Multi-Granularity Feature Fusion for Enhancing Encrypted Traffic Classification

Quan Ding[1], Zhengpeng Zha[2]*, Yanjun Li[3], Zhenhua Ling[4]
State Grid Anhui Electric Power Co. Ltd., Electric Power Science Research Institute, Hefei, China[1]
University of Science and Technology of China, Institute of Advanced Technology, Hefei, China[2,3,4]

***Abstract*—Encrypted traffic classification, a pivotal process in network security and management, involves analyzing and categorizing data traffic that has been encrypted for privacy and security. This task demands the extraction of distinctive and robust feature representations from content-concealed data to ensure accurate and reliable classification. Traditional approaches have focused on utilizing either the payload of encrypted traffic or statistical features for more precise classification. While these methods achieve relative success, their limitation lies in not harnessing multi-grained features, thus impeding further advancements in encrypted traffic classification capabilities. To tackle this challenge, ET-CompBERT is presented, an innovative framework specifically designed for the fusion of multi-granularity features in encrypted traffic, encompassing both payload and global temporal attributes. The extensive experiments reveal that our approach significantly enhances classification performance in data-rich scenarios (achieving up to a +4.43% improvement in certain cases over existing methods) and establishes state-of-the-art results on training sets with different sizes. The source codes will be released after paper acceptance.**

***Keywords*—Encrypted traffic classification; BERT; multi-granularity fusion**

I. Introduction

Recently, the widespread use of traffic encryption has become instrumental in protecting the privacy and anonymity of Internet users [1], [2]. While this advancement is vital for security and confidentiality, it concurrently presents significant challenges to traffic classification. The increasing utilization of privacy-enhancing encryption techniques, such as Tor and VPNs, by both legitimate users and malicious actors, complicates the task of distinguishing benign from harmful traffic. Encrypted traffic classification thus emerges as a crucial tool in this landscape. It enables the identification and mitigation of malware and cybercriminal activities that exploit encryption to bypass surveillance systems, without compromising the privacy and integrity of legitimate communications. This delicate balance between user privacy and cybersecurity underscores the indispensable role of sophisticated traffic classification methodologies in maintaining a secure digital environment.

Traditional cleartext traffic classification methods [3], [4], [5] primarily rely on deep packet inspection, capturing patterns and keywords within data packets from the payload. Nevertheless, the advent of encrypted traffic poses a significant challenge to these methodologies. The inherent unreadability of encrypted traffic renders traditional cleartext classification ineffective. Recent study [6] proposes leveraging unencrypted protocol field information. This approach involves extracting key features such as device type, certificate details, packet size, and temporal characteristics to represent each data flow. However, this strategy has its limitations. In virtual communication networks, these fingerprints are susceptible to tampering, leading to misinterpretation and a consequent failure in accurately classifying encrypted traffic.

The field of machine learning has witnessed rapid advancement, prompting numerous security researchers [7], [8] to explore statistical methods to enhance the accuracy of encrypted traffic classification. Predominantly, these machine-learning approaches for encrypted traffic classification rely on the meticulous selection of handcrafted features, followed by the application of statistical machine-learning algorithms for classification purposes. For instance, Flowprint [8] leverages statistical features of packet sizes to train random forest classifiers, while BIND [7] utilizes statistical features related to temporality. However, we contend that these methods are overly dependent on selecting handcrafted features. Designing universally applicable features that can effectively address the increasing complexity of numerous applications and websites is a challenging endeavor. Moreover, these methods typically provide only a generalized perspective to the algorithm, limiting the coarse-grained capability of encrypted traffic classification. This inherent limitation underscores the need for more adaptable approaches in this rapidly evolving domain.

These limitations have increasingly steered researchers toward adopting end-to-end deep learning methodologies for encrypted traffic classification. The utilization of supervised deep learning for encrypted traffic classification has emerged as a predominant approach, primarily due to its ability to automatically extract discriminative features, thus diminishing the reliance on manual feature design. In previous research, such as DF [9], convolutional neural networks have been employed to autonomously derive representations from raw packet size sequences in encrypted traffic. The remarkable achievements of BERT[10] in the natural language processing domain have inspired analogous advancements in network traffic analysis. ET-BERT[11] introduces a novel network traffic representation, termed BURST, defined as a sequence of temporally adjacent network packets originating from either the request or response in a single session flow. This approach also incorporates a similar learning task, positioning ET-BERT as a pioneering method in applying a pre-train and fine-tune model to encrypted traffic classification. Despite these advancements, it is crucial to acknowledge that current pre-trained methodologies often focus on the payload of encrypted traffic but neglect the global attributes. This oversight leads to models achieving suboptimal accuracy in encrypted traffic

*Corresponding authors.

classification, highlighting a critical area for improvement in this evolving field.

To address the challenge mentioned above, we introduce a novel framework known as Encrypted Traffic comprehensive Bidirectional Encoder Representations from Transformer (ET-CompBERT). As depicted in Fig. 1, we innovatively introduce a multi-grained learning strategy, termed *comprehensive fusion-guided (CFA)* learning. This strategy synergistically combines a fine-grained understanding of encrypted traffic payloads with a broader, coarse-grained analysis, thereby enhancing the overall comprehension of encrypted traffic. To the best of our knowledge, this represents the inaugural effort in integrating multi-grained features for pre-trained encrypted traffic classification. Drawing inspiration from the rapidly evolving field of prompt tuning in both the computer vision [12] and natural language processing [13] communities, we propose the *global-feature-aware* (GFA) learning strategy, significantly enhancing the robust classification on different data sizes capabilities of ET-CompBERT. Our contributions are outlined as follows:

- This study represents a pioneering effort in a novel multi-grained learning approach as comprehensive fusion-guided learning. This innovative strategy enables fusing the global temporal attributes into the extracted representations of encrypted traffic payloads.
- We introduce an innovative GFA learning strategy, that effectively fuses the representations of local traffic payloads and the representations of global temporal attributes for encrypted traffic classification.
- Extensive experimental evaluations demonstrate the effectiveness of our framework. The results indicate that our approach surpasses existing state-of-the-art methods in all tasks. Notably, our model achieves consistent advantages on training sets with different sizes of datasets, underscoring its versatility and robustness.

The remainder of this paper is organized as follows. Section II provides a detailed review of the relevant literature and background information on encrypted traffic classification. Section III describes the methodology applied in our study, including comprehensive fusion-guided learning and global-feature-aware learning. The results are presented and discussed in Section IV, where we analyze comparisons with existing methods, ablation studies, and other analyses. Finally, Section V concludes the paper with a summary of the findings, and implications of our work.

II. Related Work

A. Encrypted Traffic Classification

Encrypted traffic classification aims to discern the services operating behind obfuscated network traffic, thereby enhancing both network service quality and security assurance. Contemporary methodologies in this domain predominantly fall into two principal categories, those grounded in machine learning techniques [7], [14], [15] and those leveraging deep learning paradigms. However, traditional machine learning-based approaches for encrypted traffic classification are often constrained by their dependency on expert-derived feature extraction and selection, which can impede generalization and further development. This limitation has led researchers to gravitate towards end-to-end deep learning Encrypted Traffic Classification methodologies increasingly.

In contrast to methods based on traditional machine learning, deep learning-based approaches offer a comprehensive solution for encrypted traffic classification by autonomously learning feature representations. This shift towards deep learning methodologies enhances robustness and addresses the inherent complexities in encrypted traffic analysis. Wang et al. [16] exemplify this trend by proposing an application of convolutional neural networks (CNN). Their method involves using the initial 784 bytes of each traffic flow as input, enabling the CNN to extract and learn feature representations effectively, thus showcasing the potential of deep learning in this domain. Given the remarkable success of the BERT [10] model within the natural language processing community, researchers [11] are exploring the application of its structural principles in the realm of encrypted traffic classification through learning approaches. However, a common challenge these methods face is their reliance on substantial volumes of labeled data to ensure optimal performance, limiting their applicability to new, unseen classes that diverge from the training dataset. This challenge gives rise to the need for few-shot learning approaches, capable of classifying new encrypted traffic types with a minimal reliance on labeled data, thus presenting a promising solution to these constraints.

In our study, we introduce a novel approach to encrypted traffic classification that diverges from traditional single-granularity pre-trained methods. Our method centers around enhancing a pre-trained encrypted traffic classification model which proposes to enhance encrypted traffic classification by multi-granularity feature fusion. Furthermore, We introduce an innovative fine-tuning method, GFA learning, which empowers the model with robust classification capabilities under different data sample scenarios.

III. Methods

A. Comprehensive Fusion-guided Learning

In the initial phase of our study, we implement comprehensive fusion-guided learning to cultivate ET-CompBERT in Fig. 1, which integrates global temporal attributes into the payload representations of encrypted traffic. Starting with payload encoding, we utilize the BURST structure, identified as a sequence of temporally contiguous network packets emanating from either a request or a response in a single session flow. This structure is employed to precisely depict encrypted traffic, thereby forming the input for our pre-trained ET-BERT model, mirroring the strategy delineated in [11]. This approach culminates in the creation of an encrypted traffic embedding, denoted as e_b. Following this, we utilize the straightforward two-layer Multi-Layer Perceptron (MLP) for projecting the encrypted traffic embedding into f_b.

In the comprehensive fusion-guided learning procedure, for a given piece of encrypted traffic, we utilize both the payload and its global temporal attributes. During the global-feature-aware learning process, the global temporal attributes concatenate a special classification token and the BURST representation as input to the pre-trained ET-CompBERT model. It is important to highlight that the '//' symbol is used to denote

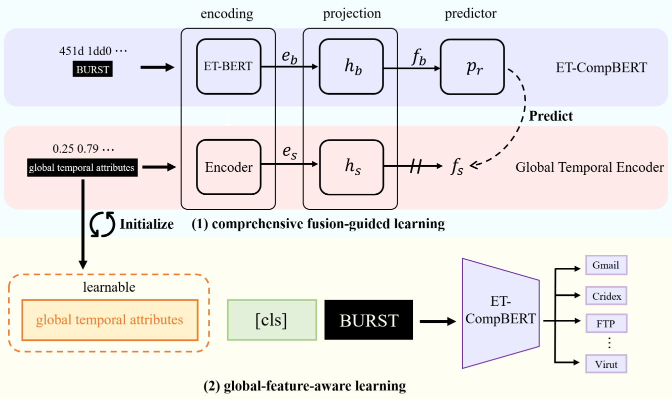

Fig. 1. Framework of ET-CompBERT.

the stop gradient, a critical measure implemented to prevent the model from adopting shortcut learning methods.

To design the global temporal attributes encoding procedure, we employ an encoding strategy akin to that utilized in payload encoding, to acquire embeddings reflective of global temporal properties. Initially, key properties are extracted from network packets, a subset of which is detailed in Table I. These properties are subjected to min-max normalization, resulting in a normalized feature vector. Importantly, we utilize a fully connected network (FCN) to align the dimensions of the global temporal attributes with those of the Transformer Encoder [17], ensuring dimensional compatibility. Following this, the features are replicated as the global temporal attributes and passed through the original Transformer Encoder [17], thereby obtaining the preliminary global temporal embedding e_s. These initial global temporal embeddings are further processed via a two-layer Multi-Layer Perceptron (MLP), culminating in the generation of the final global temporal embedding f_s.

TABLE I. DESCRIPTION OF GLOBAL TEMPORAL ATTRIBUTES

Feature	Description
duration	The duration of the flow.
fiat	Forward Inter Arrival Time (mean, min, max, std).
biat	Backward Inter Arrival Time (mean, min, max, std).
flowiat	Flow Inter Arrival Time (mean, min, max, std).
active	The amount of time a flow was active (mean, min, max, std).
idle	The amount of time a flow was idle (mean, min, max, std).
fb_psec	Flow Bytes per second.
fp_psec	Flow packets per second.

In our pursuit to enhance our framework's capability to identify global attributes of encrypted traffic, while also retaining a profound understanding of its payload, we have developed a comprehensive fusion-guided learning strategy. This innovative approach introduces a novel prediction methodology, meticulously designed to empower the encrypted traffic model with the capability of category classification. Consequently, this facilitates a fundamental enhancement in the fine-grained interpretation of encrypted traffic information, thus enabling a comprehensive and coarse-grained understanding of encrypted traffic dynamics. This strategic design marks a significant advancement in the nuanced analysis of encrypted traffic data. Upon obtaining the encrypted traffic embedding f_b and the global temporal attributes embedding f_s, our framework employs a two-layer Multi-Layer Perceptron (MLP) as a predictor. This MLP is specifically tailored to categorize the global temporal attributes. Given the model's proficiency in acquiring an in-depth understanding of encrypted traffic from both holistic and detailed perspectives, we refer to the ET-BERT, subjected to the aforementioned comprehensive learning, as ET-CompBERT. It is imperative to highlight that our methodology incorporates a stop-gradient strategy designed to prevent the model from gravitating towards shortcut solutions. This critical implementation is pivotal in safeguarding the robustness and integrity of our approach. Such a strategy underpins a more effective learning process, perfectly in sync with the core objectives of encrypted traffic analysis. This meticulous attention to the learning process ensures the reliability and efficacy of our model in challenging scenarios.

B. Global-feature-aware Learning

Following the learning procedure, we fine-tune the ET-CompBERT on downstream datasets. However, conventional learning strategies, while effective under sufficient data conditions, often falter in a few data scenarios. The framework after the comprehensive fusion-guided learning can understand the payload of the encrypted traffic and the global temporal properties. We innovatively introduce the GFA learning strategy.

Initially, global temporal attributes are deployed to initialize the learnable token, akin to their usage in the comprehensive fusion-guided learning paradigm. These attributes are encoded into a token via a shared-weight Fully Connected Network (FCN), ensuring a coherent and efficient representation for subsequent processing. This token is concatenated alongside the special $[class]$ symbol and the BURST to constitute the primary input. To ensure dimensional coherence, we employ a simple one-layer MLP, maintaining the dimensionality adapted to the ET-CompBERT. Through the self-attention mechanism, the global temporal attributes alongside the $[class]$ symbol acquire knowledge from BURST. This process culminates in the formation of the final learnable global temporal attributes, specifically designed to bolster the classification capabilities for encrypted traffic. The effectiveness of this enhancement is continually assessed and refined under the guidance of the cross-entropy loss function. The robust enhancement in encrypted traffic classification performance achieved by our learning and learning methodology is demonstrated in our experimental results. We propose two learning strategies that can also enhance the encrypted traffic classification model for varied classification scenarios.

C. Inference

During the inference phase, we omit the projection and predictor components, retaining only the final global temporal attributes which concatenate with the input to elevate the encrypted traffic classification ability. These attributes encapsulate the optimal encrypted traffic classification capabilities as evidenced by the GFA learning results. In the final layer of the ET-CompBERT, we employ an FCN layer coupled with a softmax function to generate the probability distribution across various categories.

TABLE II. SUMMARY OF DATASETS USED IN ENCRYPTED TRAFFIC CLASSIFICATION EXPERIMENTS

Task	**Dataset**	**Flow**	**Packet**	**Label**
General Encrypted Application Classification	Cross-Platform (iOS) [6]	20,858	707,717	196
	Cross-Platform (Android) [6]	27,846	656,044	215
Encrypted Malware Classification	USTC-TFC [26]	9,853	97,115	20
Encrypted Application Classification on Tor	ISCX-Tor [10]	3,021	80,000	16

TABLE III. PERFORMANCE COMPARISON OF DIFFERENT METHODS ON CROSS-PLATFORM (IOS) AND CROSS-PLATFORM (ANDROID) DATASETS

Dataset	**Cross-Platform(iOS)**				**Cross-Platform(Android)**			
Method	AC	PR	RC	F1	AC	PR	RC	F1
AppScanner [8]	0.3205	0.2103	0.2173	0.2030	0.3868	0.2523	0.2594	0.2440
CUMUL [15]	0.2910	0.1917	0.2081	0.1875	0.3525	0.2221	0.2409	0.2189
BIND [7]	0.3770	0.2566	0.2715	0.2484	0.4728	0.3126	0.3253	0.3026
K-fp [25]	0.2155	0.2037	0.2069	0.2003	0.2248	0.2113	0.2104	0.2052
FlowPrint [6]	0.9254	0.9438	0.9254	0.9260	0.8698	0.9007	0.8698	0.8702
DF [9]	0.3106	0.2232	0.2179	0.2140	0.3862	0.2595	0.2620	0.2527
FS-Net [22]	0.3712	0.2845	0.2754	0.2655	0.4846	0.3544	0.3365	0.3343
GraphDApp [20]	0.3245	0.2450	0.2392	0.2297	0.4031	0.2842	0.2786	0.2703
TSCRNN [21]	-	-	-	-	-	-	-	-
Deeppacket [23]	0.9204	0.8963	0.8872	0.9034	0.8805	0.8004	0.7567	0.8138
PERT [24]	0.9789	0.9621	0.9611	0.9584	0.9772	0.8628	0.8591	0.8550
ET-BERT(flow) [11]	0.9844	0.9701	0.9632	0.9643	0.9865	0.9324	0.9266	0.9246
ET-BERT(packet) [11]	0.9810	0.9757	0.9772	0.9754	0.9728	0.9439	0.9119	0.9206
ET-CompBERT(flow)	**0.9964**	**0.9978**	0.9871	0.9924	0.9954	0.9611	**0.9712**	**0.9661**
ET-CompBERT(packet)	0.9945	0.9911	**0.9975**	**0.9943**	**0.9982**	**0.9627**	0.9671	0.9649

TABLE IV. PERFORMANCE COMPARISON OF DIFFERENT METHODS ON ISCX-TOR AND USTC-TFC DATASETS

Dataset	**ISCX-Tor**				**USTC-TFC**			
Method	AC	PR	RC	F1	AC	PR	RC	F1
AppScanner[8]	0.6722	0.3756	0.4422	0.3913	0.8954	0.8984	0.8968	0.8892
CUMUL [15]	0.6606	0.3850	0.4416	0.3918	0.5675	0.6171	0.5738	0.5513
BIND [7]	0.7185	0.4598	0.4515	0.4511	0.8457	0.8681	0.8382	0.8396
K-fp [25]	0.6472	0.5576	0.5849	0.5522	-	-	-	-
FlowPrint [6]	0.9092	0.3820	0.3661	0.3654	0.8146	0.6434	0.7002	0.6573
DF [9]	0.7533	0.6228	0.6010	0.5850	0.7787	0.7883	0.7819	0.7593
FS-Net [22]	0.6071	0.5080	0.5350	0.4590	0.8846	0.8846	0.8920	0.8840
GraphDApp [20]	0.6836	0.4864	0.4823	0.4488	0.8789	0.8226	0.8260	0.8234
TSCRNN [21]	-	0.9490	0.9480	0.9480	-	0.9870	0.9860	0.9870
Deeppacket [23]	0.7449	0.7549	0.7399	0.7473	0.9640	0.9650	0.9631	0.9641
PERT [24]	0.7682	0.4424	0.4446	0.4345	0.9909	0.9911	0.9910	0.9911
ET-BERT(flow) [11]	0.8311	0.5564	0.6448	0.5886	0.9929	0.9930	0.9930	0.9930
ET-BERT(packet) [11]	0.9921	0.9923	0.9921	0.9921	0.9915	0.9915	0.9916	0.9916
ET-CompBERT(flow)	0.8365	0.5598	0.6415	0.5914	0.9916	0.9947	0.9987	0.9967
ET-CompBERT(packet)	**0.9946**	**0.9979**	**0.9957**	**0.9968**	**0.9969**	**0.9964**	**0.9978**	**0.9971**

TABLE V. ABLATION STUDY OF FLOW-LEVEL LEARNING ON CROSS-PLATFORM (IOS) AND CROSS-PLATFORM (ANDROID) DATASETS

Dataset	**Cross-Platform (iOS)**				**Cross-Platform (Android)**			
Method	AC	PR	RC	F1	AC	PR	RC	F1
ET-CompBERT(flow)	**0.9964**	**0.9978**	0.9871	**0.9924**	**0.9954**	**0.9611**	**0.9712**	**0.9661**
−GFA learning	0.9855	0.9874	**0.9887**	0.9881	0.9947	0.9529	0.9648	0.9588
−GFA learning −CFG learning	0.9844	0.9701	0.9632	0.9643	0.9865	0.9324	0.9266	0.9246

TABLE VI. ABLATION STUDY OF FLOW-LEVEL LEARNING ON ISCX-TOR AND USTC-TFC DATASETS

Dataset	**ISCX-Tor**				**USTC-TFC**			
Method	AC	PR	RC	F1	AC	PR	RC	F1
ET-CompBERT(flow)	**0.8365**	**0.5598**	0.6415	**0.5979**	0.9916	0.9947	0.9987	0.9967
−GFA learning	0.8325	0.5577	0.6314	0.5971	**0.9982**	**0.9987**	**0.9992**	**0.9989**
−GFA learning −CFG learning	0.8311	0.5564	**0.6448**	0.5886	0.9929	0.9930	0.9930	0.9930

TABLE VII. ABLATION STUDY OF PACKET-LEVEL LEARNING ON CROSS-PLATFORM (IOS) AND CROSS-PLATFORM (ANDROID) DATASETS

Dataset	**Cross-Platform (iOS)**				**Cross-Platform (Android)**			
Method	AC	PR	RC	F1	AC	PR	RC	F1
ET-CompBERT(packet)	**0.9945**	**0.9911**	**0.9975**	**0.9943**	**0.9982**	**0.9627**	**0.9671**	**0.9649**
−GFA learning	0.9867	0.9814	0.9748	0.9781	0.9910	0.9472	0.9421	0.9446
−GFA learning −CFG learning	0.9844	0.9701	0.9632	0.9643	0.9865	0.9324	0.9266	0.9246

TABLE VIII. ABLATION STUDY ON PACKET-LEVEL LEARNING ON ISCX-TOR AND USTC-TFC DATASETS

Dataset	**ISCX-Tor**				**USTC-TFC**			
Method	AC	PR	RC	F1	AC	PR	RC	F1
ET-CompBERT(packet)	**0.9946**	**0.9979**	**0.9957**	**0.9968**	**0.9969**	**0.9964**	**0.9978**	**0.9971**
−GFA learning	0.9904	0.9577	0.9535	0.9556	0.9841	0.9834	0.9847	0.9840
−GFA learning −CFG learning	0.9865	0.9324	0.9266	0.9246	0.9729	0.9756	0.9731	0.9733

IV. EXPERIMENTS

A. Dataset and Metrics

To validate the efficacy and broad applicability of ET-CompBERT, we conducted a series of experiments across three established encrypted traffic classification tasks, utilizing four publicly accessible datasets. Table II delineates the specifics of these datasets. The General Encrypted Application Classification task focuses on categorizing application traffic under standard encryption protocols. Our evaluations were conducted on the Cross-Platform datasets for both iOS and Android, encompassing 196 and 215 applications respectively. The Encrypted Malware Classification task involves the analysis of encrypted traffic comprising both malware and benign applications. In this context, the USTC-TFC dataset is particularly noteworthy, as it features 10 categories each of benign and malicious traffic, providing a comprehensive framework for the assessment of encryption-based malware detection capabilities. The Encrypted Application Classification on Tor task is centered around classifying encrypted traffic using the Onion Router to enhance communication privacy. The relevant dataset, termed ISCX-Tor, comprises 16 distinct applications, offering a unique landscape for assessing privacy-preserved encrypted traffic analysis.

B. Implementation Details

During comprehensive fusion-guided learning, approximately 30GB of traffic data is utilized for pre-training purposes. The dataset is divided into two segments: (1) roughly 15GB of traffic data sourced from public datasets [18], [19]; (2) an equivalent volume of traffic data, approximately 15GB, obtained through passive collection within the China Science and Technology Network (CSTNET). the batch size is set at 32, and the total number of steps is 500,000. The learning rate is established at 2×10^{-5}, with a warmup ratio of 0.1. For learning, we utilize the AdamW optimizer across 10 epochs, applying a learning rate of 6×10^{-5} for flow-level and 2×10^{-5} for packet-level tasks. The batch size remains at 32, and the dropout rate is set to 0.5. All experiments are conducted using Pytorch 1.8.0 on eight NVIDIA Tesla V100 GPUs. In our approach, we implement two distinct learning strategies for the ET-CompBERT model to adapt to different levels of traffic data granularity, the ET-CompBERT(flow) and the ET-CompBERT(packet). Here the $e_b, e_s \in \mathbb{R}^{768}$, $f_b, f_s \in \mathbb{R}^{120}$.

For testing, we maintained consistency in the dataset across both strategies, ensuring a fair and objective comparison with other methodologies. The pivotal difference between these strategies lied in the granularity of the fine-tuning input traffic information. Our method employed a dataset comprising a concatenated sequence of M consecutive packets within a flow, where M is predefined as 5 in our experimental setup.

C. Comparison with Existing Methods

Tables III and IV showcase the performance comparison of our framework with existing frameworks. Our framework sets a new benchmark in state-of-the-art performance, outperforming the preceding leading method in the F1-score across all four datasets with flow-level fine-tuning. The margin of enhancement ranges from +0.28% to +4.43%, as substantiated by the results enumerated in both tables. Moreover, it is imperative to highlight that the results of our experiments surpass all previous state-of-the-art methods in terms of F1-score in the packet-level fine-tuning. These outcomes attest to the proficiency of our framework in synthesizing coarse-grained and fine-grained insights into encrypted traffic analysis. The observed discrepancy in classification capability may be attributed to the inherently more fine-grained nature of packet-level fine-tuning compared to flow-level fine-tuning. This granularity enables the self-attention mechanism to capture subtler details within the encrypted traffic, potentially leading to enhanced model classification performance.

D. Ablation Studies

In this study, we evaluate the impact of individual components within our framework in both flow-level and packet-level learning. Crucially, omitting the comprehensive fusion-guided (CFG) learning and global-feature-Aware (GFA) learning from ET-CompBERT reverts it to its baseline counterpart, ET-BERT. Tables V and VI detail the performance implications of these components in flow-level learning scenarios. As delineated in Table V, the omission of GFA learning precipitates a minimum decline of -0.43% in the F1-score for the Cross-Platform (Android) dataset. More strikingly, the simultaneous removal of both GFA and CFG learning induces a minimum downturn of approximately -3.42%in the F1-score. Such outcomes accentuate the critical role of our proposed CFG and GFA methodologies in the effective assimilation of multi-grained features for encrypted traffic classification.

We assess the performance of our model using four fundamental metrics: Accuracy (AC), Precision (PR), Recall (RC), and F1-Score. This involves calculating the mean values of AC, PR, RC, and F1 for each category, ensuring a more equitable and round evaluation framework.

The packet-level learning, as shown in Tables VII and VIII, ET-CompBERT demonstrates robust performance across all four datasets. All classification abilities decline when removing GFA learning or removing both GFA learning and CFG learning. A discernible decline in performance is noted with the removal of the GFA learning, evidenced by a marked reduction of -4.12% in the F1-score, most notably in the ISCX-Tor dataset. They were similarly, eliminating both the GFA learning and CFG learning results in a further decrease, with the F1-score dropping by -3.10% in the same dataset. These observations underscore the substantial contributions of both

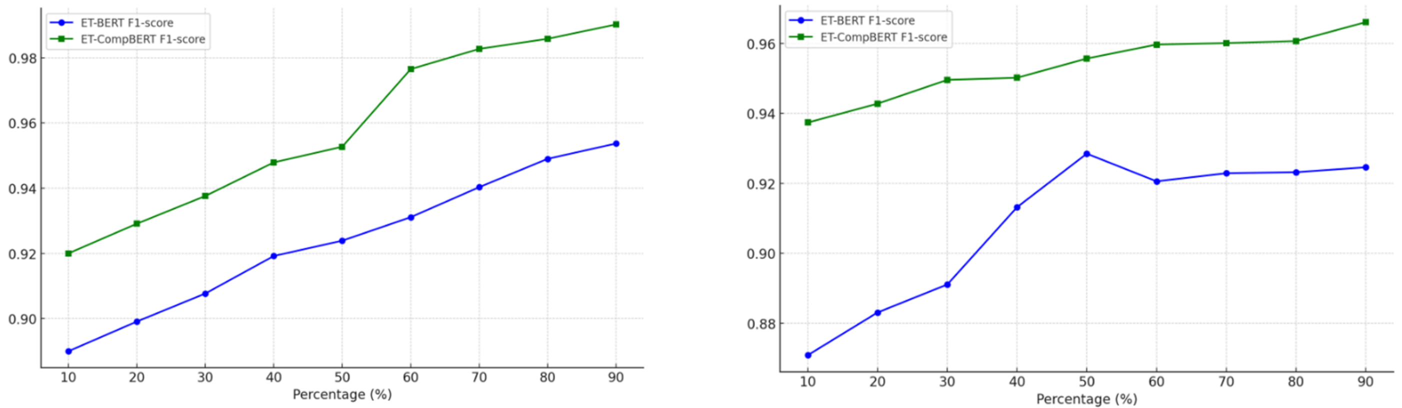

Fig. 2. F1-scores of flow-level ET-BERT and ET-CompBERT using varying percentages of training samples on (a) Cross-Platform (iOS) and (b) Cross-Platform (Android) datasets.

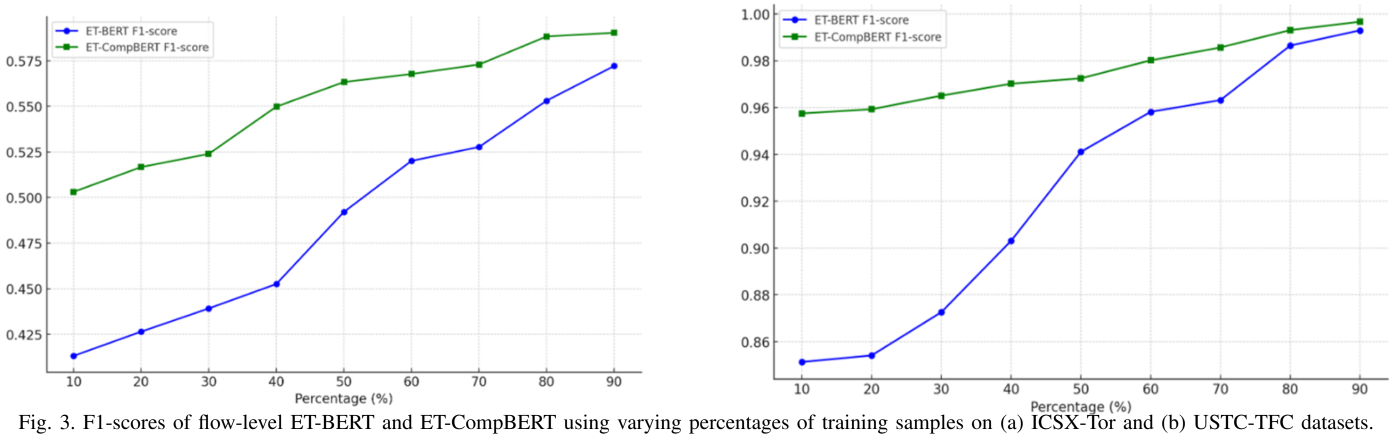

Fig. 3. F1-scores of flow-level ET-BERT and ET-CompBERT using varying percentages of training samples on (a) ICSX-Tor and (b) USTC-TFC datasets.

TABLE IX. Performance Comparison of Different Pre-training Learning Methods on Cross-Platform (ios) and Cross-Platform (Android) Datasets

Dataset	**Cross-Platform(ios)**				**Cross-Platform(Android)**			
Method	AC	PR	RC	F1	AC	PR	RC	F1
ET-CompBERT(flow)	**0.9964**	**0.9978**	0.9871	0.9924	0.9954	0.9611	**0.9712**	**0.9661**
ET-CompBERT(packet)	0.9945	0.9911	**0.9975**	**0.9943**	**0.9982**	**0.9627**	0.9671	0.9649
ET-BERT+GFA(flow)	0.9851	0.9745	0.9701	0.9721	0.9870	0.9331	0.9294	0.9312
ET-BERT+GFA(packet)	0.9824	0.9784	0.9842	0.9813	0.9754	0.9511	0.9187	0.9346
ET-BERT(flow) [11]	0.9844	0.9701	0.9632	0.9643	0.9865	0.9324	0.9266	0.9246
ET-BERT(packet) [11]	0.9810	0.9757	0.9772	0.9754	0.9728	0.9439	0.9119	0.9206

the GFA learning and CFG learning to the overall effectiveness of our framework in encrypted traffic classification tasks.

E. Analysis

We conduct an extensive analysis of the impact of varying dataset volumes under flow-level fine-tuning, as depicted in Fig. 2 and Fig. 3, and packet-level fine-tuning, illustrated in Fig. 4 and Fig. 5. Additionally, we perform a comparative evaluation of the classification performances using different late fusion methods, which are systematically presented in Table IX and X.

Fig. 2 and Fig. 3 demonstrate the comparative performance of ET-BERT and ET-CompBERT across four datasets under flow-level fine-tuning. It is observed that with the reduction in dataset volumes, ET-CompBERT's classification performance consistently outperforms that of ET-BERT. Significantly, our proposed framework exhibits a notable enhancement over ET-BERT, evidenced by an approximate increase of +0.37% in classification accuracy under 90% GFA learning data volumes. This improvement is even more marked at lower dataset volumes, especially at 10%. Particularly, under 10% volumes of the USTC-TFC dataset, as shown in Fig. 3(b), the classification capability of our framework substantially surpasses ET-BERT by about +10.62%. These findings clearly illustrate the superior robustness and improved classification efficacy of our proposed framework compared to ET-BERT when applied to flow-level fine-tuning. In the packet-level fine-tuning scenario in Fig. 3 and Fig. 4, the ascending trends observed across the four datasets indicate that ET-CompBERT significantly outperforms in scenarios with reduced dataset volumes, particularly evident in the ISCX-Tor dataset, where it surpasses ET-BERT by a notable +9.21% in Fig. 4(a). These experimental results validate that our comprehensive fusion-guided learning

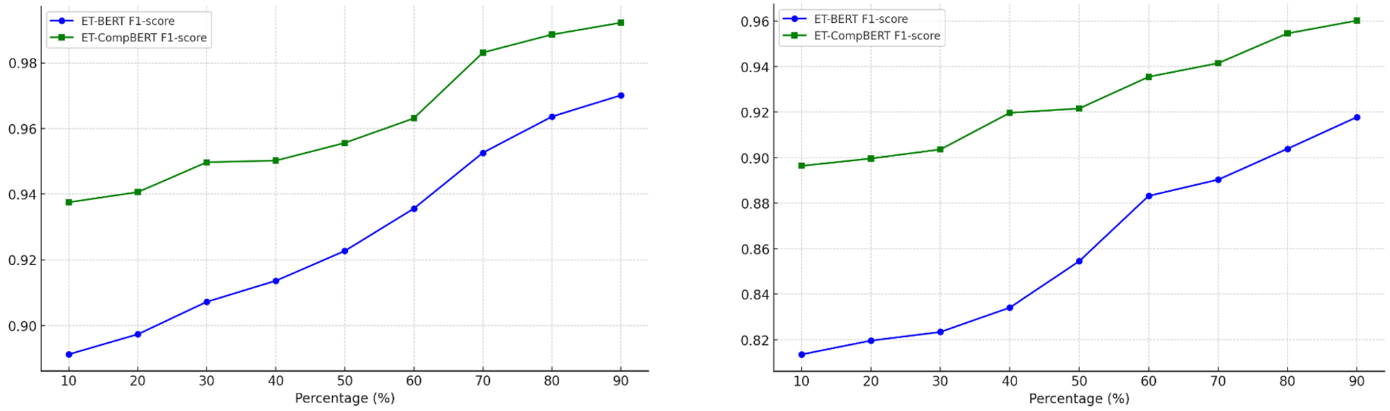

Fig. 4. F1-scores of packet-level ET-BERT and ET-CompBERT using varying percentages of training samples on (a) Cross-Platform (iOS) and (b) Cross-Platform (Android) datasets.

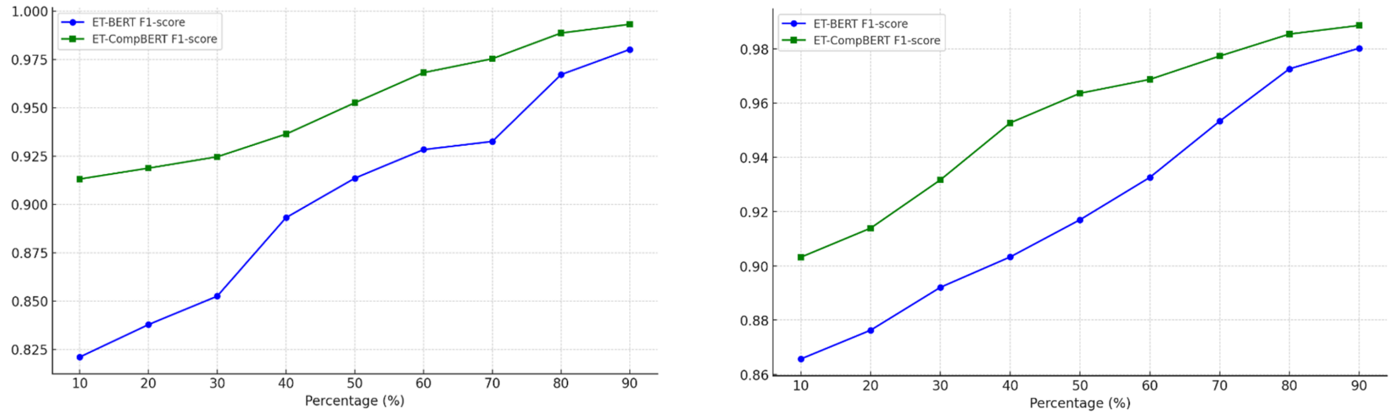

Fig. 5. F1-scores of packet-level ET-BERT and ET-CompBERT using varying percentages of training samples on (a) ICSX-Tor and (b) USTC-TFC datasets.

TABLE X. PERFORMANCE COMPARISON OF DIFFERENT PRE-TRAINING LEARNING METHODS ON ISCX-TOR AND USTC-TFC DATASETS

Dataset	ISCX-Tor				USTC-TFC			
Method	AC	PR	RC	F1	AC	PR	RC	F1
ET-CompBERT(flow)	0.8365	0.5598	0.6415	0.5914	0.9916	0.9947	0.9987	0.9967
ET-CompBERT(packet)	**0.9946**	**0.9979**	**0.9957**	**0.9968**	**0.9969**	**0.9964**	**0.9978**	**0.9971**
ET-BERT+GFA(flow)	0.8301	0.5407	0.6319	0.5828	0.9878	0.9908	0.9898	0.9903
ET-BERT+GFA(packet)	0.9934	0.9947	0.9931	0.9939	0.9807	0.9911	0.9936	0.9923
ET-BERT(flow) [11]	0.8311	0.5564	0.6448	0.5886	0.9929	0.9930	0.9930	0.9930
ET-BERT(packet) [11]	0.9921	0.9923	0.9921	0.9921	0.9915	0.9915	0.9916	0.9916

approach effectively enables the framework to comprehend coarse-grained global temporal attributes, building upon its understanding of fine-grained information from the encrypted traffic payload. Concurrently, these global temporal attributes contribute to enhancing the pre-trained model's proficiency in encrypted traffic classification, underscoring the synergy between different granularities of data in improving model encrypted traffic classification performance.

In addition to our primary methodology, we examine an alternative late fusion technique to replace the comprehensive fusion-guided learning. This approach computes an arithmetic mean of the prediction probabilities derived from the fine-grained payload encoder, ET-BERT, and the encoder capturing global temporal attributes. Designated as ET-BERT+GFA, this advanced late fusion methodology strives to amalgamate multi-grained informational aspects. In this technique, the encoder responsible for capturing global temporal attributes computes a probability distribution that is the arithmetic mean of its own output and that of ET-BERT. Unfortunately, as Tables IX and X reveal, this method falls short of achieving the desired performance in encrypted traffic classification. Table X, in particular, illustrates that ET-BERT+GFA's classification efficacy does not surpass that of ET-BERT. The underlying cause of this shortfall may be attributed to the simplistic nature of this late fusion approach, which potentially disrupts the model's ability to integrate coarse-grained features without preserving the intricate fine-grained details of the encrypted traffic. These findings affirm the effectiveness of our comprehensive fusion-guided learning in enabling the model to assimilate coarse-grained temporal attributes without compromising the nuanced information acquired from the fine-grained payload of encrypted traffic.

V. Conclusion

In this work, we introduce a novel framework for encrypted traffic classification, named ET-CompBERT. This framework innovatively integrates the fine-grained payload characteristics of encrypted traffic with the coarse-grained global temporal attributes. We also introduce innovative GFA learning which endows our framework with robust encrypted traffic classification results under different data sizes. Extensive experimental results validate the effectiveness of our approach.

References

[1] S. Rezaei and X. Liu, "Deep learning for encrypted traffic classification: An overview," *IEEE Communications Magazine*, vol. 57, no. 5, pp. 76–81, 2019.

[2] P. Velan, M. Čermák, P. Čeleda, and M. Drašar, "A survey of methods for encrypted traffic classification and analysis," *International Journal of Network Management*, vol. 25, no. 5, pp. 355–374, 2015, Wiley Online Library.

[3] T. Hu, C. Xu, S. Zhang, S. Tao, and L. Li, "Cross-site scripting detection with two-channel feature fusion embedded in self-attention mechanism," *Computers & Security*, vol. 124, pp. 102990, 2023, Elsevier.

[4] N. Thalji, A. Raza, M. S. Islam, N. A. Samee, and M. M. Jamjoom, "AE-Net: Novel Autoencoder-Based Deep Features for SQL Injection Attack Detection," *IEEE Access*, vol. 11, pp. 135507–135516, 2023.

[5] G. S. Nilavarasan and T. Balachander, "XSS Attack Detection using Convolution Neural Network," in *Proceedings of the 2023 International Conference on Artificial Intelligence and Knowledge Discovery in Concurrent Engineering (ICECONF)*, pp. 1–6, 2023, IEEE.

[6] T. van Ede, R. Bortolameotti, A. Continella, J. Ren, D. J. Dubois, M. Lindorfer, D. Choffnes, M. van Steen, and A. Peter, "Flowprint: Semi-supervised mobile-app fingerprinting on encrypted network traffic," in *Network and Distributed System Security Symposium (NDSS)*, vol. 27, 2020.

[7] K. Al-Naami, S. Chandra, A. Mustafa, L. Khan, Z. Lin, K. Hamlen, and B. Thuraisingham, "Adaptive encrypted traffic fingerprinting with bi-directional dependence," in *Proceedings of the 32nd Annual Conference on Computer Security Applications*, pp. 177–188, 2016.

[8] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "Robust smartphone app identification via encrypted network traffic analysis," *IEEE Transactions on Information Forensics and Security*, vol. 13, no. 1, pp. 63–78, 2017.

[9] P. Sirinam, M. Imani, M. Juarez, and M. Wright, "Deep fingerprinting: Undermining website fingerprinting defenses with deep learning," in *Proc. 2018 ACM SIGSAC Conf. Comput. Commun. Secur.*, 2018, pp. 1928–1943.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: learning of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[11] X. Lin, G. Xiong, G. Gou, Z. Li, J. Shi, and J. Yu, "Et-bert: A contextualized datagram representation with learning transformers for encrypted traffic classification," in *Proc. ACM Web Conf. 2022*, 2022, pp. 633–642.

[12] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, PMLR, 2021, pp. 8748–8763.

[13] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[14] V. F. Taylor, R. Spolaor, M. Conti, and I. Martinovic, "Appscanner: Automatic fingerprinting of smartphone apps from encrypted network traffic," in *Proc. 2016 IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, 2016, pp. 439–454.

[15] A. Panchenko, F. Lanze, J. Pennekamp, T. Engel, A. Zinnen, M. Henze, and K. Wehrle, "Website Fingerprinting at Internet Scale," in *Proc. NDSS*, 2016.

[16] W. Wang, M. Zhu, J. Wang, X. Zeng, and Z. Yang, "End-to-end encrypted traffic classification with one-dimensional convolution neural networks," in *Proc. 2017 IEEE Int. Conf. Intell. Secur. Inform. (ISI)*, 2017, pp. 43–48.

[17] 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. Adv. Neural Inform. Process. Syst.*, vol. 30, 2017.

[18] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related," in *Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP)*, pp.

[19] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *ICISSp*, vol. 1, pp. 108–116, 2018.

[20] M. Shen, J. Zhang, L. Zhu, K. Xu, and X. Du, "Accurate decentralized application identification via encrypted traffic analysis using graph neural networks," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2367–2380, 2021.

[21] K. Lin, X. Xu, and H. Gao, "TSCRNN: A novel classification scheme of encrypted traffic based on flow spatiotemporal features for efficient management of IIoT," *Computer Networks*, vol. 190, pp. 107974, 2021, Elsevier.

[22] C. Liu, L. He, G. Xiong, Z. Cao, and Z. Li, "Fs-net: A flow sequence network for encrypted traffic classification," in *Proc. IEEE INFOCOM 2019 - IEEE Conf. Comput. Commun.*, 2019, pp. 1171–1179.

[23] M. Lotfollahi, M. Jafari Siavoshani, R. Shirali Hossein Zade, and M. Saberian, "Deep packet: A novel approach for encrypted traffic classification using deep learning," *Soft Computing*, vol. 24, no. 3, pp. 1999–2012, 2020, Springer.

[24] H. Y. He, Z. G. Yang, and X. N. Chen, "PERT: Payload encoding representation from transformer for encrypted traffic classification," in *Proceedings of the 2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K)*, pp. 1–8, 2020, IEEE.

[25] J. Hayes and G. Danezis, "k-fingerprinting: A robust scalable website fingerprinting technique," in *Proceedings of the 25th USENIX Security Symposium (USENIX Security 16)*, pp. 1187–1203, 2016.

[26] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in *Proc. 2017 International Conference on Information Networking (ICOIN)*, pp. 712–717, 2017.