个与现存挑战高度适配的方向，对后续研究展开系统性分析与展望。

5.2.1　自适应抗漂移防御技术

针对概念漂移导致固定防御策略持续失效的问题，现有方案多局限于填充概率等表层参数的微调，底层逻辑固化，易被攻击者通过重训练实现策略逆向。为此，自适应抗漂移防御依托动态感知与环境响应，维持Tor指纹防护的长期有效性。当前主流技术路线分为三类：一是以Proteus[117]为代表的集成化方案，通过特征对齐与自适应伪标签提升防御稳定性；二是CellShift，结合RTT感知增强模型对动态环境的泛化能力[138]；三是TrapFlow，通过数据中毒，强化对抗防御[96]。此外，CertTA依托多模态随机平滑构建理论鲁棒边界，从可证明安全层面完善抗漂移防御体系[139]。

尽管如此，现有方案仍面临毫秒级实时响应能力不足、多源复合漂移覆盖度有限，以及状态感知、策略决策与效能验证的自治闭环尚未打通等共性瓶颈。基于上述瓶颈，该方向后续研究将围绕实时性、轻量性、长效鲁棒性三大维度推进。其中，Proteus与TrapFlow侧重提升对抗鲁棒性，CellShift侧重轻量化实现，但毫秒级在线自适应机制仍是当前亟待突破的核心难点。

5.2.2　生成式人工智能增强防御

针对确定性策略被重训练破解的困境，生成式AI增强防御以不可学习的伪装为核心突破方向。当前，生成式AI在Tor网站指纹防御中的应用尚处起步阶段。现有方案（如Traffic Morphing、WF-GAN、Surakav及WFD-EST）虽通过重构流量时序特征提升了防护效能，但其生成器一经训练即固化，输入与输出之间仍存在确定性映射，尚未从根本上消除可学习性，且在开放世界与多标签场景下泛化能力受限[14]。理论层面，Transformer架构的长时序依赖建模能力与预训练模型的跨场景知识迁移潜力，为突破上述局限提供了新的技术范式。但受限于数据匮乏、算力开销及黑盒不可解释性，基于大语言模型的防御探索尤为薄弱。

针对上述瓶颈，未来研究需沿三条路径突破：一是引入随机扰动，破除生成过程的确定性映射以阻断逆向拟合；二是构建端到端对抗框架，将攻击反馈嵌入生成损失，实现伪装与干扰的协同优化；三是探索离散策略搜索，利用大语言模型的组合泛化能力诱导策略跳变，规避攻击者的增量学习追踪。上述路径旨在实现生成式伪装的不可学习性，将有效完善攻击模型误导的技术谱系。

5.2.3　轻量级可证明安全防御

针对“安全强度-系统开销-部署可行性”三元困境，轻量级可证明安全防御致力于在形式化安全上界内实现开销最小化。以Tamaraw为代表的早期方案确立了分析框架，Adaptive Tamaraw通过聚类与差异化参数分配，在维持安全上界的同时显著降低开销[47]，验证了理论与工程相结合的可行性。

针对现有方案静态安全边界难以适配动态网络、复杂场景缺乏完备证明、计算成本难以支撑终端部署等瓶颈，未来研究需重点推动形式化安全体系与自适应机制的深度耦合。一方面，面向多标签、持续对抗等真实场景完善威胁模型，构建普适

的安全证明框架；另一方面，设计低复杂度的专用算法，适配移动及物联网等资源受限终端。上述探索旨在推动具备严格理论保障的防御范式从理论模型向工程实践跨越。

5.2.4 复杂场景通用防御

针对多标签并发与开放世界场景对封闭世界假设的冲击，复杂场景通用防御需整合四大类技术方案以构建跨场景防护框架。现有防御多基于单标签封闭世界范式设计，难以适配多标签并发、开放世界访问及Conflux流量拆分等复杂场景[133-134]。

针对上述瓶颈，后续研究需整合四类防御技术核心组件：一是研发低开销并发流量联合混淆算法，融合流量混淆与对抗干扰技术，破解多标签解混与标签一致性维护难题；二是构建动态背景流量生成机制，依托集群匿名化思路构造高保真噪声，缓解开放世界类别失衡压力；三是实现防护逻辑与传输调度深度耦合，平衡路径随机化与跨路径流量协同混淆需求。上述探索旨在构建适配多元异构环境的通用防御范式，推动防御体系从单场景专用向跨场景通用演进。

5.2.5 Tor原生可集成防御

针对评估失真问题与部署可行性困境，Tor原生可集成防御通过真实环境持续验证与分层渐进式部署来同步化解双重瓶颈。现有防御方案普遍面临评估失真与部署壁垒的双重制约：实验环境多依赖合成数据集与简化假设，易致防护效能被高估；协议层改造则需通过严苛的安全审计并兼容异构存量节点，学术原型难以规模化落地。尽管跨层协同优化研究已证实将防护逻辑嵌入协议栈可兼顾开销控制与长效稳定[76]，但现有集成思路仍缺少标准化实测验证手段，Retracer、CellShift 等评测框架虽搭建了真实流量测试环境[116,137]，却未形成完整的审计与验证流程，难以支撑新型防御全维度校验。

该方向后续可采用分层渐进式集成思路推进落地。短期优先考虑在无需改动内核的可插拔传输层部署自适应填充等成熟防御机制，降低改造成本；中期可结合多路径传输协议开展协同设计，补齐多流场景下的防护短板；长期则需将成熟防护逻辑纳入 Tor 原生协议规范，内置自适应调整能力。整套演进路径打通离线仿真、真实流量实测与协议原生改造全流程，同步解决实验失真、协议兼容困难等问题，可持续提升 Tor 整体匿名防护水平。

6 总结

Tor网络作为关键匿名通信基础设施，其隐私价值凸显，但网站指纹识别技术持续威胁其匿名性根基。本文系统综述了面向Tor网络的网站指纹防御技术，其主要贡献如下：

首先，本文从Tor网络网站指纹攻防博弈演进的角度出发，构建了一个体系化、可扩展的防御技术分类框架，依据Tor网络的多层级脆弱性与防御机理，将领域内技术系统地划分为流量特征混淆、攻击模型误导、匿名集增强及跨层协同优化四大类，再依据技术实现路径进一步细分，在各大类下划分若干子类。框架以防御目标为牵引，完整覆盖了从经典静态填充到前沿生成式伪装的全维度方案，清晰揭示了不同技术的内在原理与区别、演进关系、适用场景以及根本局限等，为新兴技术提供了具备良好扩展性的归类基础。

其次，综合防御效果、系统开销与部署特性三个维度，剖析了该领域的技术演进规律。研究发现，技术发展呈现两条核心主线。一是流量特征的改造与增强：路径从固定规则的静态填充，演进为感知网络状态的动态自适应混淆，并进一步探索多路径传输等架构特性进行流量分散；匿名集增强技术可视作此主线的分支，通过用户行为混合扩大匿名集以间接“改造”特征。二是攻击模型的干扰与破坏：技术纵深从流量层的对抗性扰动，发展到利用生成式人工智能合成高仿真流量进行深度伪装，再前置到通过数据中毒干扰训练阶段，实现了从“被动隐藏”向“主动攻击”的思想转变。跨层协同与信息论优化则为上述主线提供理论支撑，指导参数优化或从协议底层设计轻量级混淆。本质上，两条主线技术均在“防御效果、系统开销、部署可行性”构成的不可能三角中寻求动态平衡：流量混淆型部署简易但面临概念漂移挑战；攻击误导型效果显著但需应对对抗训练升级；匿名集增强型思路独特但存在场景适配与元数据泄露风险；跨层协同型理论先进但通常以高计算复杂度与低部署可行性为代价。

最后，本文提炼了当前核心挑战并指明了未来趋势。现有研究仍面临着概念漂移引发的防御衰减、对抗训练导致的自适应破解、资源约束下的多目标权衡困境、复杂真实场景适配性不足以及评估框架失真等共性难题。随着Tor网络网站指纹攻防

持续演进，防御技术正朝着动态自适应、理论可证明与智能深度协同的方向发展，将重点围绕在线自适应防御、基于生成式人工智能的智能深度伪装、轻量级可证明安全机制构建、通用化防御框架设计以及协议原生平滑部署开展研究。这标志着该领域正从静态规则应用，迈向动态、鲁棒、智能、理论严谨与工程可落地的新发展阶段。

参考文献：

[1] SONG B, POKHREL S R, DENG M, et al. Digital privacy under attack: Challenges and enablers[J/OL]. ACM Computing Surveys, 2026, 58(4): 1-35.

[2] 隋嘉祺, 扈红超, 史鑫, 等. 基于相似性感知的Tor网络路径选择算法 [J]. 计算机科学, 2025, 52(3): 392-399.
SUI J Q, HU H C, SHI X, et al. Tor network path selection algorithm based on similarity perception[J]. Computer Science, 2025, 52(3): 392-399.

[3] DINGLEDINE R, MATHEWSON N, SYVERSON P. Tor: The second-generation onion router:[R/OL]. Fort Belvoir, VA: Defense Technical Information Center, 2004.

[4] TAN Q, WANG X, SHI W, et al. An anonymity vulnerability in tor[J/OL]. IEEE/ACM Transactions on Networking, 2022, 30(6): 2574-2587.

[5] MURDOCH S J, DANEZIS G. Low-cost traffic analysis of tor[C/OL]// 2005 IEEE Symposium on Security and Privacy (S&P'05. Oakland, CA, USA: IEEE, 2005: 183-195.

[6] CUI Y, WANG G, VU K, et al. A comprehensive survey of website fingerprinting attacks and defenses in tor: Advances and open challenges [A/OL]. arXiv, 2025.

[7] JOHNSON A, WACEK C, JANSEN R, 等 . Users get routed: Traffic correlation on tor by realistic adversaries[C/OL]//Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security - CCS '13. Berlin, Germany: ACM Press, 2013: 337-348.

[8] 梅汉涛，程光，朱怡霖，等 . Tor 被动流量分析综述 [J]. 软件学报，2025, 36 (1): 253-288.
MEI H T, CHENG G, ZHU Y L, et al. Survey on Tor passive traffic analysis[J]. Journal of Software, 2025, 36(1): 253-288.

[9] KARUNANAYAKE I, AHMED N, MALANEY R, et al. De-anonymisation attacks on tor: A survey[J/OL]. IEEE Communications Surveys & Tutorials, 2021, 23(4): 2324-2350.

[10] 杨宏宇，宋成瑜，王朋，等。洋葱路由器网站指纹攻击与防御研究综述 [J]. 电子与信息学报，2024, 46 (9): 3474-3489.
YANG H Y, SONG C Y, WANG P, et al. Website fingerprinting attacks and defenses on Tor: A survey[J]. Journal of Electronics & Information Technology, 2024, 46(9): 3474-3489.

[11] WANG T, CAI X, NITHYANAND R, et al. Effective attacks and provable defenses for website fingerprinting[C]//Proceedings of the 23rd USENIX Security Symposium, San Diego, USA, 2014: 143-157.

[12] JIN Z X, LU T B, LUO S, et al. Transformer-based Model for Multi-tab Website Fingerprinting Attack[C]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery, 2023: 1050-1064.

[13] 蒋首志，曹金璇，殷浩展，等。基于 MHA 与 SDAE 的 Tor 网站指纹识别模型 [J]. 信息网络安全，2022,22 (10):8-14.
JIANG S Z, CAO J X, YIN H Z, et al. Identifying Tor Website Fingerprinting Model Based on MHA and SDAE[J]. Netinfo Security, 2022, 22(10): 8-14.

[14] ZHANG D F, RAO C, HUANG J N, et al. A dynamic Website Fingerprinting Defense by Emulating Spatio-Temporal Traffic features[J/OL]. Electronics, 2025, 14(22): 4441.

[15] SHEN M, JI K X, WU J H, et al. Real-time Website Fingerprinting Defense via Traffic Cluster Anonymization[C/OL]//2024 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2024: 3238-3256

[16] JIANG M, CUI B, FU J, et al. RUDOLF: An efficient and adaptive defense approach against website fingerprinting attacks based on soft actor-critic algorithm[J/OL]. IEEE Transactions on Information Forensics and Security, 2024, 19: 7794-7809.

[17] 黄斌，杜彦辉。基于流量分割与填充的网站指纹防御技术 [J]. 北京理工大学学报，2024,44 (7):750-760.
HUANG B, DU Y H. An Effective Website Fingerprinting Defense Based on Traffic Splitting and Padding [J]. Transactions of Beijing Institute of Technology, 2024, 44 (7): 750-760.

[18] 邹鸿程，苏金树，魏子令,等。网站指纹识别与防御研究综述 [J]. 计算机学报，2022, 45 (10): 2243-2278.
ZOU H C, SU J S, WEI Z l ,et al. A Review of the Research of Website Fingerprinting Identification and Defense [J]. Chinese Journal of Computers, 2022, 45 (10): 2243-2278.

[19] LIU P, HE L T, LI Z J. A survey on deep learning for website fingerprinting attacks and defenses[J/OL]. IEEE Access, 2023, 11: 26033-26047.

[20] 蔡满春，席荣康，朱懿，等。一种 Tor 网站多网页多标签指纹识别方法 [J]. 信息网络安全，2024, 24 (7): 1088-1097.
CAI M C, XI R K, ZHU Y, et al. A Fingerprint Identification Method of Multi-Page and Multi-Tag Targeting Tor Website[J]. Netinfo Security, 2024, 24(7): 1088-1097.

[21] 谢翛然，蔡满春，王珑皓，等。边界样本优先重放的 Tor 网站指纹增量识别方法 [J]. 计算机应用研究，2025, 42 (9): 1-10.
XIE X R, CAI M C, WANG L H, et al. Incremental Tor website fingerprint recognition method with boundary sample priority replay[J]. Application Research of Computers, 2025, 42(9): 1-10.

[22] 朱俞翡，胡宇翔，陈博，等。基于 Tor 的匿名网络技术研究综述 [J]. 通信学报，2025, 46 (5): 218-236.
ZHU Y F, HU Y X, CHEN B, et al. Survey of research on Tor-based anonymous networking technologies[J]. Journal on Communications, 2025, 46(5): 218-236.

[23] GONG J J, ZHANG W Q, ZHANG C, et al. WFDefProxy: Real World Implementation and Evaluation of Website Fingerprinting Defenses[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 1357-1371.

[24] SHEN M, JI K X, GAO Z B, et al. Subverting Website Fingerprinting Defenses with Robust Traffic Representation[C]//Proceedings of the 32nd USENIX Security Symposium. Anaheim: USENIX Association, 2023: 607-624.

[25] DENG X H, YIN Q L, LIU Z T, et al. Robust Multi-tab Website Fingerprinting Attacks in the Wild[C]//2023 IEEE Symposium on Security and Privacy (SP). San Francisco: IEEE, 2023: 1005-1022.

[26] RAHMAN M S, IMANI M, MATHEWS N, et al. Mockingbird: Defending Against Deep-Learning-Based Website Fingerprinting Attacks With Adversarial Traces[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 1594-1609.

[27] SIRINAM P, IMANI M, JUAREZ M, et al. Deep Fingerprinting: Undermining Website Fingerprinting Defenses with Deep Learning[C]// Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. Toronto: ACM, 2018: 1928-1943.

[28] MATHEWS N, HOLLAND J K, HOPPER N, et al. LASERBEAK: Evolving Website Fingerprinting Attacks With Attention and Multi-Channel Feature Representation[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 9285-9300.

[29] MATHEWS N, HOLLAND J K, OH S E, et al. SoK: A critical evaluation of efficient website fingerprinting defenses[C]//Proceedings of the 2023 IEEE Symposium on Security and Privacy (SP). San Francisco: IEEE, 2023: 154-169.

[30] HERRMANN D, WENDOLSKY R, FEDERRATH H. Website fingerprinting: attacking popular privacy enhancing technologies with the multinomial naïve-bayes classifier[C]//Proceedings of the 2009 ACM Workshop on Cloud Computing Security. Chicago: ACM, 2009: 31-42.

[31] HAYES J, DANEZIS G. k-fingerprinting: a robust scalable website fingerprinting technique[C]//Proceedings of the 25th USENIX Security Symposium. Austin: USENIX Association, 2016: 1-18.

[32] RIMMER V, PREUVENEERS D, JUAREZ M, et al. Automated website fingerprinting through deep learning[C]//Proceedings of the 25th Annual Network and Distributed System Security Symposium. San Diego: The Internet Society, 2018.

[33] SIRINAM P, MATHEWS N, RAHMAN M S, et al. Triplet fingerprinting: more practical and portable website fingerprinting with N-shot learning[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London: ACM Press, 2019: 1-18.

[34] ZHOU Q, WANG L M, ZHU H J, et al. Few-shot website fingerprinting attack with cluster adaptation[J]. Computer Networks, 2023, 229: 109780.

[35] ZHOU Q, WANG L M, ZHU H J, et al. WF-Transformer: learning temporal features for accurate anonymous traffic identification by using transformer networks[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 30-43.

[36] YIN Q L, LIU Z T, LI Q, et al. An Automated Multi-Tab Website Fingerprinting Attack[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(6): 3656-3670.

[37] BHAT S, LU D, KWON A, et al. Var-CNN: a data-efficient website fingerprinting attack based on deep learning[J]. Proceedings on Privacy Enhancing Technologies, 2019, 2019(4): 292-310.

[38] CHERUBIN G, JANSEN R, TRONCOSO C. Online Website Fingerprinting: Evaluating Website Fingerprinting Attacks on Tor in the Real World[C]//Proceedings of the 31st USENIX Security Symposium. Boston: USENIX Association, 2022: 753-770.

[39] RAHMAN M S, SIRINAM P, MATHEWS N, et al. Tik-Tok: the utility of packet timing in website fingerprinting attacks[J]. Proceedings on Privacy Enhancing Technologies, 2020, 2020(3): 5-24.

[40] CAI X, NITHYANAND R, WANG T, et al. A Systematic Approach to Developing and Evaluating Website Fingerprinting Defenses[C]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale: ACM Press, 2014: 227-238.

[41] DYER K P, COULL S E, RISTENPART T, et al. Peek-a-Boo, I Still See You: Why Efficient Traffic Analysis Countermeasures Fail[C]//Proceedings of 2012 IEEE Symposium on Security and Privacy. San Francisco: IEEE Press, 2012: 332-346.

[42] JUAREZ M, Imani M, PERRY M, et al. Toward an Efficient Website Fingerprinting Defense[C]//Computer Security – ESORICS 2016. Heraklion: Springer, 2016: 374-391.

[43] WANG T, GOLDBERG I. Walkie-Talkie: An Efficient Defense Against Passive Website Fingerprinting Attacks[C]//Proceedings of 26th USENIX Security Symposium. Vancouver: USENIX Association, 2017: 1375-1390.

[44] AL-NAAMI K, EL-GHAMRY A, ISLAM M S, et al. BiMorphing: A bi-directional bursting defense against website fingerprinting attacks [J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18 (2): 505-517.

[45] GONG J J, ZHANG W Q, ZHANG C, et al. Surakav: Generating Realistic Traces for a Strong Website Fingerprinting Defense[C]//2022 IEEE Symposium on Security and Privacy. San Francisco: IEEE, 2022: 188-205. DOI:10.1109/SP46214.2022.9833722.

[46] SINGH K P, PILLI E S, LAXMI V, et al. ADEFTOR: Adaptive Adversarial Example Generation for Website Fingerprinting Defense in Tor [J]. Journal on Communications, 2025, 46(8):12-25.

[47] KHAJAVI K, WANG T. Lightening the Load: A Cluster-Based Framework for A Lower-Overhead, Provable Website Fingerprinting Defense [C]//Proceedings of the 2026 Network and Distributed System Se- curity Symposium (NDSS). San Diego: Internet Society, 2026.

[48] ATHANASIOU A, CHATZIKOKOLAKIS K, PALAMIDESSI C. Self-defense: Optimal QIF solutions and application to website fingerprinting[A/OL]. arXiv, 2024

[49] LIU L, LIU X, ZHAO Y, HU N. TLD-WF: Surfing freely under website fingerprinting attacks[C]//Proceedings of the 2022 7th IEEE International Conference on Data Science in Cyberspace. Piscataway: IEEE, 2022.

[50] PULLS T. Towards Effective and Efficient Padding Machines for Tor: The good, the bad, and the ugly[EB/OL]. 2020-11-30.

[51] HOLLAND J K, HOPPER N. RegulaTor: A Straightforward Website Fingerprinting Defense[J]. Proceedings on Privacy Enhancing Technologies, 2022, 2022(2): 344-362.

[52] LIBERATORE M, LEVINE B N. Inferring the Source of Encrypted HTTP Connections[C]//Proceedings of the 13th ACM Conference on Computer and Communications Security, Alexandria, Virginia, USA, 2006. New York: ACM, 2006: 255-263.

[53] WRIGHT C V, COULL S E, MONROSE F. Traffic Morphing: An Efficient Defense Against Statistical Traffic Analysis[C]//Proceedings of the 16th Annual Network and Distributed System Security Symposium, San Diego, California, USA, 2009. San Diego: NDSS Symposium, 2009: 237-250.

[54] CAI X, NITHYANAND R, JOHNSON R. CS-BuFLO: A Congestion Sensitive Website Fingerprinting Defense[C]//Proceedings of the 13th Workshop on Privacy in the Electronic Society, Scottsdale, Arizona, USA, 2014. New York: ACM, 2014.

[55] NITHYANAND R, CAI X, JOHNSON R. Glove: A Bespoke Website Fingerprinting Defense[C]//Proceedings of the 13th Workshop on Privacy in the Electronic Society, Scottsdale, Arizona, USA, 2014. New

York: ACM, 2014.

[56] GONG J, WANG T. Zero-delay Lightweight Defenses against Website Fingerprinting[C]//Proceedings of the 29th USENIX Security Symposium, Virtual Event, 2020. Berkeley: USENIX Association, 2020: 717-734.

[57] CHERUBIN G, HAYES J, JUAREZ M. Website fingerprinting defenses at the application layer[J]. Proceedings on Privacy Enhancing Technologies, 2017, 2017(2): 186-203.

[58] LU D, BHAT S, KWON A, DEVADAS S. DynaFlow: An efficient website fingerprinting defense based on dynamically-adjusting flows [C]//Proceedings of the 2018 Workshop on Privacy in the Electronic Society (WPES'18). Toronto, Canada: ACM, 2018: 1-5.

[59] LIANG J Y, YU C S, SUH K W, et al. Tail time defense against website fingerprinting attacks[J]. IEEE Access, 2022, 10: 18516-18525.

[60] HONG X S, MA X K, LI S Y, et al. A Website Fingerprint defense tech- nology with low delay and controllable bandwidth[J]. Computer Communications, 2022, 193: 332-345.

[61] LI W H, ZHANG X Y, BAO H F, et al. Prism: Real-Time Privacy Protection Against Temporal Network Traffic Analyzers[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 2524-2537.

[62] YANG Z Y, XIAO X, ZHANG B, et al. AAP: Defending Against Website Fingerprinting Through Burst Obfuscation[C]//Proceedings of the 28th European Symposium on Research in Computer Security (ESORICS 2023). Cham: Springer, 2023: 145-164.

[63] XIONG H H, MAN D P, WANG H R, et al. PEZD: A practical and effective zero-delay defense against website fingerprinting[J]. Computers & Security, 2024, 139: 103428.

[64] ALSABAH M, BAUER K, ELAHI T, et al. The path less travelled: overcoming Tor's bottlenecks with traffic splitting[C]//Proceedings of the 13th International Symposium on Privacy Enhancing Technologies, Bloomington, 2013: 143-163.

[65] DE LA CADENA W, MITSEVA A, PENNEKAMP J, et al. POSTER: traffic splitting to counter website fingerprinting[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, London, 2019. New York: ACM, 2019.

[66] HENRI S, GARCIA-AVILES G, SERRANO P, et al. Protecting against website fingerprinting with multihoming[J]. Proceedings on Privacy Enhancing Technologies, 2020, 2020(2): 89-110. DOI:10.2478/popets-2020-0019.

[67] DE LA CADENA W, MITSEVA A, HILLER J, et al. TrafficSliver: fighting website fingerprinting attacks with traffic splitting[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, 2020. New York: ACM, 2020.

[68] REUTER S, HILLER J, PENNEKAMP J, et al. Demo: traffic splitting for Tor — a defense against fingerprinting attacks[C]//Electronic Communications of the EASST: Proceedings of Conference on Networked Systems 2021, 2021, 80: 1-4.

[69] WANG M, KULSHRESTHA A, WANG L, et al. Leveraging strategic connection migration-powered traffic splitting for privacy[J]. Proceedings on Privacy Enhancing Technologies, 2022, 2022(3): 498-515.

[70] SEO M J, YOU M S, KIM J H, et al. MUFFLER: Secure tor traffic obfuscation with dynamic connection shuffling and splitting[C/OL]// IEEE INFOCOM 2025 - IEEE Conference on Computer Communications. London, United Kingdom: IEEE, 2025: 1-10.

[71] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. arXiv:1412.6572, 2014.

[72] IMANI M, RAHMAN M S, MATHEWS N, et al. Adv-DWF: Defending against deep-learning-based website fingerprinting attacks with adversarial traces[EB/OL]. 2019-02-18

[73] NASR M, BAHROMALI A, HOUMANSDAR A. Defeating DNN-based traffic analysis systems in real-time with blind adversarial perturbations[C]// Proceedings of the 30th USENIX Security Symposium. Berkeley: USENIX, 2021: 2705-2722.

[74] HOU C S, SHI J Z, CUI M X, et al. Universal website fingerprinting defense based on adversarial examples[C/OL]//2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom). Shenyang, China: IEEE, 2021: 99-106.

[75] SADEGZADEH A M, SHIRAVI S, JALILI R. Adversarial network traffic: Towards evaluating the robustness of deep-learning-based network traffic classification[J/OL]. IEEE Transactions on Network and Service Management, 2021, 18(2): 1962-1976.

[76] SHAN S, BHAGOJI A N, ZHENG H T, ZHAO B Y. A real-time defense against website fingerprinting attacks[EB/OL]. 2021-02-08.

[77] LI D, ZHU Y F, CHEN M H, WANG J. Minipatch: Undermining DNN-based website fingerprinting with adversarial patches[J]. IEEE Transactions on Information Forensics and Security, 2022, 17: 2437-2451.

[78] TANG R, SHEN G, GUO C,et al.SAD: Website fingerprinting defense based on adversarial examples[J/OL]. Security and Communication Networks, 2022, 2022: 1-12.

[79] SADEGZADEH A M, TAJALI B, JALILI R. AWA: Adversarial Website Adaptation[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 3109-3122.

[80] QIAO L T, WU B, YIN S J, et al. Resisting DNN-Based Website Fingerprinting Attacks Enhanced by Adversarial Training[J]. IEEE Transactions on Information Forensics and Security, 2023, 18: 5375-5386.

[81] QIAO L T, GAO C Y, WU B, et al. Trace-agnostic and adversarial training-resilient website fingerprinting defense[C/OL]//IEEE INFOCOM 2024 - IEEE Conference on Computer Communications. Vancouver, BC, Canada: IEEE, 2024: 211-220.

[82] ABUSNAINA A, JANG R, KHORMALI A, et al. DFD: Adversarial Learning-based Approach to Defend Against Website Fingerprinting [C]//IEEE INFOCOM 2020 - IEEE Conference on Computer Communications. Toronto, Canada: IEEE, 2020: 2459-2468.

[83] JIANG M Y, CUI B J, FU J S, et al. KimeraPAD: A novel low-overhead real-time defense against website fingerprinting attacks based on deep reinforcement learning[J/OL]. IEEE Transactions on Network and Service Management, 2024, 21(3): 2944-2961.

[84] LING Z, XIAO G, LUO L, et al. WFGuard: An effective fuzzing-testing-based traffic morphing defense against website fingerprinting [C/OL]//IEEE INFOCOM 2024 - IEEE Conference on Computer Communications. Vancouver, BC, Canada: IEEE, 2024: 441-450.

[85] XIE R T, XIE K, ZHAO P C, et al. GAPDiS: Gradient-assisted perturbation design via sequence editing for website fingerprinting defense [C/OL]//Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. Taipei Taiwan: ACM, 2025: 3012-3026.

[86] HUANG J N, LIU W W, LIU G J, et al. WF-A2D: Enhancing privacy with asymmetric adversarial defense against website fingerprinting[J/OL]. IEEE Transactions on Information Forensics and Security, 2025,

20: 4739-4754.

[87] HOU C S, GOU G P, SHI J Z, et al. WF-GAN: Fighting back against website fingerprinting attack using adversarial learning[C/OL]//2020 IEEE Symposium on Computers and Communications (ISCC). Rennes, France: IEEE, 2020: 1-7.

[88] ZHANG Z W, YE D P. Defending against deep-learning-based flow correlation attacks with adversarial examples[J/OL]. Security and Communication Networks, 2022, 2022: 1-11.

[89] Holland J K, Carpenter J, Oh S E, et al. DeTorrent: An adversarial padding-only traffic analysis defense[J/OL]. Proceedings on Privacy Enhancing Technologies, 2024, 2024(1): 98-115.

[90] JIN M, APOSTOLAKI M. PANTS: Practical adversarial network traffic samples against ML-powered networking classifiers[EB/OL]. (2024-09-07)

[91] HUANG J N, LIU W W, LIU G J, et al. STAP: Leveraging state-transition adversarial perturbations for asymmetric website fingerprinting defenses[J/OL]. IEEE Transactions on Network and Service Management, 2025, 22(6): 6200-6214.

[92] HUANG S K, XIAO J C, GOU G P, et al. TMGAN: A GAN-based traffic morphing defense against website fingerprinting[C/OL]//2024 IEEE International Conference on High Performance Computing and Communications (HPCC). Wuhan, China: IEEE, 2024: 1273-1281.

[93] NING R, XIN C S, WU H Y. TrojanFlow: A neural backdoor attack to deep learning-based network traffic classifiers[C/OL]//IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. London, United Kingdom: IEEE, 2022: 1429-1438.

[94] SEVERI G, BOBOILA S, OPREA A, et al. Poisoning network flow classifiers[A/OL]. arXiv, 2023.

[95] LIANG S Y, GONG J J, FANG T M, et al. Red pill and blue pill: controllable website fingerprinting defense via dynamic backdoor learning [J/OL]. (2024-12-22).

[96] LIANG S Y, GONG J J, FANG T M, et al. TrapFlow: Controllable website fingerprinting defense via dynamic backdoor learning[J/OL]. IEEE Transactions on Information Forensics and Security, 2026, 21: 2610-2625.

[97] NIE L H, DONG X D, SHI L L, et al. Stinger: A light-weight website fingerprinting defense through poisoning packet sequences[J/OL]. IEEE Transactions on Information Forensics and Security, 2026, 21: 560-577.

[98] CHAN-TIN E, KIM T, KIM J. Website fingerprinting attack mitigation using traffic morphing[C/OL]//2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). Vienna: IEEE, 2018: 1575-1578.

[99] WANG Z H, LI Q, XIE G R, et al. Minos: a lightweight and dynamic defense against traffic analysis in programmable data planes[C]//Proceedings of the 2025 USENIX Annual Technical Conference (ATC). Berkeley: USENIX Association, 2025: 1-18.

[100] PANCHENKO A, NIESSEN L, ZINNEN A, et al. Website fingerprinting in onion routing based anonymization networks[C/OL]//Proceedings of the 10th annual ACM workshop on Privacy in the electronic society. Chicago Illinois USA: ACM, 2011: 103-114.

[101] CUI W Q, YU J M, GONG Y M, et al. Realistic cover traffic to mitigate website fingerprinting attacks[C/OL]//2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS). Vienna: IEEE, 2018: 1579-1584.

[102] LI S, GUO H J, HOPPER N. Measuring information leakage in web site fingerprinting attacks and defenses[A/OL]. arXiv, 2019.

[103] WANG R, LING Z, LIU G C, et al. Cease at the ultimate goodness: towards efficient website fingerprinting defense via iterative mutual information minimization[C]//Proceedings of the 2026 Network and Distributed System Security Symposium. San Diego: Internet Society, 2026: 1-16.

[104] DEGABRIELE J P, MELLONI A, MÜNCH J P, et al. Counter Galois Onion: Fast Non-Malleable Onion Encryption for Tor[C]//Advances in Cryptology – EUROCRYPT 2024: 43rd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Zurich, May 26 – 30, 2024. Proceedings, PartIII. Cham: Springer, 2024: 327-357.

[105] MCGUAN C, YU C, SUH K. Practical and lightweight defense against website fingerprinting[J/OL]. Computer Communications, 2024, 228: 107976.

[106] LING Z, XIAO G, WU W J, et al. Towards an efficient defense against deep learning based website fingerprinting[C/OL]//IEEE INFOCOM 2022 - IEEE Conference on Computer Communications. London, United Kingdom: IEEE, 2022: 310-319.

[107] JUAREZ M, AFROZ S, ACAR G, et al. A critical evaluation of website fingerprinting attacks[C/OL]//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. Scottsdale Arizona USA: ACM, 2014: 263-274.

[108] DENG X X, CHEN J Y, YU L X, et al. Beyond a single perspective: Towards a realistic evaluation of website fingerprinting attacks[A/OL]. arXiv, 2025.

[109] WANG T. The one-page setting: A higher standard for evaluating website fingerprinting defenses[C/OL]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. Virtual Event Republic of Korea: ACM, 2021: 2794-2806.

[110] CHERUBIN G. Bayes, not naïve: Security bounds on website fingerprinting defenses[J/OL]. Proceedings on Privacy Enhancing Technologies, 2017, 2017(4): 215-231.

[111] VEICHT A, RENGGLI C, BARRADAS D. DeepSE-WF: Unified security estimation for website fingerprinting defenses[A/OL]. arXiv, 2022.

[112] DI A. Hydra: An efficient, provably secure website fingerprinting defense based on traffic splitting[R]. Summer STEM Institute, 2020.

[113] BAUER K, McCOY D, SHERR M, et al. ExperimenTor: A testbed for safe and realistic Tor experimentation[C]//Proceedings of the 4th USENIX Workshop on Cyber Security Experimentation and Test. Berkeley: USENIX Association, 2011: 1-8.

[114] PULLS T, WITWER E. Maybenot: A framework for traffic analysis defenses[A/OL]. arXiv, 2024.

[115] AMINUDDIN M A I M, ZAABA Z F. WFP-Collector: Automated dataset collection framework for website fingerprinting evaluations on Tor Browser[J]. Journal of King Saud University - Computer and Information Sciences, 2023, 35(9): 101778.

[116] JANSEN R, WAILS R, JOHNSON A. A measurement of genuine Tor traces for realistic website fingerprinting[C]//Proceedings of the 27th International Conference on Passive and Active Measurement. Cham: Springer, 2026: 277-293.

[117] DENG X H, ZHANG Y X, LI Q, et al. Enhancing Website Fingerprinting Attacks against Traffic Drift[C]//Proceedings of the 2026

Network and Distributed System Security Symposium (NDSS). San Diego: NDSS, 2026.

[118] SHEN M, WU J H, AI J Y, et al. Swallow: A transfer-robust website fingerprinting attack via consistent feature learning[C/OL]//Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. Taipei Taiwan: ACM, 2025: 1574-1588.

[119] KADIAKAKIS G, POLYZOS T, PERRY M, et al. Tor circuit fingerprinting defenses using adaptive padding[A/OL]. arXiv, 2022.

[120] WANG T. High precision open-world website fingerprinting[C/OL]// 2020 IEEE Symposium on Security and Privacy (SP). San Francisco, CA, USA: IEEE, 2020: 152-167.

[121] DENG X H, LI Q, XU K.Robust and reliable early-stage website fingerprinting attacks via spatial-temporal distribution analysis[A/OL]. arXiv, 2024.

[122] CAI X, ZHANG X C, JOSHI B, et al. Touching from a distance: Website fingerprinting attacks and defenses[C/OL]//Proceedings of the 2012 ACM conference on Computer and communications security. Raleigh North Carolina USA: ACM, 2012: 605-616.

[123] WANG T, GOLDBERG I. Improved website fingerprinting on tor[C/OL]//Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society. Berlin Germany: ACM, 2013: 201-212.

[124] WANG T, GOLDBERG I. On realistically attacking tor with website fingerprinting[J/OL]. Proceedings on Privacy Enhancing Technologies, 2016, 2016(4): 21-36.

[125] PANCHENKO A, LANZE F, ZINNEN A, et al. Website fingerprinting at internet scale[C/OL]//Proceedings 2016 Network and Distributed System Security Symposium. San Diego, CA: Internet Society, 2016.

[126] OH S E, SUNKAM S, HOPPER N. p-FP: Extraction, classification, and prediction of website fingerprints with deep learning[A/OL]. arXiv, 2018.

[127] BAHRAMALI A, BOZORGI A, HOUMANSADR A. Realistic website fingerprinting by augmenting network traces[C/OL]//Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security. Copenhagen Denmark: ACM, 2023: 1035-1049.

[128] ZHAO X Y, DENG X H, LI Q, et al. owards fine-grained webpage fingerprinting at scale[A/OL]. arXiv, 2024.

[129] MITSEVA A, PANCHENKO A. Stop, Don't Click Here Anymore: Boosting Website Fingerprinting By Considering Sets of Subpages [C]//Proceedings of the 33rd USENIX Security Symposium. Philadelphia: USENIX Association, 2024.

[130] DENG X H, ZHAO X Y, YIN Q L, et al. Towards robust multi-tab website fingerprinting[A/OL]. arXiv, 2025.

[131] SHADBEH M, KHAJAVI K, WANG T. Reality check for tor website fingerprinting in the open world[A/OL]. arXiv, 2026.

[132] YUAN X. Mechanisms and modeling of concept drift in torbasedWebsite fingerprinting[C/OL]//Proceedings of the 4th International Conference on Artificial Intelligence and Intelligent Information Processing. Qingdao China: ACM, 2025: 875-880.

[133] SHUSTERMAN A, DAVID R, OREN Y. Understanding and addressing concept drift in website fingerprinting[J/OL]. Computer Networks, 2026, 275: 111811.

[134] GRINI A, TAHERI O, EL KHAMLICHI B, et al. Constrained network adversarial attacks: Validity, robustness, and transferability[A/OL]. arXiv, 2025.

[135] WITWER E, HOLLAND J, HOPPER N. Padding-only defenses add delay in tor[A/OL]. arXiv, 2022.

[136] GUAN Z, XIONG G, GOU G P, et al.BAPM: Block attention profiling model for multi-tab website fingerprinting attacks on tor[C/OL]// Annual Computer Security Applications Conference. Virtual Event USA: ACM, 2021: 248-259.

[137] JANSEN R, WAILS R, JOHNSON A.Repositioning real-world website fingerprinting on tor[C/OL]//Proceedings of the 23rd Workshop on Privacy in the Electronic Society. Salt Lake City UT USA: ACM, 2023: 124-140.

[138] JANSEN R. CELLSHIFT: RTT-Aware Trace Transduction for Real-World Website Fingerprinting [C]//Proceedings of the 33rd Network and Distributed System Security Symposium (NDSS 2026). San Diego, CA, USA: The Internet Society, 2026.

[139] YAN J Z, LIU Z T, XIE Y Y, et al. CertTA: Certified Robustness Made Practical for Learning-Based Traffic Analysis [C]//Proceedings of the 34th USENIX Security Symposium (USENIX Security' 25). Seattle, WA, USA: USENIX Association, 2025: 7349-7368.

王康旭（2002– ），男，黑龙江牡丹江人，信息工程大学博士生，主要研究方向为网络空间安全，隐私保护。

扈红超（1982- ），男，河南商丘人，博士，信息工程大学教授、博士生导师，主要研究方向为网络空间安全等。

程国振（1986– ），男，山东菏泽人，博士，信息工程大学副教授，主要研究方向为网络主动防御、软件多样化。

任权（1994– ），男，湖南常德人，博士，信息工程大学助理研究员，主要研究方向为网络空间安全、内生安全防御等。

周大成（1995– ），男，河南息县人，博士，信息工程大学助理研究员，主要研究方向为网络空间安全、云计算等。

常慧妍（2000– ），女，河南三门峡人，信息工程大学硕士生，主要研究方向为隐蔽通信和图像隐写。

刘文彦（1986-），男，河南周口人，博士，信息工程大学副研究员，主要研究方向为网络主动防御、云计算安全和匿名网络安全。

杨晓晗（1989-），女，河南南阳人，博士，信息工程大学助理研究员，主要研究方向为网络空间安全、拟态防御、云计算安全等。