TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 04/38 pp1381–1392
DOI: 10.26599/TST.2025.9010167
Volume 31, Number 3, June 2026

Research Article

Beyond a Single Perspective: Towards a Realistic Evaluation of Website Fingerprinting Attacks

Xinhao Deng, Jingyou Chen, Linxiao Yu, Yixiang Zhang, Zhongyi Gu, Changhao Qiu, Xiyuan Zhao, Ke Xu, and Qi Li*

Abstract: Website Fingerprinting (WF) attacks exploit patterns in encrypted traffic to infer the websites visited by users, posing a serious threat to anonymous communication systems. Although recent WF techniques achieve over 0.9 accuracy in controlled experimental settings, most studies remain confined to single scenarios, overlooking the complexity of real-world environments. This paper presents the first systematic and comprehensive evaluation of existing WF attacks under diverse realistic conditions, including defense mechanisms, traffic drift, multi-tab browsing, early-stage detection, open-world settings, and few-shot scenarios. Experimental results show that many WF techniques with strong performance in isolated settings degrade significantly when facing other conditions. Since real-world environments often combine multiple challenges, current WF attacks are difficult to apply directly in practice. This study highlights the limitations of WF attacks and introduces a multidimensional evaluation framework, offering critical insights for developing more robust and practical WF attacks.

Key words: traffic analysis; Website Fingerprinting (WF); deep learning

1 Introduction

Website Fingerprinting (WF) attacks have emerged as an important and active research area in anonymous communication over the Internet[1, 2]. By analyzing encrypted traffic features, such as packet size, direction, and timing patterns, adversaries can infer the websites visited by users, thereby undermining the privacy guarantees of encrypted communication. Over the past decade, WF techniques have evolved from statistical learning approaches to deep learning models, demonstrating steadily improving performance under controlled experimental conditions[3–7]. Existing research typically frames WF attacks as a classification problem, employing machine learning methods (e.g., Support Vector Machines (SVM) and Random Forests (RF)) or deep learning architectures (e.g., Convolutional Neural Networks (CNNs) and Transformers) to automatically extract traffic features. However, most reported results achieve high accuracy only within isolated experimental settings, lacking systematic evaluation under realistic conditions.

• Xinhao Deng, Yixiang Zhang, Zhongyi Gu, Xiyuan Zhao, and Qi Li are with Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing 100084, China. E-mail: dengxinhao@tsinghua.edu.cn; zhangyix24@mails.tsinghua.edu.cn; guzy24@mails.tsinghua.edu.cn; zhaoxy23@mails.tsinghua.edu.cn; qli01@tsinghua.edu.cn.
• Jingyou Chen and Linxiao Yu are with School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China. E-mail: 230240004@seu.edu.cn; yulinxiaobbb@seu.edu.cn.
• Changhao Qiu is with National Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha 410073, China. E-mail: qch@nudt.edu.cn.
• Ke Xu is with Department of Computer Science, Tsinghua University, Beijing 100084, China. E-mail: xuke@tsinghua.edu.cn.
* To whom correspondence should be addressed.
Manuscript received: 2025-09-22; revised: 2025-10-27; accepted: 2025-11-03

© The author(s) 2026. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/).

In practice, multiple challenges significantly affect the performance of WF attacks. Juarez et al.[8] shown that attack accuracy decreases substantially when tested against larger website sets[9], diverse Tor versions[10], more complex user behaviors[11–13] and temporal drift[7, 14]. In other words, many high-accuracy results rely on unrealistic assumptions, such as single-tab browsing, static website content, or consistency between training and testing data[8]. Moreover, adversaries are often assumed to possess complete knowledge of a user's browsing scope, an assumption that rarely holds in open network environments[15]. Studies based on real Tor traffic further reveal that even state-of-the-art models experience dramatic performance degradation in more realistic open-world scenarios. For example, when the monitored set includes only five popular websites, accuracy can exceed 0.95; yet when expanded to 25 websites, accuracy drops below 0.8[16]. This illustrates the practical difficulty of reliably identifying and monitoring large-scale website visits.

To address these limitations, this paper proposes a more comprehensive and realistic WF attack evaluation framework. We identify and categorize six key challenges: (1) defense mechanisms, (2) traffic drift, (3) multi-tab browsing, (4) early-stage detection, (5) open-world scenarios, and (6) the few-shot settings. Based on these challenges, we construct a unified experimental evaluation system that integrates them into WF research and systematically analyzes mainstream methods under complex conditions. Experimental results in Fig. 1 demonstrate that many WF techniques achieving strong performance in single scenarios suffer substantial accuracy degradation in other contexts. This finding underscores the limitations of current methods and highlights the need for future research to conduct comprehensive evaluations in diverse and dynamic environments, thereby advancing the development of more practical and robust WF attack techniques.

2 Background & Related Work

2.1 WF attacks

Traditional approaches. Early WF attacks primarily relied on handcrafted traffic features in combination with traditional machine learning classifiers. Machine learning models are widely used in the fields of security and privacy[17]. In these approaches, attackers

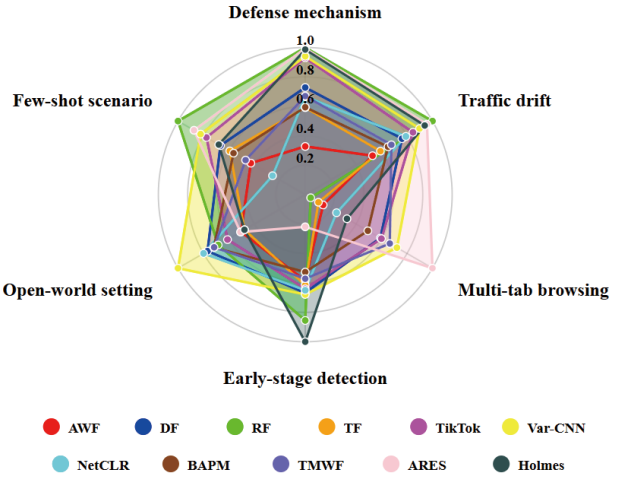


Fig. 1 Comparative evaluation of WF attacks across key real-world scenarios.

first collect encrypted traffic traces from various websites and then extract statistical features, such as sequence length, packet size distribution, traffic directionality, and inter-packet timing. Liberatore and Levine[18] pioneered the use of packet-size frequency histograms, while Panchenko et al.[19] improved accuracy through refined distance statistics and SVMs. Later, Cai et al.[20] introduced edit-distance and template-matching techniques. These methods generally involve manual feature engineering, followed by classification using algorithms such as k-Nearest Neighbors (k-NN), Naïve Bayes, SVMs, or random forests. Although they achieved relatively high accuracy in small-scale closed-world settings, their performance is highly sensitive to feature selection and parameter tuning. Moreover, their effectiveness plateaus in large-scale or more complex environments, limiting their practical deployment.

Deep learning approaches. With the rise of deep learning, researchers increasingly shifted toward automated feature extraction, thereby reducing reliance on manual engineering. A variety of neural architectures have been proposed for WF attacks. Sirinam et al.[21] introduced Deep Fingerprinting (DF), which leverages CNNs to directly learn from raw packet sequences. Building on this, Rimmer et al.[7] explored residual and recurrent neural architectures. These models automatically capture discriminative traffic patterns and have achieved remarkable accuracy in closed-world settings, with accuracy often exceeding 0.95. For example, Sirinam et al.[14]'s CNN model reported over 0.98 accuracy on undefended Tor traffic. Subsequent variants, such as Var-CNN[22] and TF[14],

further enhance temporal robustness and inter-class separability. Despite these advances, deep learning models generally require large-scale training datasets and relatively stable operating conditions. Their performance tends to degrade under data scarcity or distribution shifts, issues that will be revisited later in this study.

These specialized efforts represent significant progress toward practical WF attacks, but often optimize for isolated challenges. Their limited adaptability under complex, real-world conditions motivates the broader evaluation undertaken in this work.

2.2 Focal aspects of attacks

WF attacks identify websites by learning discrepancies in traffic patterns. Early attacks, such as AWF[7], focus on undefended traffic. With the development of defense mechanisms, later works proposed models robust to obfuscated or distorted traffic, such as TF[14] and NetCLR[23]. However, most existing approaches assume that the user visits only a single website per trace, and that the adversary can capture the entire traffic trace, which may not hold in real-world scenarios. To address these limitations, BAPM[13] and ARES[11] consider multi-tab browsing, where the user accesses multiple websites concurrently. Furthermore, Holmes[24] introduces a model that remains effective even when only partial traffic is captured. We detail these attacks below and summarize their primary focus and addressed challenges in Table 1.

AWF[7]. AWF provides the first systematic exploration of deep learning algorithms, such as CNNs and Long Short-Term Memory Networks (LSTMs), to automatically extract website traffic features and select hyperparameters. It evaluates models on a large open-world dataset to demonstrate competitive performance. Moreover, it evaluates traffic from different time periods to assess robustness against concept drift.

DF[21]. DF employs a deep network architecture consisting of multiple consecutive CNN blocks, thereby improving its capability for more effective feature extraction. Furthermore, DF integrates advanced deep learning techniques, such as batch normalization and dropout, to mitigate overfitting and enhance the capture of richer traffic information. As a result, DF exhibits superior performance in overcoming defenses, including Website Traffic Fingerprinting Protection with Adaptive Defense (WTF-PAD).

Var-CNN[22]. Var-CNN recognizes that packet sequences often exhibit long-term dependencies, making it challenging to model global information

Table 1 Summary of representative WF attacks and their addressed challenges.

Attack	Year	Description	Defense	Traffic drift	Multi-tab	Early-stage	Open-world	Few-shot
AWF[7]	2018	Deep learning–based WF with automated hyperparameter tuning for optimal model selection.	×	√	×	×	√	×
DF[21]	2018	CNN-based approach to capture richer traffic features and mitigate overfitting.	√	×	×	×	√	×
Var-CNN[22]	2019	Ensemble of softmax outputs from packet timing and direction, reducing training data requirements.	√	√	×	×	√	√
Tik-Tok[25]	2020	Histogram-based burst features to improve resilience against padding defenses.	√	×	×	×	√	×
TF[14]	2019	Triplet-network feature extractor for robust and transferable traffic representations.	√	√	×	×	√	√
BAPM[13]	2021	Attention mechanism to segment multi-tab traffic into single-tab flows for classification.	×	×	√	×	√	×
ARES[11]	2023	Multi-label framework for tab-level traffic division and website recognition without prior knowledge.	√	√	√	×	√	×
NetCLR[23]	2023	Burst-based augmentation to simulate diverse network conditions and enhance robustness to drift.	√	√	×	×	√	√
TMWF[12]	2023	Transformer-based WF for multi-tab traffic with calibrated metrics for realistic evaluation.	×	×	√	×	√	×
RF[26]	2023	TAM jointly encodes time and direction to improve defense robustness.	√	×	×	×	√	×
Holmes[24]	2024	Early-stage WF using supervised contrastive learning to extract website features from limited traffic.	√	×	×	√	√	×

using Recurrent Neural Network (RNNs) or LSTMs due to their high computational cost. To overcome this limitation, Var-CNN employs dilated convolutions, which effectively capture temporal dependencies within packet sequences while avoiding the added training overhead associated with RNNs and LSTMs. By combining the softmax outputs of packet timing and direction, Var-CNN achieves competitive performance, requiring less training data compared to previous methods.

Tik-Tok[25]. Tik-Tok builds upon the CNN architecture used in DF, enhancing it with a more sophisticated input representation that combines packet direction and raw timing information. This modification allows Tik-Tok to detect more subtle patterns in the data, improving both its accuracy and efficiency compared to the original DF attack. By incorporating both directional and temporal features, Tik-Tok demonstrates superior performance in overcoming defenses.

TF[14]. TF employs a Triplet Network[27] to learn discriminative traffic representations by minimizing intra-class distances and maximizing inter-class distances. During classification, the triplet network encodes the traffic traces into embedding vectors. This separation of feature extraction from classification reduces the need for frequent retraining.

BAPM[13]. BAPM focuses on detecting multi-tab WF attacks. Given a trace generated by sequentially opening multiple tabs, BAPM divides the trace into blocks, each representing partial traffic features. It then uses an attention mechanism[28] to model the relationships between these blocks. Blocks with higher attention scores are grouped together to identify the originating website. While BAPM effectively distinguishes mixed traffic, it requires prior knowledge of the number of tabs, which limits its practical applicability.

ARES[11]. ARES redefines multi-tab WF as a multi-label classification problem. Rather than using a single model, it deploys a dedicated model, Trans-WF, for each website. ARES aggregates the predictions from all Trans-WFs to identify the visited websites. To enhance detection, it employs a multi-head Top-m attention mechanism that strengthens local patterns within the same website while suppressing patterns from different websites. This approach ensures robust detection across varying numbers of tabs, eliminating the need for prior knowledge of tab counts.

NetCLR[23]. NetCLR leverages data augmentation to simulate traffic under various network conditions. For each burst, it generates multiple variants that capture dynamic content and fluctuating Tor bandwidth by modifying, inserting, or merging bursts. During training, NetCLR first performs pre-training on unlabeled traces to learn robust feature representations, then fine-tunes the model using labeled traces. Experimental results demonstrate that NetCLR enhances the robustness of WF attacks against network variability and improves generalization capabilities.

TMWF[12]. TMWF utilizes the Transformer[28] as its classification module, while employing DF as the feature extractor. For websites that fall outside the monitored set, TMWF assigns a "no-tab" label to filter out irrelevant predictions. Furthermore, TMWF introduces new metrics to recalibrate accuracy, precision, and recall in the presence of unmonitored websites, thereby aligning more closely with realistic attacker objectives. The method relies on a predefined parameter, N, representing the maximum number of tabs. Setting N to a sufficiently large value enhances the method's practicality.

RF[26]. RF introduces a novel representation, TAM, which encodes packet counts and directions within fixed time slots. TAM is designed to withstand obfuscation techniques, such as padding and delays, making it robust against a variety of defenses. The model combines 2-D and 1-D CNNs with 2×2 max-pooling and replaces fully connected layers with Global Average Pooling (GAP)[29] layers to mitigate overfitting. RF demonstrates superior open-world performance, particularly in defensive settings.

Holmes[24]. In real-world scenarios, only partial traffic traces may be observable, causing models trained on complete traces to perform poorly under such conditions. Holmes addresses this issue by identifying websites using "early-stage traffic" collected during the initial page load. The authors demonstrate that early-stage traffic contains enough information for effective classification. To enhance early-stage traces, Holmes utilizes temporal data augmentation and applies supervised contrastive learning[30] to align early and complete traces in latent space. During deployment, Holmes adaptively classifies traces once confidence exceeds a predefined threshold, thereby enhancing robustness against various defenses.

3 Threat Model & Problem Statement

3.1 Threat model

Figure 2 presents the threat model for WF attacks, which aim to infer visited websites by analyzing encrypted traffic patterns, thereby undermining anonymous communication. We assume a local passive adversary (such as an Internet service provider or a compromised Tor entry relay) who can monitor all traffic between the user and the Tor network[8]. The adversary cannot access plaintext content but can collect metadata (packet timing, direction, and size), and seeks to reconstruct the user's browsing activity from these observable traces.

In a closed-world setting, WF is formulated as a multi-class classification problem over a predefined set of monitored websites: the adversary collects traffic samples per site and trains a classifier mapping observed traces to one of these classes. In an open-world setting, an additional "other" class represents unmonitored sites, transforming the task into an $(N+1)$-class classification or, equivalently, a detection problem in which the adversary must both identify monitored sites and reject unknowns. In both cases, the adversary's objective is to maximize identification accuracy under given constraints.

3.2 Problem statement

While WF attacks perform well when website traffic is statistically distinctive and ample training data exist, their effectiveness is notably hindered by several practical constraints:

- **Defense mechanisms.** Traffic obfuscation techniques, such as adaptive padding, random delays, and other shaping defenses, diminish discriminative features and introduce noise, often leading to significant accuracy degradation unless mitigated by specialized countermeasures[31].
- **Traffic drift.** Temporal variations in website content and network conditions lead to distribution shifts, causing models trained on outdated data to degrade in performance unless they are regularly updated[32].
- **Multi-tab browsing.** When users simultaneously access multiple websites through different browser tabs, the single-traffic classification task evolves into a multi-label classification problem[6, 33–35].
- **Early-stage detection.** The adversary can capture only the traffic from the initial stage of webpage loading, which contains limited website information, thereby increasing the difficulty of fingerprint identification[24].
- **Open-world setting.** Users may visit a vast number of unmonitored websites that are unknown to the adversary, with the majority of network traffic originating from these unmonitored sources. Therefore, the adversary must accurately differentiate between traffic associated with monitored and unmonitored websites[9].
- **Few-shot scenario.** Each monitored website possesses only a limited amount of labeled traffic for model training, which consequently constrains the model's performance[14, 36].

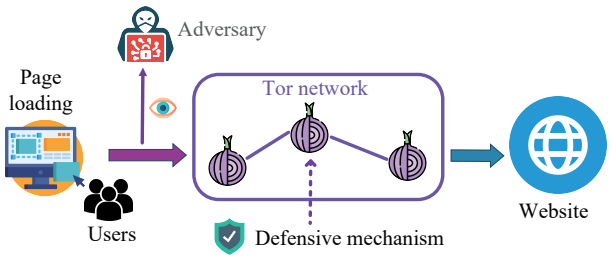


Fig. 2 WF attack model.

4 Attack Analysis

In this section, we evaluate the performance of existing WF attacks across six practical scenarios: defense mechanisms, traffic drift, multi-tab browsing, early-stage detection, open-world, and few-shot setting.

4.1 Dataset construction

We construct six evaluation datasets to provide a comprehensive assessment of WF methods. The defense dataset includes traffic from 95 websites and covers three defense techniques: WTF-PAD, Front, and Walkie-Talkie[25]. To evaluate robustness in the face of traffic drift, we use a traffic dataset from 102 websites spanning from March 2024 to July 2025, totaling 140 000 samples. For the multi-tab browsing scenario, we collect traffic data from sessions involving 2–5 tabs across 100 websites, resulting in 58 000 samples[11]. To assess early detection performance, we generated test data by partially masking traffic from 95 websites at 20%, 40%, 60%, and 80% of page loading progress[24]. For open-world evaluation, we utilize the GTT23 dataset[15], which includes large-scale traffic data from non-monitored websites. Finally, to evaluate performance in the few-shot setting, we limit each

monitored website to only 10 labeled traffic samples for model training or fine-tuning, allowing us to assess the model's adaptability to scenarios with limited data.

4.2 Evaluation under defense mechanisms

To evaluate the performance of WF attacks under different defense mechanisms, we employ a dataset that integrates three representative defense schemes: WTF-PAD, Front, and Walkie-Talkie. This dataset comprises traffic traces from 95 websites, each contributing approximately 1000 samples. For every trace, one of the three defense mechanisms is applied with equal probability. Such a hybrid defense dataset allows for a comprehensive assessment of robustness against composite defense strategies.

As shown in Table 2, the accuracy varies substantially with the choice of feature representation in the WF model. When only packet direction sequences are used, the accuracy ranges from 0.3198 to 0.7042. Incorporating both packet direction and timestamp information considerably enhances performance, yielding accuracies between 0.8910 and 0.9007. Further improvement is achieved through feature aggregation, which raises accuracy to between 0.9458 and 0.9582.

The observed performance pattern primarily arises from the inherent properties of the defense mechanisms. These defenses introduce numerous dummy packets into the traffic traces, which collectively obscure both packet direction and temporal characteristics, thereby diminishing their discriminative effectiveness. In contrast, aggregating features through statistical analysis within fixed-length time windows mitigates the temporal distortions caused by these defenses and consequently enhances the overall robustness of the model.

Table 2 Performance comparison of WF attacks under defense mechanisms. The best results are highlighted in bold.

Attack	Accuracy	Precision	Recall	F1-score
AWF	0.3198	0.3203	0.3199	0.3141
DF	0.7042	0.7053	0.7066	0.6982
Var-CNN	0.9007	0.9130	0.8995	0.9019
TikTok	0.8910	0.8971	0.8918	0.8922
TF	0.5767	0.5772	0.5693	0.5673
BAPM	0.5750	0.6023	0.5718	0.5697
ARES	0.9530	0.9560	0.9537	0.9542
NetCLR	0.6490	0.6424	0.6403	0.6382
TMWF	0.6550	0.6580	0.6486	0.6423
RF	**0.9582**	**0.9605**	**0.9591**	**0.9593**
Holmes	0.9458	0.9469	0.9458	0.9458

4.3 Evaluation under traffic drift

To evaluate model robustness under temporal distribution shifts, all models are and validated using traffic data from Day 0 and tested on uniformly sampled traces collected from later checkpoints.

As shown in Table 3, all WF attacks demonstrate a noticeable decline in performance over time, which highlights the inherent challenge of maintaining stable classification accuracy as traffic drift. Among the evaluated models, RF, ARES, and Holmes, which rely on comprehensive and diverse feature representations, show the strongest robustness, with RF achieving the highest accuracy of 0.6911. TikTok and Var-CNN incorporate temporal information into directional sequences and obtain moderate improvements, yet their accuracy remains below that of RF and ARES. In comparison, AWF and TF, which rely solely on directional sequences without additional contextual information, achieve the weakest performance and experience the largest accuracy decline.

These findings suggest that the robustness of website fingerprinting models under traffic drift is closely tied to the richness and diversity of their feature representations.

4.4 Evaluation under multi-tab browsing

To evaluate the robustness of WF attacks under realistic browsing behaviors, we analyze four multi-tab

Table 3 Performance comparison of WF attacks in the context of traffic drift. The best results are highlighted in bold.

Attack	Before drift		After drift	
	Accuracy	F1-score	Accuracy	F1-Score
AWF	0.5438	0.5369	0.3706	0.3602
DF	0.7358	0.7319	0.5247	0.5179
Var-CNN	0.8112	0.8097	0.6188	0.6122
TikTok	0.8033	0.7989	0.5851	0.5767
TF	0.5908	0.5856	0.4101	0.4025
BAPM	0.6421	0.6375	0.4519	0.4431
ARES	0.8629	0.8613	0.6595	0.6517
NetCLR	0.7512	0.7491	0.5466	0.5387
TMWF	0.6571	0.6539	0.4676	0.4612
RF	**0.8762**	**0.8688**	**0.6911**	**0.6810**
Holmes	0.8338	0.8287	0.6458	0.6397

datasets, denoted as 2-tab, 3-tab, 4-tab, and 5-tab datasets. Each dataset is collected by visiting 100 monitored websites over the Tor network within a closed-world setting. To emulate realistic user activity, 25% of the traces from each dataset are randomly mixed to form a composite dataset, which ultimately contains 46 980 training traces, 5220 validation traces, and 5800 testing traces. This setup reflects a more authentic and complex browsing scenario, where concurrent page loads introduce significant variations in traffic patterns.

As summarized in Table 4, most existing WF methods demonstrate only moderate robustness when evaluated on multi-tab datasets. Specifically, their AP@avg values range from 0.03 to 0.57, while P@avg typically remains below 0.40. Similarly, AUC scores span from 0.50 to 0.81, underscoring a notable decline in classification reliability under concurrent browsing conditions. This degradation primarily arises from elevated noise levels, inter-tab interference, and overlapping traffic flows, which collectively obscure website-specific patterns and hinder accurate model inference. In contrast, ARES exhibits strong robustness in this challenging scenario. Its top-m attention mechanism effectively emphasizes discriminative features within mixed flows, allowing the model to extract relevant signals while suppressing background noise.

Table 4 Performance comparison of WF attacks under multi-tab browsing settings. The best results are highlighted in bold.

Attack	AUC	Precision	AP
AWF	0.6049	0.0950	0.1104
DF	0.7610	0.3257	0.4630
Var-CNN	0.8100	0.4110	0.5650
TikTok	0.7689	0.3295	0.4693
TF	0.5433	0.0701	0.0814
BAPM	0.7790	0.2869	0.3858
ARES	**0.9432**	**0.6439**	**0.7839**
NetCLR	0.6594	0.1535	0.1930
TMWF	0.8417	0.3908	0.5211
RF	0.4972	0.0339	0.0336
Holmes	0.6405	0.1939	0.2571

In conclusion, our findings highlight that multi-tab browsing remains a fundamental obstacle for the practical deployment of WF attacks. Despite recent advances, accurately identifying websites under realistic, concurrent browsing conditions continues to challenge even the sophisticated models, underscoring the need for more adaptive WF approaches in future research.

4.5 Evaluation with early-stage detection

For the early-stage evaluation, we use a comprehensive traffic dataset collected from 95 monitored websites. Each website contributes more than 1000 individual traffic traces. To simulate different levels of page loading, each trace is divided into four incremental stages corresponding to 20%, 40%, 60%, and 80% of the total page load time, determined by packet timestamps. This segmentation enables a systematic assessment of model performance when only partial traffic information is available.

As shown in Table 5, Holmes demonstrates a clear advantage even during the earliest loading stage. When only 20% of the total load time is available, Holmes achieves an F1-score of 0.5292, which is substantially higher than the scores of RF at 0.2793 and ARES at 0.0263. When the load time increases to 40%, the F1-

Table 5 Comparison of WF attack performance at different page loading stages. The best results are highlighted in bold.

Attack	20% loaded				40% loaded				60% loaded				80% loaded			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
AWF	0.0950	0.3072	0.0953	0.1185	0.3200	0.4678	0.3201	0.3465	0.6335	0.6834	0.6334	0.6411	0.8717	0.8803	0.8716	0.8691
DF	0.1289	0.3780	0.1298	0.1481	0.4070	0.5670	0.4073	0.4293	0.7457	0.7835	0.7457	0.7475	0.9294	0.9392	0.9295	0.9282
Var-CNN	0.1225	0.4267	0.1227	0.1431	0.4323	0.6590	0.4322	0.4624	0.7655	0.8102	0.7655	0.7658	0.9371	0.9434	0.9372	0.9354
TikTok	0.1172	0.3751	0.1174	0.1325	0.3721	0.5288	0.3719	0.3848	0.7044	0.7498	0.7043	0.7080	0.9188	0.9262	0.9188	0.9163
TF	0.0933	0.3328	0.0938	0.1083	0.3291	0.5099	0.3295	0.3583	0.6877	0.7369	0.6876	0.6947	0.9095	0.9228	0.9095	0.9081
BAPM	0.0645	0.2613	0.0647	0.0805	0.2090	0.4022	0.2092	0.2334	0.5556	0.6449	0.5553	0.5694	0.8613	0.8862	0.8613	0.8629
ARES	0.0265	0.0633	0.0273	0.0263	0.0814	0.1382	0.0751	0.0742	0.1894	0.2902	0.1743	0.1834	0.4867	0.5334	0.4708	0.4552
NetCLR	0.1242	0.3314	0.1246	0.1400	0.4093	0.5325	0.4095	0.4190	0.7167	0.7546	0.7167	0.7160	0.9212	0.9298	0.9212	0.9195
TMWF	0.0911	0.2908	0.0918	0.1029	0.2822	0.4401	0.2826	0.2977	0.6353	0.6900	0.6350	0.6348	0.8963	0.9127	0.8962	0.8937
RF	0.2474	0.5748	0.2477	0.2793	0.7029	0.8068	0.7025	0.7163	0.9175	0.9272	0.9175	0.9170	0.9742	0.9769	0.9742	0.9740
Holmes	**0.5132**	**0.6673**	**0.5129**	**0.5292**	**0.9048**	**0.9090**	**0.9046**	**0.9020**	**0.9661**	**0.9676**	**0.9661**	**0.9659**	**0.9784**	**0.9787**	**0.9784**	**0.9784**

score of Holmes rises sharply to 0.902, further confirming its superiority over other WF attack models. These results highlight Holmes's strong ability to infer webpage identities from incomplete traffic data, reflecting greater robustness compared with other attacks.

Overall, the findings reveal a fundamental limitation of traditional WF attacks, which often rely on complete webpage traffic to ensure reliable classification. In real-world networks, however, traffic traces are frequently truncated, challenging the assumption of full data availability. Maintaining detection accuracy under such incomplete traffic conditions remains a key obstacle in developing practical WF attack models.

4.6 Open-world evaluation

To evaluate the model under open-world conditions, we utilize the GTT23 dataset. The training set comprises traffic data from 100 monitored websites, each containing approximately 100 traffic traces, as well as 5000 unmonitored traffic traces. The test set includes 100 traffic traces for each monitored website and 20 000 unmonitored traces. Importantly, there is no overlap between the training and testing datasets. This separation ensures a fair assessment of generalization and more accurately reflects realistic open-world environments.

Following prior studies[37], we adopt the r-precision metric, denoted as π_r. This metric integrates the ratio $r = N_N/N_P$, where N_N and N_P represent the numbers of unmonitored and monitored websites, respectively. The inclusion of this ratio adjusts the evaluation to account for the inherent imbalance between monitored and unmonitored classes that typically arises in open-world scenarios. In our experiments, we report π_1 and π_5, corresponding to commonly used open-world ratios.

As illustrated in Fig. 3, the π_1 scores of the evaluated methods range from 0.240 to 0.503, whereas the π_5 scores are markedly lower, with the best-performing model achieving 0.363. Among all approaches, single-tab attacks, such as DF, Var-CNN, and NetCLR consistently, outperform multi-tab attacks. Var-CNN attains the highest π_1 and π_5 values of 0.503 and 0.363, respectively, followed by NetCLR and DF. Conversely, multi-tab attack models, including BAPM, TMWF, and ARES, exhibit significantly weaker performance. These findings indicate that current multi-tab WF attacks suffer greater performance degradation in open-world settings, highlighting the challenges of maintaining robustness when faced with realistic, large-scale, and diverse traffic distributions.

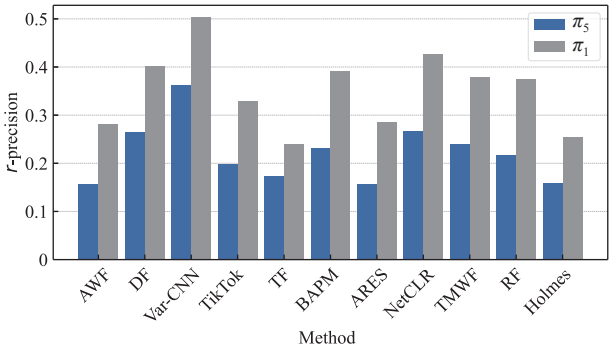


Fig. 3 Comparison of WF attacks in the open-world setting.

4.7 Evaluation under few-shot setting

To evaluate the learning performance of the model in the few-shot setting under temporal drift, we utilize a dataset collected in a traffic-drift scenario. Data collection began on 13 March 2024 and continued at six subsequent checkpoints: Days 14, 30, 90, 150, 270, and 480. The WF model was initially trained on the Day 0 data and later fine-tuned using only ten labeled samples per website from Days 270 and 480. This setup aims to emulate a realistic environment in which models must adapt to evolving data distributions while having access to limited labeled data.

As shown in Table 6, the RF and ARES achieve the most competitive performance among all the methods evaluated. The RF maintains relatively stable accuracy, achieving 0.8134 on Day 270 and 0.7004 on Day 480. ARES follows a similar pattern, but exhibits a steeper decline, reaching 0.6130 on Day 480. In contrast, the AWF, BAPM, and NetCLR perform considerably worse in the few-shot setting. Although the TF is specifically designed to improve few-shot learning, it also suffers a marked drop in accuracy under traffic drift, reaching only 0.4290 on Day 480.

These findings underscore the challenge of maintaining the stability and generalization capacity of the WF model when only a small amount of labeled data is available in dynamic real-world environments affected by traffic drift. They also highlight that robustness to distributional shifts is vital for ensuring long-term model reliability.

5 Discussion

No single WF attack can effectively address all the

Table 6 Performance comparison of WF attacks in the few-shot setting. The best results are highlighted in bold.

Attack	Day 270				Day 480			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
AWF	0.3722	0.3723	0.3717	0.3625	0.2944	0.2867	0.2943	0.2829
DF	0.5897	0.6002	0.5897	0.5795	0.432	0.4738	0.4321	0.4264
Var-CNN	0.7042	0.7090	0.7025	0.699	0.5412	0.5947	0.5411	0.5428
TikTok	0.6647	0.6715	0.6654	0.6586	0.5183	0.5457	0.5181	0.5134
TF	0.4914	0.4982	0.4914	0.4791	0.429	0.4279	0.4289	0.4162
BAPM	0.4916	0.5202	0.4927	0.483	0.3914	0.4027	0.3912	0.3680
ARES	0.7084	0.7422	0.7090	0.7097	0.6130	0.6541	0.6131	0.6103
NetCLR	0.2214	0.2419	0.2226	0.1945	0.2317	0.2377	0.2316	0.1935
TMWF	0.2594	0.4556	0.2616	0.2713	0.4419	0.4727	0.4419	0.4367
RF	**0.8134**	**0.8211**	**0.8140**	**0.8087**	**0.7004**	**0.7355**	**0.7003**	**0.7007**
Holmes	0.6195	0.6302	0.6175	0.5989	0.4466	0.4771	0.4465	0.4268

challenges encountered in real-world scenarios. Our results indicate that none of the existing WF attacks consistently maintains high performance under diverse real-world conditions. Although each approach exhibits certain strengths in specific contexts, none can be considered universally effective. This observation underscores the intrinsic complexity of real-world environments, where enhancing performance in one aspect often entails trade-offs in others. The main limitation of current WF attacks lies in their inadequate robustness and limited adaptability across varying scenarios. Therefore, future research should prioritize comprehensive evaluations under heterogeneous settings and the design of WF attack methods with stronger generalization and adaptability. Such advancements would enable models to sustain reliable performance even in dynamic, complex, and unpredictable environments.

Future work. Future research should explore three key directions. First, leveraging multi-task or meta-learning for multi-scenario joint optimization could enhance model adaptability across varied conditions. Second, dynamic adversarial strategies, inspired by game theory, may enable real-time adaptation between attacks and defenses[38]. Third, the development of standardized datasets is crucial: expanding the proposed framework with additional public datasets and evaluation metrics would facilitate the creation of a unified benchmarking platform. Such developments would not only improve the robustness of WF attacks but also provide insights for designing more resilient anonymity systems.

6 Conclusion

In this paper, we provide a comprehensive and multidimensional evaluation of existing WF attacks, addressing six key and realistic challenges: defense mechanisms, traffic drift, multi-tab browsing, early-stage detection, open-world scenarios, and few-shot settings. Through our experiments, we demonstrate that while certain attack methods achieve high accuracy in controlled, isolated scenarios, their performance significantly deteriorates when subjected to more complex conditions. This performance degradation highlights a critical gap in the robustness of current WF attacks across diverse scenarios. Our findings not only expose the limitations of previous studies, which often rely on single-perspective evaluations, but also underscore the inherent resilience of anonymity networks when confronted with varied protections and dynamic, real-world environments. This study calls for a shift towards more holistic and adaptable evaluation frameworks for WF attacks, taking into account the complexities and variabilities of practical deployment contexts.

Acknowledgment

The work was supported by the National Natural Science Foundation of China (Nos. 62132011 and 62425201).

References

[1] Users-tor metrics, https://metrics.torproject.org/userstats-relay-country.html, 2024.

[2] R. Dingledine, N. Mathewson, and P. Syverson, Tor: The Second-Generation Onion Router, https://www.usenix.org/conference/13th-usenix-security-symposium/tor-second-

generation-onion-router, 2025.

[3] T. Wang, X. Cai, R. Nithyanand, R. Johnson, and I. Goldberg, Effective attacks and provable defenses for website fingerprinting, in *Proc. 23rd USENIX Security Symp*., San Diego, CA, USA, 2014, pp. 143–157.

[4] J. Hayes and G. Danezis, k-fingerprinting: A robust scalable website fingerprinting technique, in *Proc. 25th USENIX Security Symp*., Austin, TX, USA, 2016, pp. 1187–1203.

[5] K. Wang, M. Yang, W. Yang, and Y. Yin, Deep correlation structure preserved label space embedding for multi-label classification, in *Proc. 10th Asian Conf. Machine Learning*, Beijing, China, 2018, pp. 1–16.

[6] Y. Xu, T. Wang, Q. Li, Q. Gong, Y. Chen, and Y. Jiang, A multi-tab website fingerprinting attack, in *Proc. 34th Annu. Computer Security Applications Conf*., San Juan, PR, USA, 2018, pp. 327–341.

[7] V. Rimmer, D. Preuveneers, M. Juarez, T. Van Goethem, and W. Joosen, Automated website fingerprinting through deep learning, in *Proc. Network and Distributed Systems Security* (*NDSS*) *Symp. 2018*, San Diego, CA, USA, 2018, pp. 1–15.

[8] M. Juarez, S. Afroz, G. Acar, C. Diaz, and R. Greenstadt, A critical evaluation of website fingerprinting attacks, in *Proc. 2014 ACM SIGSAC Conf. Computer and Communications Security*, Scottsdale, AZ, USA, 2014, pp. 263–274.

[9] T. Wang, High precision open-world website fingerprinting, in *Proc. 41st IEEE Symp. Security and Privacy*, San Francisco, CA, USA, 2020, pp. 152–167.

[10] X. Deng, X. Zhao, Q. Yin, Z. Liu, Q. Li, M. Xu, K. Xu, and J. Wu, Towards robust multi-tab website fingerprinting, arXiv preprint:2501.12622, 2025.

[11] X. Deng, Q. Yin, Z. Liu, X. Zhao, Q. Li, M. Xu, K. Xu, and J. Wu, Robust multi-tab website fingerprinting attacks in the wild, in *Proc. 44th IEEE Symp. Security and Privacy*, San Francisco, CA, USA, 2023, pp. 1005–1022.

[12] Z. Jin, T. Lu, S. Luo, and J. Shang, Transformer-based model for multi-tab website fingerprinting attack, in *Proc. 2023 ACM SIGSAC Conf. Computer and Communications Security*, Copenhagen, Denmark, 2023, pp. 1050–1064.

[13] Z. Guan, G. Xiong, G. Gou, Z. Li, M. Cui, and C. Liu, BAPM: Block attention profiling model for multi-tab website fingerprinting attacks on tor, in *Proc. 37th Annu. Computer Security Applications Conf*., Virtual Event, 2021, pp. 248–259.

[14] P. Sirinam, N. Mathews, M. S. Rahman, and M. Wright, Triplet fingerprinting: More practical and portable website fingerprinting with N-shot learning, in *Proc. 2019 ACM SIGSAC Conf. Computer and Communications Security*, London, UK, 2019, pp. 1131–1148.

[15] R. Jansen, R. Wails, and A. Johnson, A measurement of genuine tor traces for realistic website fingerprinting, arXiv preprint arXiv: 2404.07892, 2024.

[16] G. Cherubin, R. Jansen, and C. Troncoso, Online website fingerprinting: Evaluating website fingerprinting attacks on tor in the real world, in *Proc. 31st USENIX Security Symp*., Boston, MA, USA, 2022, pp. 753–770.

[17] Z. Chen, J. Mao, Q. Lin, L. Ma, and J. Liu, Empirical analysis of remote keystroke inference attacks and defenses on incremental search, *Tsinghua Science and Technology*, vol. 30, no. 6, pp. 2434–2451, 2025.

[18] M. Liberatore and B. N. Levine, Inferring the source of encrypted HTTP connections, in *Proc. 13th ACM Conf. Computer and Communications Security*, Alexandria, VA, USA, 2006, pp. 255–263.

[19] A. Panchenko, L. Niessen, A. Zinnen, and T. Engel, Website fingerprinting in onion routing based anonymization networks, in *Proc. 10th Annu. ACM Workshop on Privacy in the Electronic Society*, Chicago, IL, USA, 2011, pp. 103–114.

[20] X. Cai, X. C. Zhang, B. Joshi, and R. Johnson, Touching from a distance: Website fingerprinting attacks and defenses, in *Proc. 2012 ACM Conf. Computer and Communications Security*, Raleigh, NC, USA, 2012, pp. 605–616.

[21] P. Sirinam, M. Imani, M. Juarez, and M. Wright, Deep fingerprinting: Undermining website fingerprinting defenses with deep learning, in *Proc. 2018 ACM SIGSAC Conf. Computer and Communications Security*, Toronto, Canada, 2018, pp. 1928–1943.

[22] S. Bhat, D. Lu, A. Kwon, and S. Devadas, Var-CNN: A data-efficient website fingerprinting attack based on deep learning, *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 4, pp. 292–310, 2019.

[23] A. Bahramali, A. Bozorgi, and A. Houmansadr, Realistic website fingerprinting by augmenting network traces, in *Proc. 2023 ACM SIGSAC Conf. Computer and Communications Security*, Copenhagen, Denmark, 2023, pp. 1035–1049.

[24] X. Deng, Q. Li, and K. Xu, Robust and reliable early-stage website fingerprinting attacks via spatial-temporal distribution analysis, in *Proc. 2024 on ACM SIGSAC Conf. Computer and Communications Security*, Salt Lake City, UT, USA, 2024, pp. 1997–2011.

[25] M. S. Rahman, P. Sirinam, N. Mathews, K. G. Gangadhara, and M. Wright, Tik-Tok: The utility of packet timing in website fingerprinting attacks, *Proc. Privacy Enhancing Technol*., vol. 2020, no. 3, pp. 5–24, 2020.

[26] M. Shen, K. Ji, Z. Gao, Q. Li, L. Zhu, and K. Xu, Subverting website fingerprinting defenses with robust traffic representation, in *Proc. 32nd USENIX Security Symp*., Anaheim, CA, USA, 2023, p. 35.

[27] F. Schroff, D. Kalenichenko, and J. Philbin, FaceNet: A unified embedding for face recognition and clustering, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, Boston, MA, USA, 2015, pp. 815–823.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, Attention is all you need, in *Proc. 31st Int. Conf. Neural Information Processing Systems*, Long Beach, CA, USA, 2017, pp. 6000–6010.

[29] M. Lin, Q. Chen, and S. Yan, Network in network, arXiv preprint arXiv: 1312.4400, 2013.

[30] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, Supervised

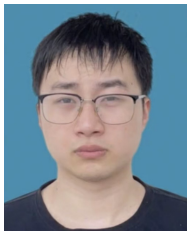
contrastive learning, in *Proc. 34th Int. Conf. Neural Information Processing Systems*, Vancouver, Canada, 2020, p. 1567.
[31] T. Wang and I. Goldberg, Walkie-talkie: An efficient defense against passive website fingerprinting attacks, in *Proc. 26th USENIX Security Symp.*, Vancouver, Canada, 2017, pp. 1375–1390.
[32] L. Yang, W. Guo, Q. Hao, A. Ciptadi, A. Ahmadzadeh, X. Xing, and G. Wang, CADE: Detecting and explaining concept drift samples for security applications, in *Proc. 30th USENIX Security Symp.*, Virtual Event, 2021, pp. 2327–2344.
[33] H. Yin, Y. Liu, Z. Guo, and Y. Wang, From traces to packets: Realistic deep learning based multi-tab website fingerprinting attacks, *Tsinghua Science and Technology*, vol. 30, no. 2, pp. 830–850, 2025.
[34] S. Huang, W. Niu, R. Wang, T. Zhang, K. Li, Y. Zhang, and X. Zhang, FlowCoPCL: Enhanced flow correlation attacks on tor using patching and contrastive learning, *Tsinghua Science and Technology*, vol. 31, no. 2, pp. 745–759, 2026.
[35] X. Zhao, X. Deng, Q. Li, Y. Liu, Z. Liu, K. Sun, and K. Xu, Towards fine-grained webpage fingerprinting at scale, in *Proc. 2024 on ACM SIGSAC Conf. Computer and Communications Security*, Salt Lake City, UT, USA, 2024, pp. 423–436.
[36] S. E. Oh, N. Mathews, M. S. Rahman, M. Wright, and N. Hopper, GANDaLF: GAN for data-limited fingerprinting, *Proc. Privacy Enhancing Technol.*, vol. 2021, no. 2, pp. 305–322, 2021.
[37] J. Zheng, Q. Li, G. Gu, J. Cao, D. K. Y. Yau, and J. Wu, Realtime DDoS defense using COTS SDN switches via adaptive correlation analysis, *IEEE Trans. Inf. Forensics Secur.*, vol. 13, no. 7, pp. 1838–1853, 2018.

Xinhao Deng received the PhD degree from Tsinghua University, Beijing, China in 2025. He is currently a joint postdoctoral researcher with Tsinghua University and Ant Group, China. His research interests include AI security and network security, with a particular focus on agent security and website fingerprinting.

Jingyou Chen received the MEng degree in software engineering from Northwestern Polytechnical University, Xi'an, China in 2024. He is currently a PhD candidate in cyber science and engineering at Southeast University, Nanjing, China. His research interests include website fingerprinting and network traffic anomaly detection.

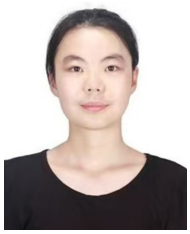
Linxiao Yu received the BEng degree in cyber science and engineering from Southeast University, Nanjing, China in 2023. He is currently a PhD candidate in cyber science and engineering at Southeast University, Nanjing, China. His research interests include encrypted network traffic classification and the theory of kernel methods.

Yixiang Zhang received the BEng degree in computer science and technology from Tsinghua University, Beijing, China in 2024. He is currently a PhD candidate at Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China. His research interests include encrypted traffic analysis, anonymous network security, and large model agent security.



Zhongyi Gu received the BEng degree in computer science and technology from Nanjing University of Science and Technology, Nanjing, China in 2024. He is currently a master student at Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China. His research interests include network traffic analysis and agent security.

Changhao Qiu received the BEng degree in software engineering from Shandong University, Jinan, China in 2021. He is currently a PhD candidate in control science and engineering at National University of Defense Technology, Changsha, China. His research interests include traffic engineering, website fingerprinting, and network topology obfuscation.



Xiyuan Zhao received the BEng degree in computer science and technology from Tsinghua University, Beijing, China in 2023. She is currently a master student at Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing, China. Her main research interest is encrypted traffic detection.

Ke Xu received the PhD degree from Tsinghua University, Beijing, China in 2001, where he is currently a professor at Department of Computer Science and Technology. He has been funded by the National Science Fund for Distinguished Young Scholars. He received the First Prize and Second Prize of the National Science and Technology Progress Award, the Second Prize of the National Technological Invention Award, and the First Prize of the China Institute of Electronics Science and Technology Award. He is a fellow of the China Institute of Electronics and an IEEE Fellow. He has published over 100 papers in the top four security conferences and leading international journals in networking and security. He has received multiple paper awards, including Distinguished Paper Awards at USENIX Security 2023, USENIX Security 2024, and NDSS 2025.

Qi Li received the PhD degree from Tsinghua University, Beijing, China in 2012. He is currently an associate professor and PhD supervisor at Institute for Network Sciences and Cyberspace, Tsinghua University, China, where he also serves as the deputy director of the National Laboratory for Internet Architecture. His research interests include network and system security, particularly Internet security, mobile security, and machine learning security. He is currently an editorial board member of the *IEEE Transactions on Dependable and Secure Computing*, *ACM Transactions on Privacy and Security*, and *ACM Digital Threats*: *Research and Practice*, and has served on the organization and program committees of various premier conferences.