Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.072184

ARTICLE

Adaptive Windowing with Label-Aware Attention for Robust Multi-Tab Website Fingerprinting

Chunqian Guo* **and Gang Chen**

School of Cyberspace Security, Zhengzhou University, Zhengzhou, 450001, China

*Corresponding Author: Chunqian Guo. Email: zzulzh23@gs.zzu.edu.cn

Received: 21 August 2025; Accepted: 11 December 2025; Published: 12 March 2026

ABSTRACT: Despite the ability of the anonymous communication system The Onion Router (Tor) to obscure the content of communications, prior studies have shown that passive adversaries can still infer the websites visited by users through website fingerprinting (WF) attacks. Conventional WF methodologies demonstrate optimal performance in scenarios involving single-tab browsing. Conventional WF methods achieve optimal performance primarily in scenarios involving single-tab browsing. However, in real-world network environments, users often engage in multi-tab browsing, which generates overlapping traffic patterns from different websites. This overlap has been shown to significantly degrade the performance of classifiers that rely on the single-tab assumption. To address this challenge, this paper proposes a Transformer-based multi-tab website fingerprinting (MT-WF) attack framework. The model employs an adaptive sliding window mechanism to capture fine-grained features of traffic direction. Additionally, it incorporates a label-aware attention mechanism designed to dynamically separate and refine entangled traffic representations, enhancing the model's ability to distinguish between overlapping traffic patterns. Furthermore, the model leverages global traffic patterns through multi-segment feature fusion and incorporates an incremental learning (IL) strategy to adapt to the continuously evolving website categories in open-world environments. Experimental results demonstrate that the proposed method achieves a top-2 precision of 0.78 in the closed-world setting. In the open-world scenario, the model attains an F1 score of 0.904, outperforming most existing baselines. The proposed method maintains superior performance even under challenging conditions, including WF defenses and concept drift.

KEYWORDS: Tor; website fingerprinting (WF); multi-tab browsing; transformer-based model; label-aware attention; traffic analysis; privacy; cybersecurity

1 Introduction

With the growing demand for personal privacy in the digital era, anonymous communication systems have become essential technologies for safeguarding online freedom and countering surveillance. Among these systems, Tor has emerged as the most widely adopted, employing a multi-hop architecture with layered encryption to obscure communication endpoints effectively [1]. Although Tor encrypts the content of communications, metadata such as packet timing and direction remain visible. This metadata can generate distinctive website fingerprints, allowing adversaries to infer the sites visited by users through website fingerprinting (WF) attacks—a critical vulnerability in Tor's anonymity guarantees.

The complexity of WF attacks varies depending on the browsing context. The following subsections outline the differences between single-tab and multi-tab attack environments, beginning with the former. In recent years, deep learning has achieved remarkable success in image classification tasks, largely due to the development of powerful neural network architectures such as AlexNet [2], ResNet [3], and Vision

Transformer [4]. These advancements not only improved performance on large-scale benchmarks like ImageNet but also enabled the transfer of deep learning techniques to other domains [5]. For instance, in natural language processing, models like BERT [6] and GPT [7] have transformed tasks such as text classification, translation, and generation. In cybersecurity, deep neural networks and Transformer-based models have shown considerable potential for tasks like intrusion detection and traffic analysis [8,9]. This cross-domain success highlights the robust feature extraction and generalization capabilities of deep learning models, providing strong motivation for their application to emerging challenges such as website fingerprinting and encrypted traffic analysis.

Previous research has predominantly concentrated on single-tab WF attacks, operating under the assumption that users visit only one website per session. Based on this assumption, numerous machine learning and deep learning approaches have been proposed [10–14], achieving identification accuracy exceeding 96%. These results validate the effectiveness of WF attacks in simplified, controlled scenarios. Representative studies include k-fingerprinting and related statistical approaches [15,16], as well as early deep learning-based fingerprinting methods [17]. These results confirm the effectiveness of WF attacks in simplified, controlled scenarios. However, this single-tab assumption often does not hold in practical settings. Due to the inherent latency of the Tor network, users frequently open multiple tabs concurrently, generating overlapping traffic from several websites. This overlap distorts the unique features of individual sites and significantly degrades the performance of traditional single-tab WF methods.

To address the recognition challenges caused by traffic overlap in multi-tab scenarios, several mixed-traffic modeling approaches have been developed [18,19]. In recent years, multi-label learning has made significant strides in capturing label dependencies. For example, graph-based methods such as ML-GCN [20] and GCN-MLP [21] enhance performance by explicitly constructing label graphs. Beyond graph neural network approaches, LabelCoRank [22] introduces a co-ranking method that simultaneously models the similarity between labels and instances. By employing a dual-graph structure to enable collaborative propagation at both the label and instance levels, it effectively alleviates the label sparsity problem and provides valuable insights for subsequent label modeling strategies. Table 1 presents a summary of representative website fingerprinting methods and highlighting their key characteristics.

Table 1: Comparison of existing methods

Method	Temporal features	Multi-tab	Direction features	Practicality*
DF [10]	✗	✗	✓	✓
TF [12]	✓	✗	✓	✓
Var-CNN [13]	✗	✗	✓	✓
Tik-Tok [14]	✓	✗	✓	✓
k-FP [15]	✗	✗	✗	✗
k-NN [16]	✗	✗	✗	✗
SDAE [17]	✗	✗	✓	✗
ARES [23]	✗	✓	✓	✓
TMWF [24]	✗	✓	✓	✓
BAPM [25]	✗	✓	✓	✓
OSCAR [26]	✗	✓	✗	✓
ADWPF [27]	✓	✗	✓	✓
FMWF [28]	✓	✓	✗	✓

Note: *Practicality indicates that the attack accounts for critical real-world complexities.

The seminal ARES framework [23] was the first to formally address multi-tab WF as a multi-class classification problem, introducing three key innovations: (i) segmenting session traffic into fixed-length sliding window fragments; (ii) independently classifying each fragment using a Transformer-based model; and (iii) aggregating the outputs of individual classifiers to achieve final recognition. This modular design enhances interpretability, achieves high F1 scores (averaging 0.91 across test datasets), and provides robustness against standard WF defenses. ARES demonstrates superior feature utilization and greater practical applicability compared to existing WF attack methods. Building on the Transformer architecture, which has been widely adopted in network security due to its strong capability in global feature modeling (e.g., Zhang et al. [8]; Kim and Pak [9]), subsequent studies have further enhanced adaptability and generalization. TMWF [24] introduced an end-to-end Transformer model that employs label-aware attention to capture correlations across fragments without relying on preset segmentation rules, thereby enhancing generalization in multi-tab browsing scenarios. Building upon this design, the present work extends the label-aware attention with adaptive feature disentanglement and incremental learning, resulting in a more robust and generalizable multi-tab WF attack framework. Moreover, BAPM [25] enhances fragment consistency through a block-level alignment mechanism that automatically aggregates relevant fragments in an unsupervised manner, reducing reliance on manual segmentation. More recently, OSCAR [26] and ADWPF [27] advanced WF toward multi-label and fine-grained recognition by modeling inter-page correlations and dynamically attending to discriminative traffic segments. In parallel, few-shot and incremental WF methods [28] were proposed to address data scarcity and concept drift. These studies collectively indicate a shift toward scalable, adaptive, and data-efficient WF models, which directly motivates the design of our proposed approach. Therefore, developing a MT-WF attack model capable of modeling label relationships, handling overlapping traffic, and adapting to dynamic network conditions has become a pressing research challenge.

Existing website fingerprinting defenses typically aim to obscure traffic patterns and reduce classifier accuracy. Common techniques include delay manipulation, dummy packet injection, and traffic rerouting. For example, Mockingbird [29] generates adversarial traces using non-target pattern synthesis combined with adversarial training, effectively deceiving classifiers. Recent work has shifted toward lightweight defenses for their practicality and low resource requirements. Representative examples include WTF-PAD [30], which inserts adaptive padding packets during idle periods, and FRONT [31], which introduces controlled randomness into dummy packets to create variable traffic signatures. Both approaches provide strong resistance to attack while maintaining throughput efficiency [32–34]. In contrast, rigid defenses such as Tamaraw [35] enforce fixed-length and constant-rate transmissions, achieving robust protection at the cost of significant bandwidth and latency overhead. This trade-off has motivated the adoption of lightweight strategies like WTF-PAD and FRONT as more practical solutions for real-world deployment.

To address the challenges of MT-WF, researchers [25,36–38] have proposed several approaches that segment entire session traffic into smaller fragments, assuming each fragment corresponds to a single website. These methods then extract features from each segment and predict potential target sites. The ARES framework [23] formalized this approach as a multi-classifier architecture, using fixed-length sliding windows to capture local features. TMWF [24] extended this idea by introducing learnable label query vectors, which extract key features associated with specific websites across multiple segments, enabling more targeted feature selection. Although these methods improve recognition performance in multi-tab scenarios, they face two main limitations: (1) Fixed segmentation strategies may discard important local traffic features, and (2) attention mechanisms often overlook weakly discriminative but valuable features. To overcome these limitations, this study proposes a novel multi-tab WF attack method that integrates label-aware attention mechanisms with adaptive segmentation. The proposed approach models feature-label relationships at a

fine-grained level while incorporating incremental learning to enhance adaptation to new website categories, ultimately boosting overall recognition performance. The main contributions of this work include:

1. **Label-aware Adaptive Segmentation Mechanism:** To address feature misalignment caused by conventional fixed sliding windows, a segment-label matching score mechanism is proposed in this work. This mechanism dynamically adjusts window positions and lengths based on correlations between traffic content features and potential labels. It effectively captures local traffic pattern variations, preserves highly discriminative features during segmentation, and improves both segmentation accuracy and representational capacity.
2. **Multi-segment Feature Fusion with Incremental Learning:** The proposed model in this study employs a dual-path architecture to encode each traffic segment from both spatial and temporal dimensions. By incorporating label-specific query vectors and spatio-temporal attention mechanisms, the model computes cross-segment label correlations for feature aggregation. When combined with incremental learning, this design mitigates catastrophic forgetting and enables rapid adaptation to new website categories. It also enhances the perception of discriminative features and improves multi-tab recognition robustness.
3. **Implementation and Evaluation:** The proposed framework is implemented and evaluated on multi-tab traffic datasets. Experimental results show that the adopted method achieves an average top-2 precision of 0.826 when users open two tabs, outperforming baseline models by approximately 25%. Furthermore, the model maintains high recognition accuracy even in the presence of WF defenses.

The remainder of this paper is organized as follows. Section 2 introduces the proposed attack model, including the adaptive segmentation, website identification, and incremental learning modules. Section 3 describes the dataset used in the experiments and the evaluation metrics. Section 4 presents the experimental results and detailed analyses under closed-world, open-world, and defense scenarios. Finally, Section 5 concludes the paper and outlines directions for future work.

2 Method

2.1 Threat Model

In multi-tab WF attack scenarios, as illustrated in Fig. 1, adversaries—ranging from local observers to ISP or Tor entry node operators—cannot decrypt Tor traffic but can monitor metadata such as packet direction and timing. Conventional threat models assume a passive observer between the client and the Tor entry node, using metadata to infer visited websites and compromise Tor anonymity. WF attacks are typically evaluated in two settings: closed-world and open-world. In closed-world settings, the task is framed as a multi-class classification problem, where the model classifies visited sites from a fixed set of monitored sites. In open-world settings, the model must not only classify known sites but also discriminate against unknown ones. Multi-tab scenarios add complexity to this problem due to overlapping traffic patterns and label co-occurrence, highlighting the need for models with robust open-set recognition capabilities.

2.2 Model Design

To address the challenges posed by overlapping traffic in multi-tab browsing, this section presents a Transformer-based website fingerprinting attack model comprising three integrated components:

(1) Traffic Segmentation Module: This module employs an adaptive sliding window approach to segment multi-tab traffic while preserving local sequential patterns. The design guarantees robust feature extraction by leveraging adjacent windows that preserve full pattern coverage, even in the presence of disruptions caused by segmentation of partial patterns.

(2) Website Identification Module: This module combines both local and global features using attention mechanisms to establish cross-feature correlations. This enables precise multi-tab classification within each traffic segment.

(3) Incremental Learning Module: This module normalizes output and uses confidence-based ranking to generate final label predictions. By applying threshold-based decision making, the model achieves accurate multi-label recognition.

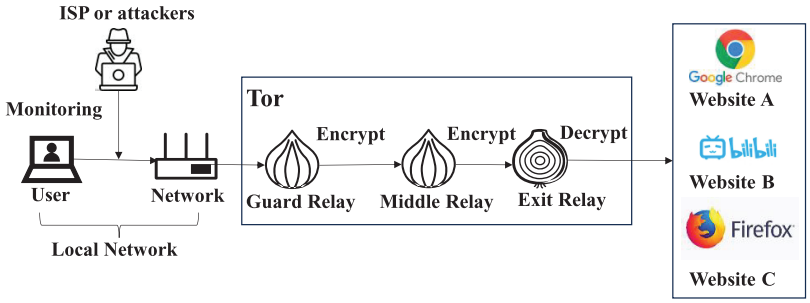

Figure 1: Threat model in multi-tab scenarios

The full architecture of the model is depicted in Fig. 2.

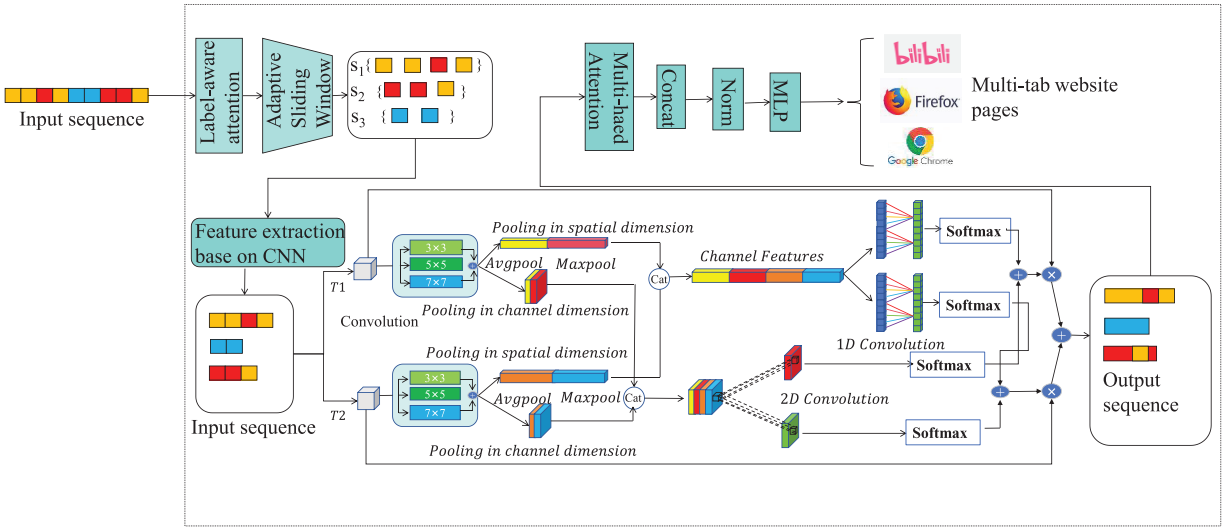

Figure 2: Architecture of the multi-label website fingerprinting attack model

2.2.1 Network Traffic Division Module

The traffic segmentation module partitions the entire session traffic into smaller segments to capture localized patterns of webpage loading events. However, fixed-length segmentation can fragment traffic sequences, potentially discarding valuable discriminative features. To address this limitation, the proposed solution in this study integrates adaptive sliding windows with label-aware attention mechanisms, as illustrated in Fig. 3.

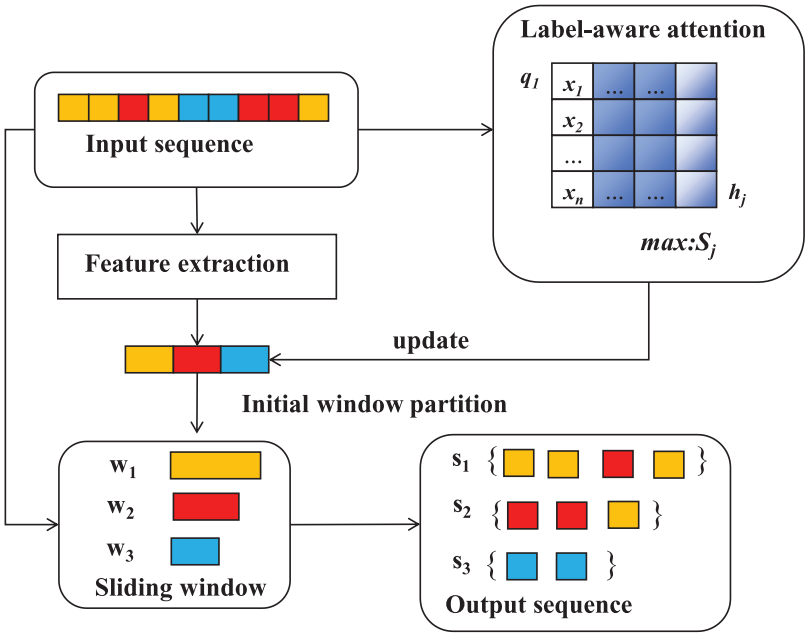

Figure 3: Adaptive sliding window mechanism

By calculating attention scores between potential labels and traffic subsequences, the system dynamically adjusts both the window length and the starting position. This optimization process ensures that only the most discriminative local features associated with target websites are preserved. In Eq. (1), S denotes a set of n sliding windows, W_i represents the i-th sliding window, and n is the total number of windows. In Eq. (2), let (c, y) represent a client-side web session instance, where c is a direction sequence of length l, and y is the label vector associated with this sequence. For instance, if the traffic sequence corresponds to the i-th monitored website, then $y_i = 1$; otherwise, $y_i = 0$.

$$S = \{W_1, \ldots, W_n\}. \tag{1}$$

$$Wi = \{c[i + j\alpha : i + j\alpha + \alpha]\}, \forall j \in [0, [l/\alpha] - 1]. \tag{2}$$

After positioning the traffic, the model introduces a label-aware attention mechanism to dynamically capture the matching relationship between each traffic segment and its corresponding candidate label. Each segmented traffic fragment is represented by h_j, while q_i denotes the learnable query vector for the i-th label. The attention weights between each label and the corresponding segments are calculated using learnable projection matrices, W_q and W_k, which are optimized during training, as shown in Eq. (3). This results in a label-specific traffic-aware representation. Subsequently, a softmax function is applied to normalize the attention distribution across all segments, as demonstrated in Eq. (4).

$$f(w_{i,j}) = \frac{(W_q q_i)^T (W_k h_j)}{\sqrt{d}} + b_{i,j} \tag{3}$$

$$\alpha_{i,j} = \frac{\exp\left(\frac{q_i^T \cdot f(w_j)}{\sqrt{d}}\right)}{\sum_{j'=1}^{n} \exp\left(\frac{q_i^T \cdot f(w_{j'})}{\sqrt{d}}\right)}. \tag{4}$$

In the above equation, q_i denotes the query vector associated with label y_i, and $f\ (w_j)$ represents the key representation of the j-th segment after linear projection. The computed value indicates the attention weight of label y_i over the j-th segment.

To better capture discriminative patterns under varying traffic dynamics, this study further introduces an adaptive window adjustment mechanism. In contrast to the use of a uniform fixed-size window across all traffic sequences, the window length for each segment is dynamically determined, taking into account both the attention entropy and the confidence derived from its label-aware matching scores. Formally, the adaptive window length for the j-th segment is defined in Eq. (5), as follows:

$$L_j = L_0 + \alpha \cdot H\left(a_j\right) + \beta \cdot \left(1 - \max a_{ji}\right) \tag{5}$$

In Eq. (5), L_0 represents the initial base window size, and α and β are scaling factors. The term $H(a_j) = -\sum a_{ji} \log a_{ji}$ represents the entropy of the attention weight distribution for segment j. The term $\max a_{ji}$ indicates the maximum attention confidence for all candidate labels, meaning that $1 - max\ a_{ji}$ penalizes segments with lower certainty. To prevent excessive window expansion, a boundary constraint is applied, as described in Eq. (6), where L_{min} and L_{max} are hyperparameters controlling the lower and upper bounds of the window size.

$$L_j^* = \min\left(L_{\max}, \max\left(L_{\min}, L_j\right)\right) \tag{6}$$

To further evaluate the significance of each segment within the overall label space, this study computes the maximum value along the label dimension of the label-aware matching score matrix, as shown in Eq. (7). This computation results in a segment score, S_j, which serves as a critical reference for guiding subsequent feature selection and modeling. However, discarding segments with low scores may result in the loss of valuable information, as these segments could still contain discriminative features for rare labels. To address this, a reweighting compensation mechanism is introduced. By setting a threshold θ and a compensation factor λ, as defined in Eq. (8), the model retains partial information from weaker segments to a certain degree. All features undergo label-aware attention fusion before being fed into the decoder.

$$S_j = \max \mathrm{f}\left(\mathrm{w_{i,j}}\right). \tag{7}$$

$$\tilde{h} = \begin{cases} \mathrm{h}_j, \text{if } S_j \geq \theta \\ \lambda \cdot h_j, \text{otherwise } S_j < \theta \end{cases}. \tag{8}$$

The decoder then receives all segment features, which have been fused through the label-aware attention mechanism, for multi-tab prediction. This approach significantly enhances the model's ability to identify locally discriminative patterns, while also effectively leveraging marginally weak features. As a result, the model achieves greater robustness and broader coverage in multi-tab recognition scenarios.

2.2.2 Website Identification Module

In multi-tab browsing scenarios, the standard self-attention mechanism falls short. It fails to distinguish between legitimate traffic overlap from sequential website visits and artificial noise introduced by fingerprinting defenses. Self-attention uses a fully connected attention layer, where the output of each input vector is determined by its correlation with all other input vectors. As a result, noisy traffic features inevitably reduce the accuracy of the final results.

To overcome this challenge, this paper introduces a module that incorporates multi-scale feature extraction. The module consists of two main components: global feature modeling and feature fusion. The

architecture of the module is depicted in Fig. 4. Specifically, the model utilizes two distinct branches, T1 and T2, to capture temporal features of traffic across varying time scales. Each branch focuses on extracting traffic characteristics from different window positions, enabling a more comprehensive analysis of the data. The temporal features extracted by these two branches are then merged, forming a more comprehensive representation. The temporal features are further processed with one-dimensional convolution to enhance the model's ability to capture long-term dependencies in the traffic data. Meanwhile, spatial features are handled using two-dimensional convolution, enabling the model to capture distribution patterns and spatial relationships within the traffic data. As a result, the model is adept at learning both temporal and spatial information embedded in the traffic.

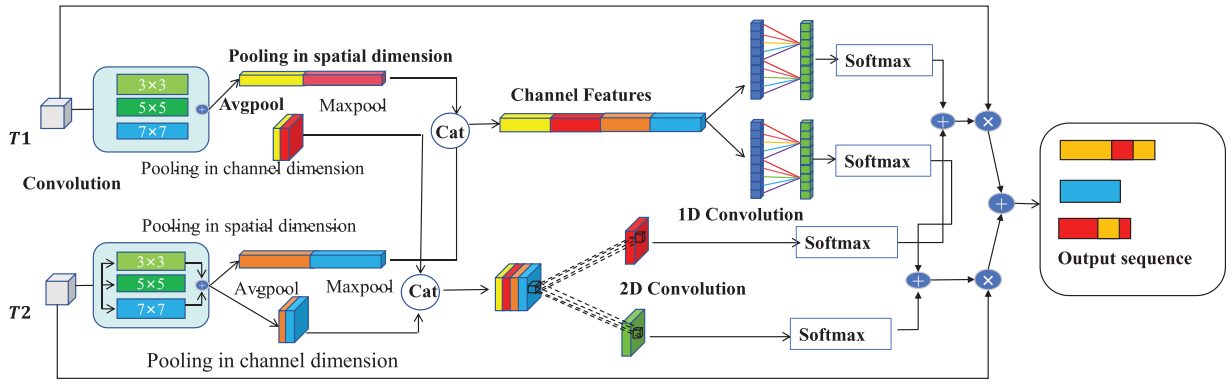

Figure 4: Architecture of the TFAM module

In the feature fusion stage, a Temporal Fusion Attention Module (TFAM) is introduced and adopted. This module combines local features using both channel attention and spatial attention, providing enhanced inputs for global modeling. The channel attention module initially computes the importance weights along the channel dimension by applying both global average pooling and global max pooling. As illustrated in Eq. (9), average pooling captures the overall trend of the features, whereas max pooling emphasizes the most prominent components. The pooled results are then fused and processed through a one-dimensional convolution layer, which compresses the dimensionality further, ultimately generating a weight distribution for each channel.

$$W_c = \text{Softmax}(\text{Conv1d}(\text{Concat}(\text{AvgPool}(x), \text{MaxPool}(x)))). \tag{9}$$

In addition, the spatial attention module, as described in Eq. (10), analyzes the distribution of fragment features along the spatial dimension to generate local spatial weighting coefficients. It computes both the mean and maximum values of the fragments across this dimension and integrates this information through a two-dimensional convolutional network, producing a spatial attention matrix. This matrix dynamically adjusts the spatial distribution of the input features, enabling the model to effectively capture local spatial characteristics related to target website patterns, while suppressing noise.

$$W_s = \text{Softmax}(\text{Conv2d}(\text{Concat}(\text{Mean}(x), \text{Max}(x)))). \tag{10}$$

Finally, to prevent overfitting, the model employs regularization techniques and uses a multi-layer perception to further process the extracted features, outputting the final multi-class classification results.

2.2.3 Incremental Learning Module

This work introduces an incremental learning approach to address concept drift, enabling continuous adaptation to evolving data distributions while mitigating catastrophic forgetting. Initially, the model is trained on a closed-world dataset to capture the baseline traffic patterns of monitored websites. It is worth noting here that iterative refinement yields a stable teacher model that retains foundational knowledge, which is then preserved for subsequent incremental updates. The complete architecture of this process is illustrated in Fig. 5.

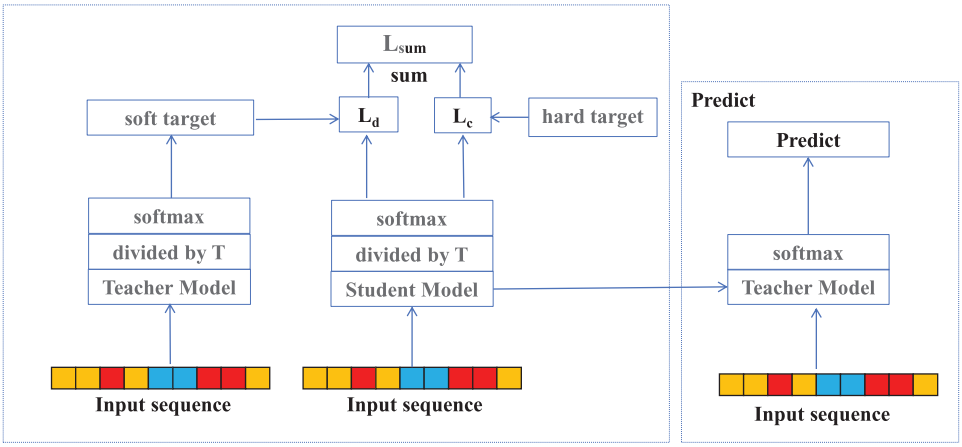

Figure 5: Incremental learning module

The second phase leverages a knowledge distillation framework to train a student model capable of recognizing emerging website categories in open-world scenarios. As formulated in Eqs. (11) and (12), the model optimizes a composite loss function that integrates both a classification loss, which is responsible for acquiring new category knowledge, and a distillation loss, which preserves previously learned representations through the teacher model's soft labels. As expressed in Eq. (13), adjustable loss coefficients control the knowledge transfer balance, allowing fine-grained regulation of learning stability. To address the data imbalance typically caused by the dominance of unmonitored websites, the training process strategically combines closed-world and open-world traffic samples. This approach allows the model to dynamically adapt to evolving data distribution without the need for complete retraining. It preserves existing knowledge while enabling the rapid integration of new categories. At the same time, it effectively mitigates the effects of concept drift.

$$L_{\mathrm{d}} = KL\left(\frac{\exp\left(z_{teacher}/T\right)}{\sum \exp\left(z_{teacher}/T\right)} \middle\| \frac{\exp\left(z_{student}/T\right)}{\sum \exp\left(z_{student}/T\right)}\right). \tag{11}$$

$$Lc = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C}\left(\mathrm{y}ij\log\left(pij\right) + \left(1 - yij\right)\log\left(1 - pij\right)\right). \tag{12}$$

$$L\mathrm{sum} = \alpha L\mathrm{d} + \left(1 - \alpha\right)Lc. \tag{13}$$

The knowledge distillation loss L_d quantifies the output divergence between the output distributions of the teacher and student models using the Kullback—Leibler (KL) divergence with the temperature parameter T controlling the smoothness of the teacher's soft labels through temperature scaling. The multi-class cross-entropy loss L_c is then combined with L_d using a balancing coefficient α, which governs the degree of knowledge integration between new and previously learned categories. In Eqs. (12) and (13), N denotes the

number of samples, C the total number of classes, and y_{ij} and p_{ij} represent the ground truth and predicted probability for the j-th class of the i-th sample, respectively. The class index symbol has been unified as k throughout the paper for consistency.

3 Dataset and Metric

3.1 Dataset

This experiment uses a real-world multi-tab traffic dataset collected by Meng et al. [28], which offers greater realism and complexity compared to simulated datasets such as those by Jin et al. [24]. The dataset captures genuine browsing behaviors with varying numbers of concurrently opened tabs on Alexa Top websites. This setup accurately reflects website fingerprinting characteristics in realistic multi-tab browsing scenarios. Importantly, the dataset also incorporates realistic network conditions, including active traffic defense mechanisms, ensuring that the evaluation closely mirrors real-world operating environments. The complete dataset specifications are summarized in Table 2.

Table 2: Dataset description

Number of websites	Number of tabs	Description
100	2 to 5-tab	The closed-world evaluation selects k random websites from the Alexa Top 100 list. Automated k-tab browsing is conducted with randomized access intervals ($k \in [2, 5]$)
20000	2 to 5-tab	The open-world dataset includes k-l monitored (Alexa Top 100) and l unmonitored (Alexa Top 20K) websites. Automated sequential browsing with random intervals simulates realistic k-tab activity ($k \in [2, 5]$)
100	2-tab	Three defense mechanisms—WTF-PAD, Front, and Tamaraw—are implemented within the Tor network using WFDefProxy [34]

The experiment carried out in this work evaluates the proposed method against three baseline models: BAPM [25], AMWF [36], and ARES [23]. Table 3 provides detailed information on the model configurations.

Table 3: Model configurations

Module	Parameter	Value
Network traffic division module	Number of sliding windows	5
	Sliding window size	1000
	Input traffic sequence length	5000
	Number of blocks	3
	Number of kernels	[64, 128, 256]
	Kernel size	[8, 8, 8]
	Pool size	[8, 8, 8]
	Output dimension	256

(Continued)

Table 3 (continued)

Module	Parameter	Value
Website identification module	Channel attention dimension	512
	Feedforward network dimension	1024
	Multi-scale convolution sizes	[3, 5, 7]
Incremental learning module	Learning rate	0.0012
	Learning rate update period	30 epochs
	Distillation temperature coefficient	3
	Batch size	64
	Traffic sequence length	10000

3.2 Metric

The experiment conducted in this work adopts three standard multi-label evaluation metrics to assess model performance. Area Under the Curve (AUC) [39] is used to evaluate the model's ability to discriminate between different websites. The Precision at K (P@K) metric measures the accuracy of the model's top-K predictions. The Mean Average Precision at K (MAP@K) [40] provides a comprehensive assessment of the model's overall ranking performance. The corresponding mathematical formulations for these metrics are presented in Eqs. (14) and (15):

$$P@K = \frac{1}{k} \sum_{l \epsilon rk(\bar{y})} y_l. \tag{14}$$

$$MAP@K = \frac{\sum_{i=1}^{k} P@i}{k}. \tag{15}$$

4 Analysis and Discussion

4.1 Multi-Tab Website Fingerprint Attack in a Closed-World Environment

To validate the effectiveness of the proposed method [41], this study examines and compares MWF-IL with existing multi-tab WF attack models, including BAPM, ARES, and AMWF. The dataset is partitioned into 90% for training, 5% for validation, and 5% for testing to ensure both reliability and reproducibility of the experimental results. To highlight the advantages of the multi-tab approach over traditional single-tab methods, the proposed model is also evaluated against the classic single-tab model, DF. In addition to AUC and P@K, this study employs Precision, Recall, and F1-score to comprehensively assess the classification performance of the models. This ensures rigorous evaluation and fair comparison across all models.

In addition, the experiment primarily uses AUC and MAP@K to compare the proposed model with existing approaches, including AMWF, BAPM, DF, and ARES. AUC offers a holistic assessment of classification performance across all thresholds, indicating the model's ability to distinguish positive from negative samples. Since BAPM only outputs the top-K predicted websites and does not provide a full ranking needed for AUC calculation, it is excluded from AUC-based comparisons. As shown in Fig. 6, under a closed-world setting, AUC trends indicate that the proposed method consistently outperforms the baseline models across varying numbers of concurrent tabs. In the 2-tab scenario, MWF-IL achieves an AUC of approximately

0.936, compared to 0.714, 0.821, and 0.907 for AMWF, DF, and ARES, respectively—an improvement of about 16.1%. It is worth noting here that in the most challenging 5-tab environment, MWF-IL maintains a leading AUC of 0.659. These results validate the effectiveness of the label-aware attention mechanism and dynamic window strategy in managing traffic overlap and label interference while ensuring robust generalization.

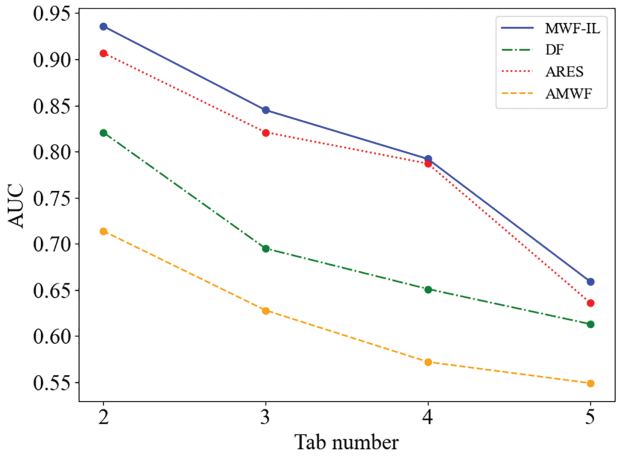

Figure 6: AUC variation across different tab numbers in the closed-world scenario

Furthermore, Fig. 7 presents the MAP@K evaluation results of each model across different multi-tab scenarios (with K representing the number of tabs). As illustrated in the figure, MWF-IL consistently outperforms all baseline models regardless of the number of open tabs. On average, MWF-IL achieves relative improvements of approximately 33.3%, 211.2%, 43.5%, and 4.5% over BAPM, AMWF, DF, and ARES, respectively. These results demonstrate that MWF-IL is particularly effective at distinguishing visited from non-visited websites in multi-tab environments, significantly enhancing the accuracy of identifying actual user activity. Notably, in more complex multi-tab scenarios, MWF-IL exhibits strong stability and robustness, consistently surpassing competing methods in recognizing genuine browsing behavior. This further confirms the model's superior ability to capture traffic features and model global dependencies, highlighting its potential for practical deployment in realistic browsing conditions.

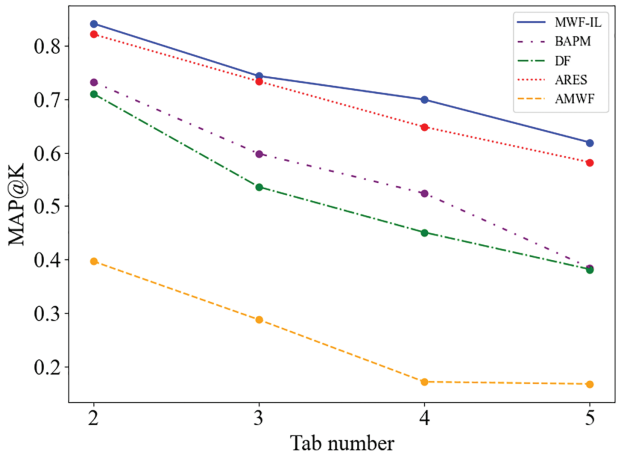

Figure 7: MAP@K variation across different tab numbers in the closed-world scenario

Furthermore, Table 4 presents the performance of various models in terms of Precision, Recall, and F1-score under different multi-tab settings. As shown in the table, the proposed MWF-IL method consistently exhibits significant performance advantages across all tab scenarios. In the 2-tab setting, MWF-IL achieves an F1-score of 0.914, compared to 0.732, 0.418, 0.714, and 0.896 for BAPM, AMWF, DF, and ARES, respectively, corresponding to average improvements of approximately 23.5%, 116.5%, 26.6%, and 2.03%. As the number of tabs increases, the complexity of the multi-tab classification task rises, leading to a general decline in F1-scores for all models. However, MWF-IL exhibits the smallest performance drop, demonstrating superior generalization ability in complex multi-tab scenarios. Under the 5-tab setting, the precision of the MWF-IL model is slightly lower than that of ARES. This is primarily attributed to MWF-IL's adaptive sliding window strategy, which may introduce occasional misclassifications in environments with highly overlapping labels, slightly reducing precision. In contrast, ARES employs a more conservative prediction approach, outputting only labels with high confidence and thereby maintaining higher precision in challenging scenarios.

Table 4: Performance comparison under varying tab settings (closed-world)

Model	2-tab			3-tab		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
MWF-IL	0.914	0.905	0.904	0.909	0.902	0.903
BAPM	0.743	0.736	0.732	0.665	0.652	0.654
AMWF	0.411	0.462	0.418	0.285	0.346	0.277
DF	0.701	0.721	0.714	0.652	0.720	0.668
ARES	0.908	0.893	0.886	0.903	0.889	0.884
Model	**4-tab**			**5-tab**		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
MWF-IL	0.906	0.897	0.898	0.814	0.806	0.808
BAPM	0.625	0.618	0.613	0.515	0.499	0.496
AMWF	0.203	0.304	0.219	0.225	0.287	0.210
DF	0.528	0.651	0.587	0.493	0.582	0.513
ARES	0.898	0.891	0.885	0.816	0.813	0.805

Moreover, P@K is a critical evaluation metric for multi-label classification, measuring the accuracy of predicting the top-K visited websites. Fig. 8 illustrates the P@K performance across different tab settings. MWF-IL consistently outperforms competing models in all scenarios. Under the 2-tab setting, MWF-IL achieves a P@2 of 0.845, while other models remain below 0.65. For the 3-tab and 4-tab settings, MWF-IL maintains P@2 scores of 0.794 and 0.764, respectively, outperforming all competing baseline models. These results indicate that MWF-IL can more accurately predict the top-K visited websites in multi-label environments. It is worth noting that the inherent decrease of P@K with larger K values, as explained in Eq. (12), reflects a property of the metric itself rather than a decline in model performance.

In summary, the experimental results demonstrate that MWF-IL achieves significant performance gains in MT-WF. Its advantages stem from its ability to capture traffic patterns across multiple temporal scales and leverage label-aware attention to distinguish concurrent tab sessions. While closed-world evaluations provide a rigorous baseline, they do not fully capture real-world conditions where users may access numerous unmonitored websites. Consequently, closed-world experiments serve primarily as a reference, whereas open-world evaluations are essential to better examine and assess practical deployment scenarios.

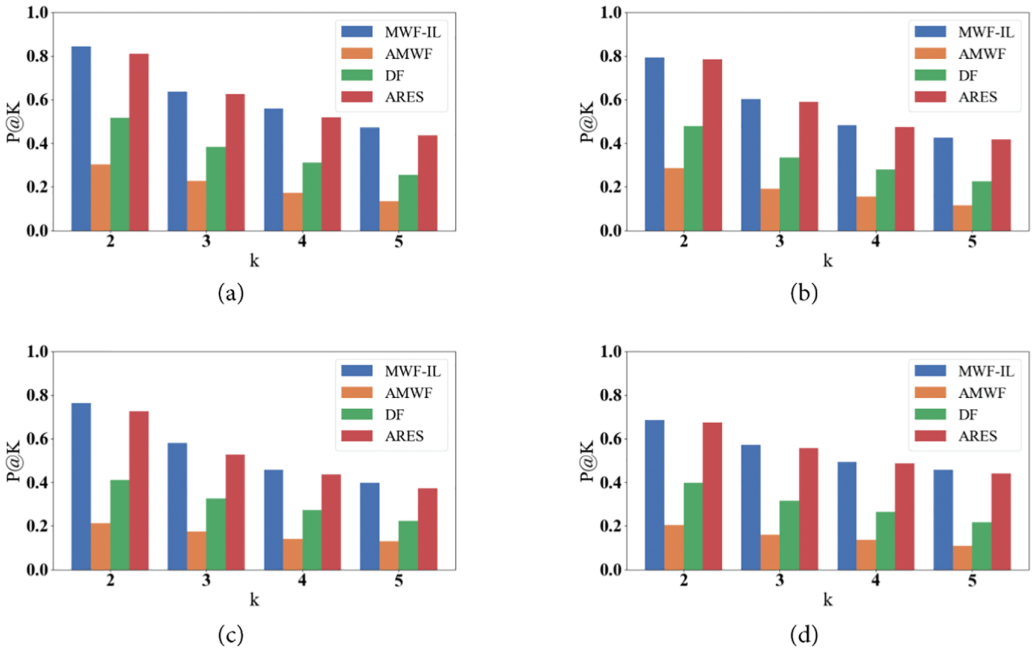

Figure 8: Top-K prediction accuracy across different tab settings: (**a**) 2-tab; (**b**) 3-tab; (**c**) 4-tab; (**d**) 5-tab

4.2 Multi-Tab Website Fingerprint Attack in an Open World Environment

In the closed-world scenario, each monitored website is treated as a distinct category. By contrast, the open-world scenario groups all unmonitored websites into a single class. This approach creates a substantial class imbalance, as the number of instances in the unmonitored class far exceeds that in each monitored class. To mitigate this, instances from both closed and open-world settings are combined while ensuring the same number of browser tabs, thereby balancing the data distribution.

Tables 5 and 6 illustrate the comparative performance of various models under different tab configurations in the open-world scenario. These comparisons are made both before and after applying incremental learning. The results indicate that all models gain some benefit from the incremental learning mechanism. Among them, the MWF-IL model exhibits the most pronounced improvements. For instance, in the 2-tab setting, the P@2 of the MWF-IL model increases from 0.754 to 0.826. Simultaneously, the MAP@2 improves from 0.830 to 0.871, reflecting relative gains of approximately 9.5% and 4.9%, respectively. In contrast, the performance enhancements of other models, such as BAPM and DF, are comparatively modest. For example, BAPM's P@2 improves by only 0.6%. Although AMWF and DF demonstrate some improvement, their overall performance remains significantly lower than that of MWF-IL. Additionally, in the 3-tab and 4-tab scenarios, the MAP@2 of MWF-IL increases from 0.739 to 0.756 and from 0.702 to 0.712, respectively. While the magnitude of improvement gradually diminishes, with more tabs, MWF-IL consistently maintains the leading performance. This trend highlights the model's strong generalization and continual learning capabilities. Even in the most challenging 5-tab scenario, where overall performance declines for all models, MWF-IL shows measurable gains. Its P@2 rises from 0.429 to 0.433, and MAP@2 increases from 0.582 to 0.586. Although these improvements are relatively modest, MWF-IL still demonstrates superior adaptability and robustness compared to other methods.

Table 5: Metric variations under different page settings in the open-world scenario

Tab number	Metric	MWF-IL	BAPM	AMWF	DF	ARES
2-tab	P@2	0.754	0.625	0.275	0.463	0.750
	MAP@2	0.830	0.662	0.354	0.620	0.822
3-tab	P@2	0.588	0.472	0.187	0.309	0.579
	MAP@2	0.739	0.571	0.265	0.472	0.735
4-tab	P@2	0.523	0.425	0.152	0.243	0.528
	MAP@2	0.702	0.514	0.219	0.405	0.711
5-tab	P@2	0.429	0.323	0.127	0.192	0.434
	MAP@2	0.582	0.409	0.189	0.327	0.599

Table 6: Metric variations after incremental learning in the open-world scenario

Tab number	Metric	MWF-IL	BAPM	AMWF	DF	ARES
2-tab	P@2	0.826	0.629	0.288	0.485	0.751
	MAP@2	0.871	0.673	0.375	0.651	0.822
3-tab	P@2	0.594	0.483	0.198	0.325	0.582
	MAP@2	0.756	0.579	0.277	0.504	0.737
4-tab	P@2	0.535	0.424	0.157	0.248	0.521
	MAP@2	0.712	0.509	0.230	0.414	0.710
5-tab	P@2	0.433	0.318	0.128	0.192	0.421
	MAP@2	0.586	0.402	0.192	0.326	0.587

In summary, incorporating incremental learning into MWF-IL significantly enhances overall recognition accuracy. The model maintains stable performance even in challenging scenarios with a high number of tabs. This demonstrates MWF-IL's strong adaptability to dynamic environments and its resilience to concept drift. The results further validate the effectiveness of the incremental learning strategy for open-world, MT-WF attacks.

This study designs experiments to investigate the impact of concept drift using a dataset collected from 200 websites over multiple time spans: 3 days, 10 days, 4 weeks, 6 weeks, and 8 weeks. The dataset allows evaluation of how concept drift affects model performance across different temporal intervals.

Fig. 9 illustrates the variations in AUC over time for different methods when exposed to concept drift. AUC serves as a key metric for assessing a model's overall discriminative ability. Higher AUC values indicate a stronger capacity to distinguish between positive and negative samples across various decision thresholds, reflecting superior recognition performance. As shown in the figure, MWF-IL consistently achieves the highest AUC across all time spans. For example, on day 3, MWF-IL attains an AUC of 0.896, which is significantly higher than DF (0.782), ARES (0.794), and AMWF (0.705). As the observation period extends to day 56, the AUC values of all methods decline, but the rate of decline varies significantly. MWF-IL experiences only a minor decrease, reaching an AUC of 0.690, which demonstrates strong robustness against concept drift. In contrast, DF and ARES drop to 0.571 and 0.564, respectively, while AMWF falls to the lowest value of 0.523. In summary, integrating an incremental learning mechanism provides MWF-IL with

enhanced robustness and adaptability in the presence of concept drift. This approach achieves superior short-term recognition performance and also maintains high stability and practical applicability in long-term monitoring scenarios.

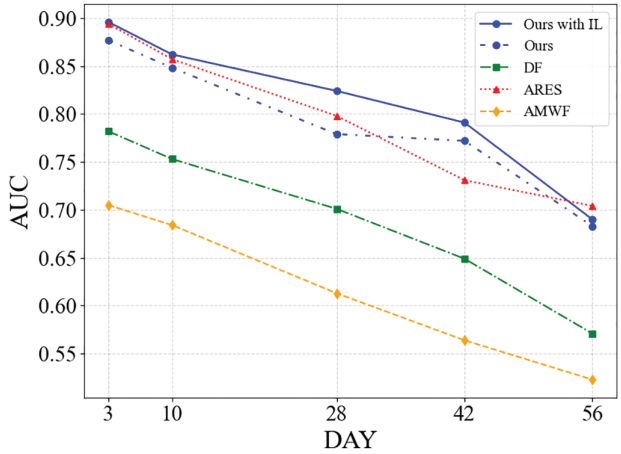

Figure 9: Comparison of AUC values for different methods under the influence of concept drift

To further validate the effectiveness of the proposed method in inferring visited websites under WF attacks, an additional set of experiments was conducted in the open-world scenario. The evaluation specifically focused on the 2-tab and 5-tab cases to capture both moderate and highly complex browsing conditions. The 2-tab setting represents a relatively simple yet realistic scenario, with limited traffic overlap between websites. In contrast, the 5-tab setting represents one of the most challenging conditions, characterized by substantial traffic overlap and feature redundancy across multiple websites. This selection demonstrates the scalability of the proposed method across varying levels of task complexity.

Moreover, as shown in Table 7, the MWF-IL model consistently outperforms other state-of-the-art baseline methods. In the 2-tab scenario, MWF-IL achieves the highest Precision (0.905), Recall (0.901), and F1-score (0.902), clearly surpassing ARES (0.879) and DF (0.709). These results indicate that MWF-IL can accurately disentangle overlapping traffic and extract discriminative features even when multiple websites are active simultaneously. In the more challenging 5-tab setting, the performance of all models decreases due to severe traffic mixing. Despite this, MWF-IL maintains superior results, achieving an F1-score of 0.795, which is significantly higher than ARES (0.779) and DF (0.493). This demonstrates that MWF-IL retains strong adaptability and robustness under high-overlap, multi-tab conditions. These findings confirm its practical applicability in real-world anonymous communication environments.

Table 7: Performance comparison under 2-tab and 5-tab settings (open-world scenario)

Model	2-tab			5-tab		
	Precision	**Recall**	**F1-score**	**Precision**	**Recall**	**F1-score**
MWF-IL	0.905	0.901	0.902	0.798	0.794	0.795
BAPM	0.729	0.723	0.719	0.493	0.472	0.466
AMWF	0.409	0.453	0.410	0.205	0.207	0.202
DF	0.693	0.715	0.709	0.483	0.542	0.493
ARES	0.901	0.887	0.879	0.789	0.787	0.779

Overall, the experiments conducted in this study demonstrate that the MWF-IL model achieves superior recognition accuracy and maintains stable performance in challenging multi-tab scenarios. These improvements are largely attributable to the incremental learning mechanism, which allows the model to adapt effectively to evolving data. Beyond the standard open-world setting, in which all unmonitored websites are grouped into a single class, the study further evaluates the model under more realistic conditions. This includes expanding the number of unmonitored websites and incorporating long-term data collection to simulate the effects of concept drift. These additional experiments highlight the adaptability of MWF-IL. The model not only manages severe class imbalance but also preserves high discriminative ability as the data distribution changes over time. These results indicate that MWF-IL is well-suited for practical application scenarios, where open-world dynamics and temporal variations are inevitable.

4.3 Multi-Tab Website Fingerprinting Attack against Defenses

This study further assesses the model's performance under three representative defense mechanisms: WTF-PAD, FRONT, and Tamaraw. It is worth noting here that WTF-PAD is a lightweight defense that uses randomized padding to protect traffic patterns. It is characterized by low computational overhead and minimal latency, representing strategies commonly adopted in practical applications. FRONT, in contrast, confuses attacker observation by modifying the order and timing of traffic packets. This approach provides a balanced trade-off between overhead and defensive effectiveness. Nevertheless, Tamaraw is a heavyweight defense mechanism that transmits padded traffic at a fixed rate, effectively eliminating predictable traffic features. However, its high latency and bandwidth consumption restrict its applicability in real-world settings. Including these three defenses allows a comprehensive assessment of the model's adaptability and robustness. It also ensures that the experimental results are both comparable and credible across lightweight and heavyweight defense settings.

As shown in Table 8, the performance of each model is evaluated under different defense mechanisms (WTF-PAD, FRONT, and Tamaraw). The abbreviation "Pre" refers to precision in the table. The experimental results indicate that MWF-IL consistently outperforms all baseline models across every defense scenario. When using the WTF-PAD defense, MWF-IL achieves an AUC of 0.852. Its MAP@2 and Precision reach 0.719 and 0.861, respectively. These results represent improvements of 1.1% and 0.35% over ARES and more than 27% higher than both DF and BAPM. Under the FRONT defense, MWF-IL attains an AUC of 0.765, with MAP@2 and Precision of 0.632 and 0.796, respectively. This corresponds to gains of 0.8% in MAP@2 and 0.51% in Precision over ARES. Moreover, MWF-IL achieves approximately 205.3% and 45.3% higher Precision compared to AMWF and BAPM, respectively. Even when subjected to Tamaraw, a more sophisticated defensive mechanism, MWF-IL demonstrates notable resilience. It attains an AUC of 0.605 and a Precision of 0.654. This reflects a relative improvement of 64.3% and 48.9% over BAPM and DF, respectively, although it represents a 3.5% decrease compared to ARES. It is noted here that the observed performance decline is attributed to Tamaraw's strategy of standardizing packet sizes and enforcing constant transmission rates. This approach suppresses temporal dynamics and burst patterns in traffic, reducing the model's ability to capture fine-grained spatio-temporal features. Additionally, MWF-IL relies on interactive modeling between temporal and directional channels to extract discriminative features. Consequently, Tamaraw's suppression of these dynamics diminishes the model's effectiveness. Furthermore, Tamaraw's traffic regularization and the injection of random noise disrupt the co-occurrence structure among labels. This interference negatively affects the performance of the multi-label learning strategy employed by MWF-IL. All relative improvements mentioned above are calculated as the percentage increase over the corresponding baseline using the formula: ((MWF-IL value − Baseline value)/Baseline value) $*$ 100%.

Table 8: Model performance under various defense mechanisms

Model	WTF-PAD			FRONT			Tamaraw		
	AUC	MAP@2	Pre	AUC	MAP@2	Pre	AUC	MAP@2	Pre
MWF-IL	0.852	0.719	0.861	0.765	0.632	0.796	0.605	0.298	0.654
BAPM	N/A	0.563	0.672	N/A	0.435	0.607	N/A	0.152	0.398
AMWF	0.574	0.245	0.449	0.549	0.207	0.384	0.517	0.125	0.300
DF	0.694	0.499	0.629	0.587	0.341	0.522	0.548	0.201	0.439
ARES	0.839	0.711	0.858	0.761	0.627	0.792	0.613	0.326	0.678

In conclusion, the proposed methodology effectively captures fine-grained traffic features, even in the presence of strong defense perturbations. It also constructs more accurate website fingerprints by leveraging locally stable traffic patterns and reducing noise. While defense mechanisms such as WTF-PAD, Walkie-Talkie, and Tamaraw significantly degrade the performance of website fingerprinting models, MWF-IL consistently outperforms baseline methods. Importantly, when evaluated in an open-world setting, the results indicate that MWF-IL remains effective under conditions that more closely resemble real-world deployment. In such scenarios, adversaries must contend simultaneously with unknown websites and active defensive strategies. These findings highlight the practical applicability and robustness of the proposed approach in real-world anonymous communication environments.

5 Conclusion

This paper presents a multi-tab WF attack method that leverages adaptive sliding window segmentation and a spatio-temporal attention mechanism. The approach addresses the challenge of identifying mixed traffic from multiple tabs in the Tor network. The adaptive sliding window module dynamically adjusts the length and position of segments based on traffic patterns, enhancing the accuracy of local pattern extraction. The label-aware attention mechanism aligns traffic segments with their corresponding labels, focusing on regions that are highly correlated with website-specific features. Additionally, a feature compensation mechanism improves the utilization of weak segments, further enhancing recognition performance. Additionally, an incremental learning strategy based on knowledge distillation is introduced. This allows the model to efficiently adapt to newly emerging labels in an open-world setting while maintaining stable recognition performance for previously learned classes. Experimental results show that the proposed method achieves superior recognition accuracy and robustness under closed-world, open-world, and defense-perturbed scenarios. It also demonstrates a notable degree of adaptability to traffic variations caused by concept drift. These findings confirm the practical potential of the method in real-world anonymous communication environments. However, several challenges remain. Parameter interference between old and new labels can occur, and the model adapts slowly to distribution shifts during continual learning under concept drift.

While the approach demonstrates promising performance, the current study identifies three key limitations that require further investigation. First, the model exhibits reduced robustness under severe class imbalance and when encountering a large number of unknown labels. Second, semantic relationships between labels are underutilized, sometimes producing redundant predictions. Third, the incremental learning component is still susceptible to catastrophic forgetting during prolonged continual learning. Future research should focus on addressing these limitations. Potential directions include: (a) modeling label dependency graphs, (b) hierarchical semantic feature fusion, and (c) meta-reinforced incremental

learning strategies. These improvements aim to enhance adaptability and resilience in dynamic network environments.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding.

Author Contributions: The authors confirm their contribution to the paper as follows: Conceptualization, Chunqian Guo and Gang Chen; methodology, software, validation, formal analysis, investigation, resources, and data curation, Chunqian Guo; writing—original draft, Chunqian Guo; writing—review and editing, Chunqian Guo and Gang Chen; visualization, Chunqian Guo; supervision, project administration, and funding acquisition, Gang Chen. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The data supporting the findings of this study are available from the corresponding author, Chunqian Guo, upon request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Pan Y, Ling Z, Zhang Y, Wang H, Liu G, Luo J, et al. TORCHLIGHT: shedding LIGHT on real-world attacks on cloudless IoT devices concealed within the tor network. arXiv:2501.16784. 2025. Available from: https://arxiv.org/abs/2501.16784.
2. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Commun ACM. 2017;60(6):84–90. doi:10.1145/3065386.
3. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/CVPR.2016.90.
4. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020. Available from: https://arxiv.org/abs/2010.11929.
5. Xu SJ, Geng GG, Jin XB, Liu DJ, Weng J. Seeing traffic paths: encrypted traffic classification with path signature features. IEEE Trans Inform Forensic Secur. 2022;17:2166–81. doi:10.1109/tifs.2022.3179955.
6. Devlin J, Chang M, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Vol. 1 Long and Short Papers. Minneapolis, MN, USA: Association for Computational Linguistics; 2019. p. 4171–86.
7. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
8. Zhang Y, Song J, Sun Y, Gao Z, Hu Z, Sun M. Federated two-stage transformer-based network for intrusion detection in non-IID data of controller area networks. Cybersecurity. 2025;8(1):29.
9. Kim T, Pak W. Deep learning-based network intrusion detection using multiple image transformers. Appl Sci. 2023;13(5):2754. doi:10.3390/app13052754.
10. Sirinam P, Imani M, Juarez M, Wright M. Deep fingerprinting: undermining website fingerprinting defenses with deep learning. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security; 2018 Oct 15–19; Toronto, ON, Canada.
11. Panchenko A, Lanze F, Zinnen A, Henze M, Pennekamp J, Wehrle K, et al. Website fingerprinting at internet scale. In: Proceedings of the 2016 Network and Distributed System Security Symposium; 2016 Feb 21–24; San Diego, CA, USA.

12. Sirinam P, Mathews N, Rahman MS, Wright M. Triplet fingerprinting: more practical and portable website fingerprinting with N-shot learning. In: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London, UK: ACM; 2019. p. 1131–48. doi:10.1145/3319535.3354217.
13. Bhat S, Lu D, Kwon A, Devadas S. Var-CNN: a data-efficient website fingerprinting attack based on deep learning. arXiv:1802.10215. 2018. Available from: https://arxiv.org/abs/1802.10215.
14. Rahman MS, Sirinam P, Mathews N, Gangadhara KG, Wright M. Tik-Tok: the utility of packet timing in website fingerprinting attacks. arXiv:1902.06421. 2019. Available from: https://arxiv.org/abs/1902.06421.
15. Hayes J, Danezis G. K-fingerprinting: a robust scalable website fingerprinting technique. In: Proceedings of the 25th USENIX Conference on Security Symposium; 2016 Aug 10–12; Austin, TX, USA.
16. Wang T, Cai X, Nithyanand R, Johnson R, Goldberg I. Effective attacks and provable defenses for website fingerprinting. In: Proceedings of the 23rd USENIX Security Symposium (USENIX Security 14); 2014 Aug 20–22; San Diego, CA, USA. p. 143–57.
17. Abe K, Goto S. Fingerprinting attack on Tor anonymity using deep learning. Proc Asia-Pac Adv Netw. 2016;42:15–20.
18. Luo C, Tang W, Wang Q, Zheng D. Few-shot website fingerprinting with distribution calibration. IEEE Trans Dependable Secure Comput. 2025;22(1):632–48. doi:10.1109/tdsc.2024.3411014.
19. Chen M, Chen Y, Wang Y, Xie P, Fu S, Zhu X. End-to-end multi-tab website fingerprinting attack: a detection perspective. arXiv:2203.06376. 2022. Available from: https://arxiv.org/abs/2203.06376.
20. Chen ZM, Wei XS, Wang P, Guo Y. Multi-label image recognition with graph convolutional networks. In: Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2019 Jun 15–20; Long Beach, CA, USA. p. 5172–81. doi:10.1109/cvpr.2019.00532.
21. Yu B, Xie H, Cai M, Ding W. MG-GCN: multi-granularity graph convolutional neural network for multi-label classification in multi-label information system. IEEE Trans Emerg Top Comput Intell. 2024;8(1):288–99. doi:10.1109/tetci.2023.3300303.
22. Yan Y, Liu J, Zhang BW. LabelCoRank: revolutionizing long tail multi-label classification with co-occurrence reranking. arXiv:2503.07968. 2025. Available from: https://arxiv.org/abs/2503.07968.
23. Deng X, Yin Q, Liu Z, Zhao X, Li Q, Xu M, et al. Robust multi-tab website fingerprinting attacks in the wild. In: 2023 IEEE Symposium on Security and Privacy (SP); 2023 May 21–25; San Francisco, CA, USA. p. 1005–22. doi:10.1109/sp46215.2023.10179464.
24. Jin Z, Lu T, Luo S, Shang J. Transformer-based model for multi-tab website fingerprinting attack. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security; 2023 Nov 26–30; Copenhagen, Denmark. p. 1050–64. doi:10.1145/3576915.3623107.
25. Guan Z, Xiong G, Gou G, Li Z, Cui M, Liu C. BAPM: block attention profiling model for multi-tab website fingerprinting attacks on Tor. In: Proceedings of the 37th Annual Computer Security Applications Conference; 2021 Dec 6–10; Virtual. p. 248–59. doi:10.1145/3485832.3485891.
26. Zhao X, Deng X, Li Q, Liu Y, Liu Z, Sun K, et al. Towards fine-grained webpage fingerprinting at scale. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security; 2024 Oct 24–28; Salt Lake City, UT, USA. p. 423–36. doi:10.1145/3658644.3690211.
27. Yuan Y, Zou W, Cheng G. Attack smarter: attention-driven fine-grained webpage fingerprinting attacks. arXiv:2506.20082. 2025. Available from: https://arxiv.org/abs/2506.20082.
28. Meng W, Ma C, Ding M, Ge C, Qian Y, Xiang T. Beyond single tabs: a transformative few-shot approach to multi-tab website fingerprinting attacks. In: Proceedings of the ACM on Web Conference 2025; 2025 Apr 28–May 2; Sydney, Australia. p. 1068–77. doi:10.1145/3696410.3714811.
29. Rahman MS, Imani M, Mathews N, Wright M. Mockingbird: defending against deep-learning-based website fingerprinting attacks with adversarial traces. IEEE Trans Inform Forensic Secur. 2021;16:1594–609. doi:10.1109/tifs.2020.3039691.
30. Juarez M, Imani M, Perry M, Diaz C, Wright M. Toward an efficient website fingerprinting defense. In: Proceedings of the 21st European Symposium on Research in Computer Security; 2016 Sep 26–30; Heraklion, Greece. p. 27–46. doi:10.1007/978-3-319-45744-4_2.

31. Gong J, Wang T. Zero-delay lightweight defenses against website fingerprinting. In: Proceedings of the 29th USENIX Security Symposium (USENIX Security 20); 2020 Aug 12–14; Boston, MA, USA. p. 717–34.
32. Yang C, Xiao X, Hu G, Ling Z, Li H, Zhang B. DTPN: a diffusion-based traffic purification network for Tor website fingerprinting. In: Proceedings of the Eighteenth ACM International Conference on Web Search and Data Mining; 2025 Mar 10–14; Hannover, Germany. p. 642–50. doi:10.1145/3701551.3703547.
33. Wickramasinghe N, Shaghaghi A, Tsudik G, Jha S. SoK: decoding the Enigma of encrypted network traffic classifiers. In: 2025 IEEE Symposium on Security and Privacy (SP); 2025 May 12–15; San Francisco, CA, USA. p. 1825–43. doi:10.1109/SP61157.2025.00165.
34. Cui Y, Wang G, Vu K, Wei K, Shen K, Jiang Z, et al. A comprehensive survey of website fingerprinting attacks and defenses in Tor: advances and open challenges. arXiv:2510.11804. 2025. Available from: https://arxiv.org/abs/2510.11804.
35. Ling Z, Xiao G, Wu W, Gu X, Yang M, Fu X. Towards an efficient defense against deep learning based website fingerprinting. In: IEEE INFOCOM 2022—IEEE Conference on Computer Communications; 2022 May 2–5; London, UK. p. 310–9. doi:10.1109/infocom48880.2022.9796685.
36. Yin Q, Liu Z, Li Q, Wang T, Wang Q, Shen C, et al. An automated multi-tab website fingerprinting attack. IEEE Trans Dependable Secure Comput. 2022;19(6):3656–70. doi:10.1109/tdsc.2021.3104869.
37. Cui W, Chen T, Fields C, Chen J, Sierra A, Chan-Tin E. Revisiting assumptions for website fingerprinting attacks. In: Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security; 2019 Jul 9–12; Auckland, New Zealand. p. 328–39. doi:10.1145/3321705.3329802.
38. Wang T, Goldberg I. On realistically attacking Tor with website fingerprinting. Proc Priv Enhancing Technol. 2016;2016(4):21–36. doi:10.1515/popets-2016-0027.
39. Ling CX, Huang J, Zhang H. AUC: a statistically consistent and more discriminating measure than accuracy. IJCAI. 2003;3:519–24.
40. Liu J, Chang WC, Wu Y, Yang Y. Deep learning for extreme multi-label text classification. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval; 2017 Aug 7–11; Tokyo, Japan. p. 115–24. doi:10.1145/3077136.3080834.
41. Li X, Guo J, Song Q, Xie J, Sang Y, Zhao S, et al. Listen to minority: encrypted traffic classification for class imbalance with contrastive pre-training. In: Proceedings of the 20th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON); 2023 Sep 11–14; Madrid, Spain.