

# Contextual Masking Distillation for Network Traffic Anomaly Detection

Xinglin Lian<sup>✉</sup>, Yu Zheng<sup>✉</sup>, *Member, IEEE*, Yan Liu, Fan Zhou<sup>✉</sup>, *Member, IEEE*, Chunlei Peng<sup>✉</sup>, *Member, IEEE*, and Xinbo Gao<sup>✉</sup>, *Fellow, IEEE*

**Abstract**—Network traffic anomaly detection is critical for cybersecurity but faces challenges in accurately identifying malicious activities. Recent zero-positive approaches, which use only normal training data under the reconstruction paradigm, have shown progress. However, encrypted network traffic obscures normal-anomalous distinctions, causing confused modeling. In addition, the “identical shortcut” problem, where models reconstruct any input with similar fidelity, produces suboptimal representations and indistinguishable detection. To address these limitations, this paper introduces ConMD, a novel Contextual Masking Knowledge Distillation framework. ConMD features distillation paradigm for discriminative representations and then pursues two objectives: effective contextual information modeling and a comprehensive anomaly metric. Specifically, we introduce context-aware local-global attention mechanisms for the student network’s backbone, which capture both intra-packet and inter-packet dependencies. Additionally, a context-enhanced masking training strategy is designed to facilitate contextual interactions in normal flows. Given the structural characteristics of network traffic, we also present a new anomaly scoring with multi-view awareness, which perceive comprehensive traffic patterns. ConMD combines insights from both packet- and flow-level views to highlight deviations in anomalous network flows, thereby improving detection accuracy. Extensive experiments on three real-world datasets validate the effectiveness of ConMD, yielding consistent improvements over state-of-the-art baselines, achieving up to 2.8% and 5.1% AUC gains on the DataCon2020 and

CIC-IDS2017 datasets, respectively. Our model code is available at <https://github.com/ikun0124/ConMD>

**Index Terms**—Network traffic anomaly detection, knowledge distillation, contextual information awareness, multi-view anomaly awareness.

## I. INTRODUCTION

NETWORK traffic anomaly detection aims to identify the potentially malicious and harmful network flows that could compromise equipment and systems on communication links [1], [2], [3]. This task particularly plays a critical role in bolstering cybersecurity, contributing to a more secure cyberspace [4], [5], enhanced network reliability [6], and robust privacy protection [7]. Traditional approaches process network traffic data as graph, sequence, and image representations [8], [9], framing this task as a supervised binary classification problem [10], [11], [12]. Despite commendable performance improvements, supervised methods still face inherent limitations in handling anomalous traffic, such as the class imbalance problem and the detection of unknown anomalies [1].

Advanced research has shifted towards a zero-positive learning setting [13], [14], [15], which only constructs normal network traffic distribution during training. Considering the high computational cost of processing graph- or sequence-based representations, most zero-positive approaches instead transform traffic data into image form. They primarily employ a reconstruction-based modeling paradigm [16], [17], [18], which reconstructs normal traffic images using autoencoder architectures. These methods rely on a fundamental hypothesis that normal traffic would exhibit significantly better reconstruction quality than anomalous traffic. As illustrated in the “Hypothesis” row of Fig. 1, the reconstruction gap serves as the key criterion for anomaly scoring. In practice, however, this expectation often fails. Recent studies [16] report that the actual reconstruction results, exemplified by samples from the DataCon2020 dataset [19] shown in the “Practice” row of Fig. 1, reveal a surprising observation. Both normal and anomalous traffic exhibit similar reconstruction quality, making them difficult to distinguish. The root cause lies in the intrinsically confused distributions of normal and anomalous traffic, with widely used encryption techniques obscuring semantic information of traffic data [16]. Moreover, the “identical shortcut” phenomenon in auto-encoder architectures [13], [20] exacerbates this problem, as the model

Received 11 December 2024; revised 2 September 2025 and 16 October 2025; accepted 13 January 2026. Date of publication 19 January 2026; date of current version 28 January 2026. This work was supported in part by the New Generation Artificial Intelligence-National Science and Technology Major Project under Grant 2025ZD0123601; in part by the National Natural Science Foundation of China under Grant 62276198, Grant U22A2035, Grant 62572097, and Grant 62176043; in part by the Natural Science Basic Research Program of Shaanxi under Grant 2025JC-YBMS-696; in part by the Innovation Capability Support Plan in Shaanxi Province under Grant 2025ZC-KJXX-22; and in part by the 111 Center under Grant B16037. The associate editor coordinating the review of this article and approving it for publication was Prof. Weizhi Meng. (*Corresponding author: Yu Zheng.*)

Xinglin Lian is with the University of Electronic Science and Technology of China, Chengdu 610054, China, and also with the School of Cyber Engineering, Xidian University, Xi’an 710126, China (e-mail: kenshin.lian24@gmail.com).

Yu Zheng and Chunlei Peng are with the School of Cyber Engineering, Xidian University, Xi’an 710126, China (e-mail: yuzheng.xidian@gmail.com; clpeng@xidian.edu.cn).

Yan Liu is with the University of Electronic Science and Technology of China, Chengdu 610054, China (e-mail: Yan.Liu@std.uestc.edu.cn).

Fan Zhou is with the University of Electronic Science and Technology of China, Chengdu 610054, China, and also with the Key Laboratory of Intelligent Digital Media Technology of Sichuan Province, Chengdu 610054, China (e-mail: fan.zhou@uestc.edu.cn).

Xinbo Gao is with the State Key Laboratory of Integrated Services Networks, School of Electronic Engineering, Xidian University, Xi’an 710071, China (e-mail: xbgao@mail.xidian.edu.cn).

Digital Object Identifier 10.1109/TIFS.2026.3655514

1556-6021 © 2026 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.

See <https://www.ieee.org/publications/rights/index.html> for more information.

Authorized licensed use limited to: Institute of Information Engineering CAS. Downloaded on March 13, 2026 at 09:34:15 UTC from IEEE Xplore. Restrictions apply.

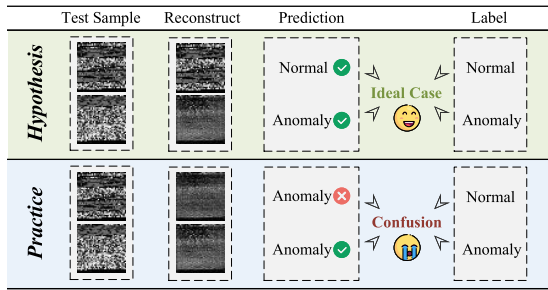


Fig. 1. Limitation in existing detection methods. Ideally, a reconstruction model would demonstrate a clear distinction in output quality between normal and anomalous network traffic. In practice, however, the reconstructions often lack of differentiation and lead to prediction confusion.

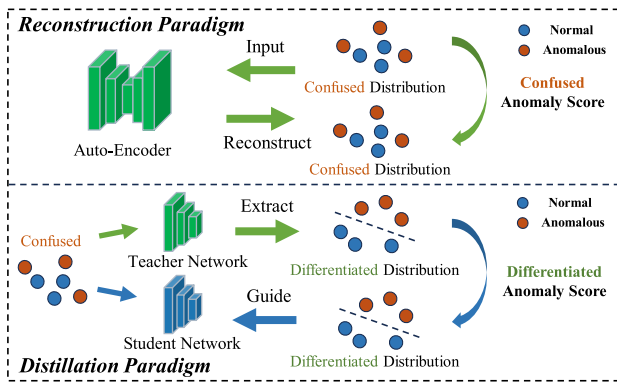


Fig. 2. Paradigm discrepancy between reconstruction- and distillation-based methods. Reconstruction-based anomaly scores rely on a “Confused-to-Confused” pipeline, whereas distillation-based methods use a pre-trained teacher to extract representations, confirming differentiated anomaly detection.

tends to converge to shortcut parameters that produce similar reconstructions for both normal and anomalous samples. Consequently, the model suffers from prediction confusion and suboptimal detection results. Although some works strengthen the representation power of autoencoders to alleviate this problem [20], [21], they still operate within the reconstruction paradigm and cannot fundamentally overcome its inherent flaws.

Recently, knowledge distillation-based anomaly detection has emerged as a promising paradigm to address this issue. By leveraging the refined knowledge extracted from a powerful pre-trained teacher network, distillation-based methods aim to provide more discriminative representations that benefit the zero-positive modeling [22], [23]. Fig. 2 illustrates the core discrepancy between reconstruction-based and distillation-based paradigms. Based on our previous analysis, we formally define the reconstruction-based paradigm as a “Confused-to-Confused” pipeline: the reconstruction model tends to generate confused output as it is driven by an inherently confused reconstruction objective. Consequently, the anomaly scores (i.e., the difference between input and reconstruction) remain inadequate for accurate anomaly detection. In contrast, the distillation-based paradigm employs a pre-trained teacher network to extract discriminative representations, thereby enabling a differentiated modeling pipeline. These representations improve zero-positive learning and amplify the discrepancy between normal and anomalous

samples [23]. This makes distillation a promising solution to mitigate prediction confusion in network traffic anomaly detection.

However, network traffic is typically represented at a flow-level unit (i.e., network flow), which encompasses a sequence of ordered packets in communication links, with strong contextual interactions between packets [9], [24]. Existing distillation techniques, primarily designed for natural images, often falter when applied to the unique characteristics of network flow data. These challenges arise from two key aspects: (1) **Insufficient Contextual Awareness**: Conventional distillation methods struggle to capture the inherent contextual relationships within network flows, such as inter-packet dependencies, causing a suboptimal model fitting and unclear boundaries between normal and anomalous flows. (2) **Overemphasis on Local Anomalies**: Current techniques tend to prioritize local anomalies, such as localized device failures or medical abnormalities [25], which are more applicable to natural image domains. However, in the context of network traffic, anomalies often manifest at both the packet-level (i.e., intra-packet patterns) and flow-level (i.e., inter-packet sequences) rather than in localized regions [11].

To address these limitations, we introduce **ConMD**, a novel **Contextual Masking Knowledge Distillation** framework. Our approach targets two primary objectives: (1) effectively modeling contextual information during training, and (2) discerning anomalies from a comprehensive perspective during inference. ConMD begins with a fine-tuned teacher network that extracts differentiated features for both normal and anomalous traffic, facilitating subsequent zero-positive modeling and the establishment of robust anomaly metrics. We then design a context-aware student network backbone equipped with local-global window attention mechanisms to capture both intra-packet and inter-packet dependencies within normal network flows. Specifically, local attention mechanisms address short-range intra-packet dependencies, while global attention mechanisms handle long-range inter-packet dependencies. To enhance contextual distillation, we propose a novel context-enhanced packet-level masking strategy, which randomly masks packet information based on a byte-strength rule. During training, the student network is encouraged to restore the masked packets from the unmasked context, thereby cultivating richer inter-packet dependencies. Since only normal traffic is available for training, the model effectively learns typical contextual patterns among normal packets.

During inference, given the structural characteristics of network traffic, ConMD achieves multi-view anomaly scoring by integrating masked and unmasked feature branches. The masked branch emphasizes packet-level anomalies: when masked anomalous packets are restored toward a normal state, the resulting deviations are amplified, making local anomalies more visible. In contrast, the unmasked branch, strengthened by context-enhanced learning during training, captures typical inter-packet contextual dependencies, enabling flow-level anomaly detection from a global perspective. By integrating these multi-view representations through a feature fusion scheme, ConMD highlights anomalous deviations across both

local (packet-level) and global (flow-level) contexts, thereby achieving comprehensive anomaly detection. In summary, our key contributions are threefold:

- We propose ConMD, a distillation framework that alleviates reconstruction confusion in network traffic. ConMD provides a more differentiated modeling pipeline and improved anomaly detection. To the best of our knowledge, this is the first attempt to explore distillation-based zero-positive anomaly detection in network traffic data.
- We introduce a context-aware student network equipped with local-global attention mechanisms to effectively capture both intra-packet and inter-packet dependencies. Additionally, we design a context-enhanced packet-level masking training strategy to facilitate deeper contextual interactions within network flow data.
- We design a multi-view anomaly scoring mechanism that can perceive both packet- and flow-level anomalies, combining their respective strengths to highlight anomalous deviations in network traffic.
- Extensive experiments on benchmark datasets show that ConMD outperforms state-of-the-art methods, with AUC improvements of 2.8% and 5.1% against the best baselines, MFAD and MMR, on the DataCon2020 and CIC-IDS2017 datasets, respectively.

The remainder of this paper is organized as follows. Section II presents a comprehensive review of related literature. Section IV details our framework design and the motivations underpinning it. In Section V, we evaluate ConMD on the benchmark datasets, comparing it with state-of-the-art methods and assessing both model detection performance and effectiveness. Finally, Section VI concludes the paper and outlines potential directions for future research.

## II. RELATED WORK

### A. Network Traffic Anomaly Detection

Anomaly detection in network traffic seeks to identify malicious or harmful network flows within network links. Traditionally, this task was approached as a supervised binary classification problem [26], [27], [28]. Neural network models such as Convolutional Neural Network (CNN) [9], [29], [30], Long Short-term Memory Networks (LSTM) [31], [32], [33], and Graph Neural Network (GNN) [34], [35], [36] were frequently used to extract spatial and temporal features. The softmax layer was employed to accomplish supervised classification [37], [38]. However, supervised learning methods face inherent limitations, particularly their vulnerability to unknown anomalies and class imbalance. These challenges have spurred a shift towards zero-positive approaches.

For example, SAFE employed a masked autoencoder to extract features and a novelty detector to identify anomalies [8]. Anomal-E leveraged a GNN to capture both local and global information [39]. One prominent zero-positive framework is GANomaly [40], which incorporates an adversarial discriminator network to improve the reconstruction of normal samples. GANomaly has been successfully applied to network traffic analysis [17], [41]. MANomaly [41] introduced a dual autoencoder adversarial training strategy for enhanced

representation learning. ARCADE [17] combined a structural similarity measure loss with the WGAN-GP framework for adversarial training. Trident [15] integrated a U-Net structure to preserve detailed input information. Overall, these studies endeavor to improve the reconstruction quality of the normal sample. However, the confused distribution between normal and anomalous sample significantly limit their paradigm intuitions and the effectiveness of practical anomaly metrics.

Recently, UnDiff [13] identified a critical “identical shortcut” issue in traffic images reconstruction, both normal and anomalous traffic tend to yield similar reconstruction quality. MFAD [18] and MFR [16] proposed low-pass-filter-based solutions that showed some promise, but they risked discarding high-frequency information crucial for accurate detection. Other anomaly detection studies [20], [21] attempted to mitigate the “identical shortcut” by enhancing feature learning to make autoencoders more sensitive to anomalies. Nevertheless, they operated within and were limited by the reconstruction paradigm, preventing a fundamental solution.

### B. Distillation-Based Anomaly Detection

While some studies in network traffic analysis introduce knowledge distillation, its typical application involves transferring a complex teacher network into a lightweight student network [42]. Recently, knowledge distillation has also emerged as a promising paradigm in anomaly detection, supporting both feature modeling and anomaly scoring. Its main advantage lies in leveraging a pre-trained teacher network to generate refined, noise-free representations that guide the student network in learning more discriminative features for normal samples [22], [23]. During inference, anomaly scores are computed based on the discrepancy between the teacher and student networks: the student outputs representations biased toward normal state, whereas the teacher produces distinct representations for both normal and anomalous samples.

The early work UStd [43] pioneered the application of knowledge distillation. STFPM [44] employed an identical student-teacher architecture with multi-scale feature matching to detect anomalies but faced challenges of homogeneous information confusion. To address this, ReverDis [22] proposed a reverse distillation paradigm based on an encoder-decoder network. AST [45] improved generalization by employing highly differentiated normalizing flows to increase the distance between student and teacher representations. Recent studies have demonstrated the effectiveness of masking mechanisms in anomaly detection. For example, SSMRKD [46] introduced two-stage masking training for reverse knowledge distillation, enhancing the learning of single-category prototype patterns. Similarly, DeSTSeg [25] incorporated a demasking process to encourage the student network to learn more robust representations. MMR [47] proposed a masked multi-scale reconstruction strategy to improve causal relationships between patches via a masked feature pyramid network.

Despite these advancements, applying existing knowledge distillation methods to anomaly detection in network traffic presents unique challenges. Unlike natural image domains, high-frequency texture characteristics of traffic images limit



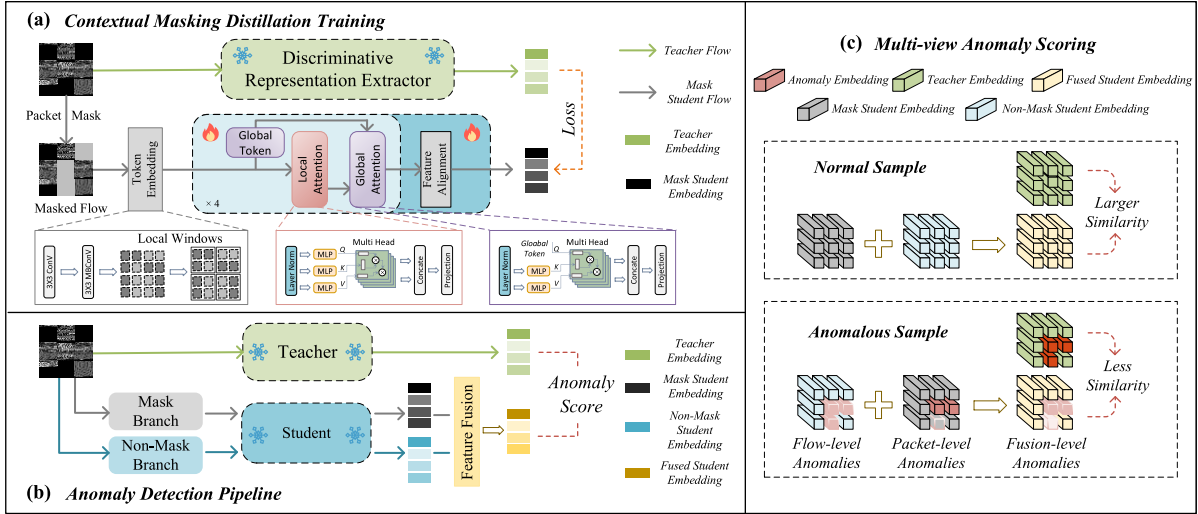


Fig. 3. Overall framework of ConMD. (a) Contextual Masking Distillation Training Process: the teacher network extracts discriminative representations to guide a more differentiated modeling pipeline. The student backbone, equipped with context-aware local-global window attention mechanisms, leverages the context-enhanced masking strategy to capture intra-packet and inter-packet dependencies. (b) Anomaly Detection Pipeline: during inference, ConMD performs multi-view anomaly scoring by fusing packet-level/masked and flow-level/non-masked branch representations. (c) Motivation for Multi-view Anomaly Scoring: fusion-level features combine the strengths of both views to amplify anomalous deviations.

the ability to explicitly represent texture information [18], causing failures in perceiving localized and multi-scale anomalies. Moreover, the nature of network data emphasizes strong contextual relationships rather than localized awareness [24].

### III. PRELIMINARIES

Let  $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$  represent a set of  $N$  normal network traffic samples. Each sample  $\mathbf{x}_i \in \mathbb{R}^{p \times d}$  represents a network flow, consisting of a sequence of network packets  $x_p \in \mathbb{R}^d$  exchanged between endpoints within a connection session, where  $p$  is the number of packets in the flow and  $d$  is the byte length of each packet instance [11], [16]. Our zero-positive detection models aim to fit the data distribution of normal flows during training. For inference, the model generates an anomaly score  $Score$  for each tested sample  $\mathbf{x}_{test} \in \mathcal{X}_{test}$ , where  $\mathcal{X}_{test}$  represents the set of test samples. This score indicates the deviations from the learned behavior of normal flow samples, with its magnitude positively correlated with the likelihood of being identified as anomalous.

### IV. METHODOLOGY

In this section, we present our proposed contextual masking distillation method, ConMD. The training and inference pipeline of ConMD is depicted in Fig. 3 (a) and 3(b), while the detection motivation is illustrated in Fig. 3 (c).

#### A. Flow-Level Representation Construction

The basic unit of network traffic is a flow, which comprises a series of ordered packets [24]. Some zero-positive works depend on manually extracted statistic features rather than raw network flow information [13], [16], which omit payload information and degrade performance. In this study, we propose a novel flow-level, end-to-end-based detection scheme. Each

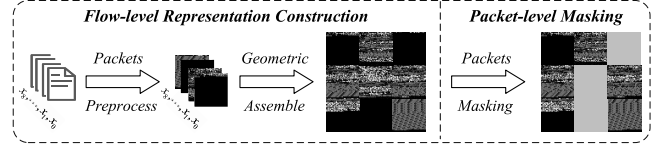


Fig. 4. The process of constructing flow-level representations and the corresponding packet-level masking strategy.

packet  $x_p$  in a flow  $\mathbf{x}_i$  is preprocessed as follows: Ethernet header is removed, IP addresses are set to zero, and the packet is padded to the maximum byte length  $d = 1,600$  using byte  $0 \times 00$  to prevent noise and overfitting impact [9], [18]. The preprocessed packets are then reshaped into a 2D image format  $x_p \in \mathbb{R}^{40 \times 40}$ , as shown in the left part of Fig. 4. To incorporate flow-level representation, we propose a new flow-level geometric assembly scheme that extracts the first nine packets of a flow and sequentially arranges them in a predefined geometric layout  $[\cdot]$ . This layout preserves the inherent flow-level contextual relationships within the image space  $\mathbf{x}_i \in \mathbb{R}^{120 \times 120}$ . The number of extracted packets is determined by comprehensively considering prior studies [16], [31]. The entire construction procedure of the flow-level image representation  $\mathbf{x}_i$  is define as follow:

$$\mathbf{x}_i = \begin{bmatrix} x_0 & x_1 & x_2 \\ x_3 & x_4 & x_5 \\ x_6 & x_7 & x_8 \end{bmatrix}, \quad (1)$$

where  $[\cdot]$  denotes geometric assembly operation.

#### B. Contextual Masking Distillation Architecture

1) **Discriminative Representation Extractor**: In the distillation paradigm [48], the teacher network's primary role is to extract discriminative clues that differentiate normal and anomalous samples, thereby guiding a more differentiated

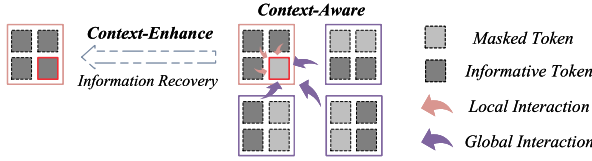


Fig. 5. The interaction illustration of local and global windows attention.

modeling pipeline and effective anomaly detection [22], [23]. Drawing from the established feature extraction techniques on the image anomaly detection domain [46], [47], we employ the pre-trained WideResNet50 [49] as the teacher network for ConMD. Considering that traffic images exhibit a unique pixels distribution is different from natural images, we perform fine-tuning on WideResNet50, which is originally pretrained on ImageNet. To prevent prior information leakage, we fine-tune it on a dataset not associated with malicious network traffic, i.e., ISCX-Tor2016 [50]. Subsequent experiments (as discussed in Section V-E) suggest avoiding excessive fine-tuning epochs, as overfitting can degrade the feature extraction ability inherent in the pre-trained model, ultimately harming the discriminative power of the learned representations.

2) *Context-Aware Student Backbone*: Current distillation techniques focus on local anomalies, failing to capture the inherent contextual relationships within images. However, network flow image data emphasizes a strong contextual nature rather than localized region awareness [11], [24]. In response, we introduce a novel local-global attention mechanism [51] and design a context-aware backbone student network that adequately models the contextual information in flow images. The design of our backbone network is detailed in the blue rectangle in Fig. 3 (a). The network consists of four layers alternating local and global attention mechanisms.

For each input flow image, we first use a  $3 \times 3$  convolutional layer with a stride of 2 to generate tokens, representing parts of the sub-packet information. Then, an MBConv layer [52] is employed to output embeddings for each token  $\mathbf{t}_{emb} \in \mathbb{R}^{h \times w \times c}$ , where  $h, w, c$  denote height, width, and channel of feature maps. Drawn inspiration from a novel local-global attention mechanism [51], we introduce a context-aware module particularly designed for flow image data. The token embeddings  $\mathbf{t}_{emb}$  are divided into multiple local windows  $w_i$  (where  $i$  denotes the index of windows), based on a fixed interval  $m$  (also denotes the window size). This process is illustrated in the Local Windows part of Fig. 3 (a). For token (sub-packet) embeddings within each local windows, we employ a self-attention mechanism layer with position bias in the standard Transformer architecture [53], as shown with a pink arrow in Fig. 5. This **Local Window Attention** mechanism adequately interacts with the short-range dependencies of sub-packet patches, facilitating the capture of intra-packet patterns. The local window attention can be expressed as follows:

$$\text{Local Attention}(\mathbf{q}_w, \mathbf{k}_w, \mathbf{v}_w) = \text{Softmax} \left( \frac{\mathbf{q}_w \mathbf{k}_w}{\sqrt{d}} + \mathbf{b} \right) \mathbf{v}_w, \quad (2)$$

where  $\mathbf{q}_w$ ,  $\mathbf{k}_w$ , and  $\mathbf{v}_w$  are *local query*, *local key* and *local value* matrices output by  $\mathbf{t}_{emb}$  with Fully Connect Layers

in local windows' tokens, respectively.  $d$  is the scaling factor.  $\mathbf{b}$  is a relative position bias term, which is sampled from the position grid  $\mathbb{R}^{(2M-1) \times (2M-1)}$ .

While the local attention perceives the intra-packet patterns, the contextual information of inter-packet relationships, which is crucial in network flow analysis, is not learned. To address this, we additionally design a global window attention. This is achieved by Global Token Generator and Global Attention modules, also as seen in the blue rectangle of Fig. 3. Global Token Generator compresses input token embeddings  $\mathbf{t}_{emb}$  into global tokens using a Fused-MBConv layer [51] and a Max-Pooling layer. A fully connected layer is then used to generate *global query*  $\mathbf{q}_g$ , which aggregates global token information from all local windows, capturing a global representation of the flow. We divide the output tokens after the local attention operation into the same-sized local windows and use fully connected layers to generate *local key*  $\mathbf{k}_l$  and *local value*  $\mathbf{v}_l$ , respectively. The *global query*  $\mathbf{q}_g$  is used to calculate relevance with *local key*  $\mathbf{k}_l$ , and output ultimate attention feature maps. The global window attention is calculated as follows:

$$\text{Global Attention}(\mathbf{q}_g, \mathbf{k}_l, \mathbf{v}_l) = \text{Softmax} \left( \frac{\mathbf{q}_g \mathbf{k}_l}{\sqrt{d}} + \mathbf{b} \right) \mathbf{v}_l, \quad (3)$$

where  $d$  and  $\mathbf{b}$  are defined as in Eq. (2).

As illustrated by the purple arrow in Fig. 5, rich span window information embedded in the *global query*  $\mathbf{q}_g$  provides an effective way of enlarging the receptive field and attending to various local windows. This **Global Window Attention** approach enables adequate interaction of span packet information (i.e., global tokens of all packet regions), significantly facilitating the long-range dependencies awareness of inter-packet patterns. Finally, to align with the teacher network's output, we apply a feature alignment layer, a  $1 \times 1$  convolutional layer, to generate identical dimensions.

3) *Context-Enhanced Masking Strategy*: Masking training is a powerful self-supervised learning approach that enhances structured relationships among features [47]. In this paper, we propose a context-enhanced packet-level masking strategy, which randomly masks entire packets within a flow, as illustrated in the right part of Fig. 4. Fig. 5 details our motivation. Our masking training, in conjunction with local and global attention mechanisms, aims to exploit informative/unmasked tokens (including local tokens within the window and global tokens outside the window) to restore the other uninformative/masked ones. This demasking process cultivates more in-depth inter-packet interactions within flow images, enhancing context awareness capability of the student backbone.

A potential issue in masking training is the risk of **model collapse**, which occurs when an excessive number of masked packets impede the model's ability to recover effectively. To avoid excessive information loss, we design a new masking rule based on packet information. First, we define a packet containing less than 240 bytes as a "weak information packet" (The maximum length is 1600 bytes). Practical distribution statistics could be found in Fig. 6. The amount of weak information packets  $N_{weak}$  follows a diverse distribution, particularly 6, 7, and 0 across three datasets. In the masking rule, the masked number  $N$  is decided by the number of weak infor-

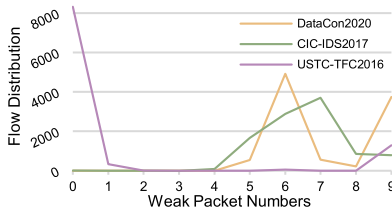


Fig. 6. The statistic distribution about weak information packets number of a flow between on three training dataset.

mation packets  $N_{weak}$ , as outlined in Eq. (4). Our subsequent ablation studies demonstrate that extra masking number  $n$  (i.e., strong information packet) must not be excessively large. An overly high value of  $n$  leads to the excessive masking of strong information packets, which can cause the demasking process to collapse. This collapse results in a failure of context awareness rather than its enhancement. The masked packet number and masked flow sample are defined as follows:

$$N = N_{weak} + n, \quad (4)$$

$$\mathbf{x}_{i\_msk} = \begin{bmatrix} x_0 & x_1 & x_{msk} \\ x_3 & x_{msk} & x_{msk} \\ x_6 & x_{msk} & x_8 \end{bmatrix}, \quad (5)$$

where  $N_{weak}$  is the number of weak-information packets in a flow (see Fig. 6),  $n$  is the extra masking number, and  $x_{msk}$  is the randomly masked packet.

### C. Training and Anomaly Scoring

1) *Training*: In zero-positive anomaly detection, teacher networks' parameters are frozen, and only the student network needs to be trained. The overall training pipeline is shown in Fig. 3 (a), with detailed steps provided in Algorithm 1. The loss function  $L$  is defined as follow:

$$\mathcal{L} = (1 - \cos(\mathbf{z}_t, \mathbf{z}_{s\_msk})) + \alpha * \|\mathbf{z}_t \ominus \mathbf{z}_{s\_msk}\|_2, \quad (6)$$

where,  $\cos(\cdot)$  denotes cosine similarity loss, and  $\|\cdot\|_2$  represents the  $L_2$  norm.  $\alpha$  is the  $L_2$  loss weight coefficient, which is set to 0.1.  $\mathbf{z}_t$  and  $\mathbf{z}_{s\_msk}$  are the output feature vectors from teacher and masked student networks, respectively. The symbol  $\ominus$  denotes element-wise subtraction.

2) *Multi-View Anomaly Scoring*: Effective anomaly scoring during inference is crucial, as it directly quantifies the anomaly degree [25]. Traditional methods rely solely on a masking mechanism to strengthen patch-level contextual connections during training and to infer multi-scale anomalies from non-masked features [46], [47]. However, this approach fails to capture the hierarchical characteristics of network traffic [11], [24]. To address this, we consider network flows data from a multi-view perspective and propose a novel anomaly scoring method that accounts for both packet- and flow-level anomalies in flow images. As shown in Fig. 3 (b), the student network is tasked with generating masked features  $\mathbf{z}_{s\_msk}$  and unmasked features  $\mathbf{z}_{s\_nmsk}$  to perceive anomalies from the packet- and flow-level perspectives, respectively. These two branches are fused using a weighted dot-add operation to further combine and amplify anomalous deviations. In this

### Algorithm 1 Contextual Masking Distillation Training

**Input:** Training set  $\mathcal{X}$

**Output:** Student network *Student* with parameters  $\theta_s$

- 1: Initialize the network parameters of student  $\theta_s$ , and fix the network parameters of teacher  $\theta_T$ .
- 2: **for** epochs **do**
- 3:   Sample a mini-batch data  $\mathcal{X}_b$  from  $\mathcal{X}$
- 4:   **for** each mini-batch **do**
- 5:     *Discriminative Representation Extraction*
- 6:     Extract teacher representations:  $\mathbf{z}_t = \text{Teacher}(\mathcal{X}_b)$
- 7:     *Packet-Level Masking and Token Generation*
- 8:     Apply packet masking:  $\mathcal{X}_{b\_msk} = \text{Packet Masking}(\mathcal{X}_b)$  from Eq. (5)
- 9:     Generate token embeddings:  $\mathbf{t}_{emb} = \text{Conv}(\mathcal{X}_{b\_msk})$
- 10:     *Contextual Interaction and Awareness*
- 11:     /\*Local Windows\*/
- 12:      $\mathbf{t}_{local} = \text{Local Attention}(\mathbf{t}_{in})$  use Eq. (2)  $\Rightarrow$  where  $\mathbf{t}_{in}$  denotes  $\mathbf{t}_{emb}$  or  $\mathbf{z}_{s\_msk}$
- 13:     /\*Global Windows\*/
- 14:      $\mathbf{z}_{s\_msk} = \text{Global Attention}(\mathbf{t}_{local})$  use Eq. (3)
- 15:     *Network Parameters Updating*
- 16:     Compute loss:  $\mathcal{L}$  for student network from Eq. (6)
- 17:     Update parameters  $\theta_s$  with Adam Optimizer
- 18:   **end for**
- 19: **end for**
- return** Trained student network *Student*

paper, we use cosine similarity to indicate the anomaly degree. The detailed calculation is defined as:

$$Score = (1 - \cos(\mathbf{z}_t, (\beta * \mathbf{z}_{s\_msk} \oplus (1 - \beta) * \mathbf{z}_{s\_nmsk}))), \quad (7)$$

where  $\beta$  is the weight coefficient for two-branch feature of the student network, and  $\oplus$  denotes the dot-add operation for two-branch feature vectors.  $\cos(\cdot)$  denotes cosine similarity.

Fig. 3 (c) illustrates the motivation behind our multi-view anomaly awareness approach. For normal samples, due to training set fitting, the feature embeddings produced by both the teacher and student networks are highly similar, showing no anomalous patterns (i.e., absence of red embeddings). In contrast, for anomalous samples, the teacher consistently generates anomalous patterns (i.e., deep red embeddings), whereas the student tends to generate embeddings closer to normal patterns (i.e., shallow red and yellow embeddings). Crucially, the greater the deviation between teacher and student embeddings, the easier it is for the model to indicate anomalies. Our approach achieves this deviation by imposing both packet- and flow-level anomaly awareness.

During inference, the non-masked representation  $\mathbf{z}_{s\_nmsk}$  is typically used to infer enhanced information [47]. Since adequate context is learned through contextual masking,  $\mathbf{z}_{s\_nmsk}$  also benefits from richer contextual representation [11]. In anomaly detection, it often causes anomalous features to be largely transformed toward normal patterns, as illustrated by the shallow red embeddings in “Flow-level Anomalies” in Fig. 3 (c). Consequently, this process highlights **flow-level anomaly awareness**, exposing the global deviation in inter-packet relationships. In contrast, the masked representation

$\mathbf{z}_{s\_msh}$  serves a distinct role. For anomalous samples, by leveraging the appropriately designed packet-level masking (e.g., the recovery process in Fig. 5), the student network removes anomalous packet information and restores these tokens to resemble normal representations. This design is particularly effective for packet-level anomalies, as shown in “Packet-level Anomalies” of Fig. 3 (c). It emphasizes deviations of masked packets, further strengthening **packet-level anomaly awareness**. Our anomaly scoring (Eq. (7)), developed for multi-view anomaly awareness, simultaneously considers anomalies from both packet- and flow-level perspectives through a weighted two-branch fusion. Multi-view anomaly awareness combines the strengths of both packet-level and flow-level anomalies, making anomalous samples appear closer to normal patterns. This further amplifies the framework’s ability to differentiate normal and anomalous flows at the anomaly metric, enabling more accurate detection.

#### D. Complexity Analysis

Model complexity is a key consideration in the online deployment of detection systems. In this section, we analyze the computational complexity of the proposed modules.

- Complexity of context-aware student backbone. Given an input feature map of  $\mathbf{t}_{emb} \in \mathbb{R}^{h \times w \times c}$  at each local-global attention mechanism with a window size of  $m \times m$ , the computational complexity  $\mathcal{O}(Window)$  within local windows remains in the same order as that of the standard Vision Transformer (ViT):

$$\mathcal{O}(Window) = \mathcal{O}(4m^2c^2 + 2m^4c). \quad (8)$$

Therefore, per-layer complexity  $\mathcal{O}(LGA)$  of the student backbone can be expressed as:

$$\begin{aligned} \mathcal{O}(LGA) &= \mathcal{O}\left((4m^2c^2 + 2m^4c) \times \left(\frac{h}{m} \times \frac{w}{m}\right)\right) \\ &= \mathcal{O}(4hwc^2 + 2hwm^2c). \end{aligned} \quad (9)$$

Obviously, compared with the standard ViT complexity  $\mathcal{O}(4hwc^2 + 2h^2w^2c)$ , our local-global attention design achieves greater computational efficiency by partitioning the feature map into multiple local windows.

- Complexity of context-enhanced masking training strategy. The masking operation is applied to each sub-packet region within a flow. Its computational cost  $\mathcal{O}(Mask)$  is negligible, with an index cost of:

$$\mathcal{O}(Mask) = \mathcal{O}(3^2) = \mathcal{O}(9). \quad (10)$$

- Complexity of anomaly metric. ConMD adopts a multi-view anomaly awareness mechanism, generating both masked and unmasked packet tokens to compute the anomaly metric during inference. Consequently, the computational complexity of this process increases to  $2 \times \mathcal{O}(Student)$  compared with the training phase.

### V. EXPERIMENTS

#### A. Experimental Setting

1) *Evaluation Datasets*: We use three publicly available network traffic anomaly detection datasets. (1) *DataCon2020*

[19] is an encrypted malware network traffic dataset. White traffic is generated by normal software (all exe types), while black traffic consists of encrypted traffic generated by malware (all exe types). (2) *CIC-IDS2017* [54] is an encrypted network intrusion detection dataset that includes seven common attacks, such as Brute Force Attack, Heartbleed Attack, Botnet, DoS, DDoS, Web Attack, and Infiltration Attack. (3) *USTC-TFC2016* [10] is an encrypted malware traffic detection dataset, where the malicious traffic is collected from public websites and normal traffic is collected from applications. For consistent evaluation, we randomly sample 10,000 normal flows as training set, as well as 5,000 normal and 5,000 anomalous flows as testing set.

2) *Fine-tuning Dataset*: We pay particular attention to potential information leakage during fine-tuning, as overly homogeneous attack categories can lead to domain overfitting. This would compromise evaluation fairness compared with other baselines by introducing prior information specific to certain attack-domain datasets. To mitigate this, we fine-tune the teacher network for binary classification on a dataset unrelated to malicious traffic, specifically *ISCX-Tor2016* [50]. This dataset contains abundant high-frequency information and is widely used for fine-tuning [11], [24]. The fine-tuning data includes 100,000 Tor and non-Tor traffic samples.

3) *Evaluation Metrics*: In alignment with recent works in network traffic anomaly detection [13], [16], we evaluate detection performance using Area Under the ROC Curve (AUC), Accuracy (ACC), Macro F1-Score (F1), Precision (Pre) and Recall (Rec) metrics.

4) *Implementation Details*: All experiments are conducted on an NVIDIA GeForce RTX 3090 GPU. To fine-tune the teacher network, we use the AdamW optimizer with a learning rate of  $1 \times 10^{-4}$  and a cross-entropy loss function. The student backbone is constructed by alternating four layers of local and global window attention, where each window size  $m$  is set to 4. During distillation training, we use the AdamW optimizer with a weight decay of 0.05 and a batch size of 128. We perform a grid search for learning rates. The learning rates is finally set to  $2 \times 10^{-5}$ ,  $6 \times 10^{-5}$ , and  $3 \times 10^{-4}$  for the three datasets, respectively. The maximum training epoch is 50, and early stopping is applied to prevent overfitting. We repeat each experiment with five different random seeds to report the standard deviation.

5) *Baselines*: Given that network traffic exhibits inherent temporal dependencies, time-series anomaly detection methods can also be relevant to this task. Accordingly, we compare ConMD with four categories of methods, covering eight state-of-the-art and two classical baselines: (1) Network traffic anomaly detection methods, such as *GANomaly* [40], *ARCADE* [17], *MFAD* [18]; (2) Knowledge distillation techniques, including *STFPM* [44], *ReverDis* [22], and *MMR* [47]; (3) Time-series anomaly detection methods, like *AnoFormer* [55] and *TSLANet* [56]; and (4) Classical machine learning methods, such as *Isolation Forest (IF)* [57] and *OCSVM* [58]. We perform a learning rate search for each baseline as specified in the original papers. All implementation details are as follows:



- GANomaly [40] is a classical comparison method used in network traffic anomaly detection with an encoder-decoder-encoder architecture. Adversarial training strategy is especially used to improve the reconstruction quality of normal samples.
- ARCADE [17] is to incorporate the WGAN-GP strategy to regularize and suppress the reconstruction of anomalous network flows. To receive the same dimensional images input, we use the 2-D convolution layer to build an encoder-decoder architecture.
- MFAD [18] leverages low-pass filters to extract texture features to benefit reconstruction modeling, enabling normal and anomalous traffic to be more distinguishable.
- STFPM [44] employs an identical student-teacher architecture with multi-scale feature matching to detect anomalies across different scales.
- ReverDis [22] proposes an asymmetric distillation framework to enhance the diversity of representations for anomalous samples. A one-class bottleneck embedding module is used to retain essential sample information.
- MMR [47] utilizes a pre-trained Vision Transformer as the student network and introduces a token-level masking mechanism to improve causal relationships.
- AnoFormer [55] highlights prior and series-wise associations in time-series data, and proposes an anomaly-attention block, which incorporates a self-attention mechanism and series characteristics.
- TSLANet [56] applies Fourier analysis to improve the capture of long- and short-term dependencies while reducing noise. It further incorporates self-supervised learning to model complex periodic patterns.
- Isolation Forest (IF) [57] constructs random decision trees to isolate anomalies, where anomalous samples typically exhibit shorter path lengths.
- OCSVM [58] encapsulates normal samples within a decision boundary and employs a decision function to reliably identify anomalous samples.

### B. Anomaly Detection Performance

1) *Comparison With SOTA Methods*: From detection results presented in Table I, we have the following observations:

(O1): ConMD consistently outperforms all strong baselines across various datasets, achieving superior results in AUC, ACC, and F1 metrics. Notably, ConMD outperforms the most competitive baseline by 2.8% and 5.1% in AUC on the DataCon2020 and CIC-IDS2017 datasets, respectively. Even on USTC-TFC2016, we achieve 99.97% in AUC and 99.00% in ACC, significantly surpassing supervised-learning's performances [24]. These improvements are attributed to our particular design for network flow data. Specifically, our teacher network provides more discriminative representation and guide student network with a more differentiated modeling pipeline. The context-aware student backbone is designed to capture crucial information for normal network samples, enhancing the differences between normal and anomalous network traffic. Moreover, the multi-view anomaly scoring further improves detection accuracy from packet- and flow-level perspectives.

TABLE I  
PERFORMANCE COMPARISONS (%) ON DATACon2020, CIC-IDS2017 AND USTC-TFC2016 DATASETS. THE BEST RESULTS ARE IN BOLD FONT AND THE SECOND UNDERLINED

	Models	AUC	ACC	F1	Pre	Rec
DataCon2020	IF	74.85±0.9	73.10±1.1	75.38±1.3	69.49±2.3	82.83±1.0
	OCSVM	52.44±2.0	63.18±3.1	69.89±1.1	59.11±3.7	85.48±4.2
	AnoFormer	51.39±0.3	52.69±0.2	67.74±0.3	51.36±0.2	<b>99.37±0.4</b>
	TSLANet	62.76±0.4	58.39±0.2	68.62±0.5	<b>91.01±0.7</b>	55.07±0.7
	STFPM	82.37±0.6	80.44±2.1	<u>80.93±1.7</u>	79.19±3.2	82.83±0.6
	ReverDis	74.53±2.4	68.80±3.1	75.30±1.1	62.71±3.7	<u>94.81±4.2</u>
	MMR	80.60±2.1	78.85±2.1	79.64±1.3	77.12±3.8	82.55±2.0
	GANomaly	81.50±1.0	79.40±2.1	79.95±1.6	78.16±3.7	82.02±2.4
	ARCADE	81.98±5.1	<u>81.48±2.0</u>	80.31±3.4	80.79±6.8	82.28±4.7
	MFAD	<u>83.16±1.9</u>	76.28±2.4	78.59±1.1	72.04±3.7	86.87±3.8
	<b>ConMD</b>	<b>86.03±0.1</b>	<b>82.58±0.1</b>	<b>82.90±0.1</b>	<u>81.41±0.4</u>	84.44±0.6
CIC-IDS2017	IF	55.27±0.4	53.88±0.3	68.23±0.5	52.04±1.1	<b>99.06±1.2</b>
	OCSVM	59.59±0.6	59.11±0.4	70.59±0.4	55.12±0.4	98.14±0.7
	AnoFormer	63.37±0.7	76.99±0.6	71.92±0.7	58.94±0.4	92.23±0.7
	TSLANet	84.45±1.7	77.57±1.9	80.78±1.5	<b>94.30±3.3</b>	70.75±2.1
	STFPM	86.29±1.7	80.00±2.1	81.50±1.1	76.26±3.6	87.83±3.2
	ReverDis	84.15±0.6	83.78±0.4	85.51±0.4	77.26±0.4	95.74±0.7
	MMR	<u>89.26±1.2</u>	<u>86.09±1.1</u>	<u>86.97±1.0</u>	81.86±1.8	92.86±2.4
	GANomaly	82.75±5.6	80.85±1.7	83.21±0.9	74.46±3.7	94.80±4.8
	ARCADE	84.85±2.6	80.15±1.6	82.78±1.0	73.29±2.6	95.32±3.0
	MFAD	86.02±0.8	81.66±1.9	83.67±1.7	75.45±1.8	93.96±3.0
	<b>ConMD</b>	<b>94.43±0.1</b>	<b>90.04±0.2</b>	<b>90.41±0.2</b>	<u>87.17±0.2</u>	93.90±0.3
USTC-TFC2016	IF	79.92±1.2	83.34±0.9	85.03±1.1	77.19±1.3	94.64±0.9
	OCSVM	62.32±0.8	69.36±0.6	67.94±1.2	71.24±1.1	64.94±1.0
	AnoFormer	88.47±0.1	92.74±0.2	93.18±0.2	87.71±0.2	<u>99.38±0.2</u>
	TSLANet	98.64±0.2	<u>97.63±0.6</u>	<u>97.64±0.6</u>	<b>98.88±0.6</b>	97.72±0.9
	STFPM	91.63±2.4	89.02±1.3	89.71±1.3	84.29±0.8	95.90±2.1
	ReverDis	98.05±0.5	95.21±0.9	95.07±0.9	98.02±1.4	92.30±0.9
	MMR	99.44±0.0	96.15±0.2	96.04±0.2	98.62±0.7	93.33±0.4
	GANomaly	95.36±1.0	91.27±2.9	91.07±3.2	93.01±4.1	89.52±5.7
	ARCADE	88.62±2.2	93.13±0.1	93.57±0.1	87.94±0.1	<b>99.97±0.0</b>
	MFAD	<u>99.73±0.0</u>	97.45±0.4	97.43±0.4	98.43±0.4	96.44±0.7
	<b>ConMD</b>	<b>99.97±0.0</b>	<b>99.00±0.2</b>	<b>99.01±0.2</b>	<u>98.74±0.5</u>	99.34±0.2

(O2): The machine learning methods, including IF and OCSVM, perform poorly, revealing their inherent limitations. Network traffic anomaly detection and distillation-based methods also perform markedly worse than ConMD. Although some methods, such as GANomaly and ARCADE, employ adversarial training strategies in enhancing reconstruction quality, they still fail to address the “Confused-to-Confused” anomaly scoring problem in encrypted traffic. Transformer-based architectures like AnoFormer and TSLANet also suffer, as encryption obscures time-series characteristics and results in confused representations. MFAD attempts to mitigate these issues through low-pass filtering, but risks information loss. Distillation-based methods exhibit superior performance on the CIC-IDS2017 dataset compared with reconstruction-based approaches. This phenomenon indicates the effectiveness of distillation framework. Unfortunately, existing distillation techniques overemphasize local anomaly awareness, resulting in suboptimal overall performance across datasets.

(O3): A notable observation is that, for most baselines, the Precision metric is explicitly lower than the Recall metric. All test samples are uniformly identified anomalous category. This finding highlights the limitations of existing methods in distinguishing normal and anomalous flows, which, in turn, results in significant false positives. Reconstruction-based methods are limited by a “Confused-to-Confused”



TABLE II

CATEGORY-SPECIFIC DETECTION ACCURACY (%) ON CIC-IDS2017 DATASET. ANOMALY LABELS REFER TO VARIOUS ATTACK CATEGORIES. CONMD-P INDICATES OUR PACKET-LEVEL BRANCH, AND CONMD-F INDICATES THE FLOW-LEVEL BRANCH

Model	Benign 5000	Brute 882	Dos 2280	DDoS 1764	Scan 10	Web 39	Botnet 18	Infiltrat 6	Heartb 1
STFPM	71.9	93.8	81.0	96.1	<b>100</b>	51.3	77.8	50.0	0
ReverDis	71.8	<b>100</b>	92.2	98.2	<b>100</b>	<b>100</b>	83.3	<b>100</b>	0
MMR	82.0	<b>100</b>	82.7	<b>100</b>	<b>100</b>	25.6	0	33.3	0
GANomaly	73.6	91.5	79.9	<b>100</b>	<b>100</b>	0	22.2	16.7	0
ARCADE	63.5	98.9	<b>96.3</b>	<b>100</b>	<b>100</b>	46.2	94.4	33.3	0
MFAD	65.8	93.7	89.1	94.0	<b>100</b>	28.2	<b>100</b>	83.3	<b>100</b>
ConMD-P	63.5	99.3	93.2	<b>100</b>	<b>100</b>	43.6	66.6	<b>100</b>	0
ConMD-F	<b>84.0</b>	98.8	92.8	98.4	<b>100</b>	12.8	22.2	<b>100</b>	0
ConMD	<b>84.0</b>	<b>100</b>	93.9	<b>100</b>	<b>100</b>	28.2	72.2	<b>100</b>	0

TABLE III

CATEGORY-SPECIFIC DETECTION ACCURACY (%) AND NUMBER STATISTICS ON USTC-TFC2016 DATASET. ANOMALY LABELS INDICATE THE TYPES OF MALICIOUS APPLICATIONS. CONMD-P INDICATES OUR PACKET-LEVEL BRANCH, AND CONMD-F INDICATES THE FLOW-LEVEL BRANCH

Model	Benign 5000	Zeus 63	Geodo 562	Shifu 37	Miuref 97	Cridex 2839	Htbot 333	Neris 627	Nsis-ay 100	Virut 342
STFPM	83.0	82.5	99.8	<b>100</b>	<b>100</b>	<b>100</b>	81.4	77.5	96.0	69.6
ReverDis	94.5	98.4	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	95.5	82.1	88.0	79.5
MMR	98.6	74.6	99.6	<b>100</b>	<b>100</b>	<b>100</b>	72.4	81.2	74.0	77.2
GANomaly	94.7	55.6	96.6	<b>100</b>	<b>100</b>	<b>100</b>	70.0	65.1	65.0	54.4
ARCADE	86.2	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>99.7</b>	<b>100</b>	<b>100</b>
MFAD	<b>98.6</b>	88.9	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	96.1	83.7	72.0	80.7
ConMD-P	97.08	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	97.8	99.2	<b>100</b>	97.4
ConMD-F	97.82	96.8	95.8	<b>100</b>	<b>100</b>	<b>100</b>	94.2	93.1	95.0	92.4
ConMD	<b>98.6</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	97.9	98.9	<b>100</b>	96.5

detection pipeline, while distillation-based methods lacks effective anomaly awareness. Benefiting from extensive designs to amplify differences between normal and abnormal flows, our method exhibits more reliable detection with higher F1 scores and more balanced Precision and Recall metrics.

2) *Category-Specific Detection*: To evaluate the sensitivity of various attack types, we count the accuracy of category-specific anomaly detection. Since the DataCon2020 dataset lacks of category-specific labels, we only assess the results on CIC-IDS2017 and USTC-TFC2016 dataset, as shown in table II and III. Overall, ConMD achieves superior accuracy in classifying benign/normal traffic. This outcome indicates that our context-aware framework effectively captures the inherent patterns of normal network flows, thereby enhancing the correct identification of normal samples. Moreover, due to its discriminative extraction and comprehensive anomaly awareness, our method also achieves superior detection for the majority of category-specific anomalies. On CIC-IDS2017, common attack types like Brute Force, Dos, DDos, and Port Scan are detected by most methods due to their distinctive high-rate malicious request patterns [54]. In contrast, low-concurrency and penetration-oriented attacks that disguise themselves as normal behavior, such as Web attack, Botnet, Heartbleed, and Infiltration, are considered hard negative samples. While ARCADE exhibits nearly 100 % detection accuracy in malicious application types, it suffers from poor identification of normal network flows. This result suggests that ARCADE is suboptimal, as it tends to classify all samples into the anomalous category. In contrast, our

method demonstrates more reliable detection performance, as it can effectively distinguish anomalous samples from the normal ones. In particular, the packet-level branch (ConMD-P) enhances anomaly sensitivity, while the flow-level branch (ConMD-F) improves normal flow discrimination. Collectively, they enable a more balanced multi-view anomaly awareness.

### C. Paradigm Analysis

To verify the effectiveness of representation extraction in our distillation approach for encrypted network traffic, we conducted a qualitative t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization to compare the embedding distributions between reconstruction-based paradigm (Fig. 7 (a)) and distillation-based paradigm (Fig. 7 (b)). Traffic flows typically exhibit encrypted high-frequency characteristics, which obscure inherent meaningful patterns [18]. Our experimental results (Fig. 7 (a)) further confirm this challenge: normal and anomalous flows display blurred boundaries and significant confusion. Such representations inevitably constrain the modeling process, leading to the “Confused-to-Confused” pipeline. As illustrated in Fig. 2, the reconstruction-based paradigm is inherently guided to reproduce the original profile of network flows—that is, the confused class distribution. This has already constrained the reliability of anomaly scores and constitutes a fundamental limitation of the reconstruction-based paradigm.

In contrast, the distillation-based paradigm prioritizes extracting discriminative representation to guide a more effective modeling process. This motivation is confirmed by Fig. 7 (b). We can observe a clearer boundary and a more separated distribution for normal and anomalous samples outputted by our teacher network. Such discriminative embeddings refine representations, suppress high-frequency noise, and enable a differentiated rather than confused pipeline (Fig 2). This constitutes the fundamental advantage of the distillation-based paradigm, which directly addresses the limitations of reconstruction-based approaches.

### D. Ablation Studies

Table IV presents ablation studies evaluating the contribution of each ConMD component by individually removing or replacing the masking strategy, anomaly scoring, architecture design, or fine-tuning dataset.

1) *Effect of Masking Training Strategy*: Overall, masking training proves effective within our framework. As a common used technique [47], random masking (w/ RMT) performs better performances than non-masked training (w/o MT). However, this strategy has a significant limitation: it is unsuitable for flow images because it neglects the contextual information inherent in network flows. In comparison, our method outperforms the random masking strategy (w/ RMT) with considerable performance advantages. This improvement is attributed to our context-enhanced design. Packet-level masking strategy effectively cultivates more in-depth inter-packet interaction within flow images, enhancing the context-aware capability of the student backbone.

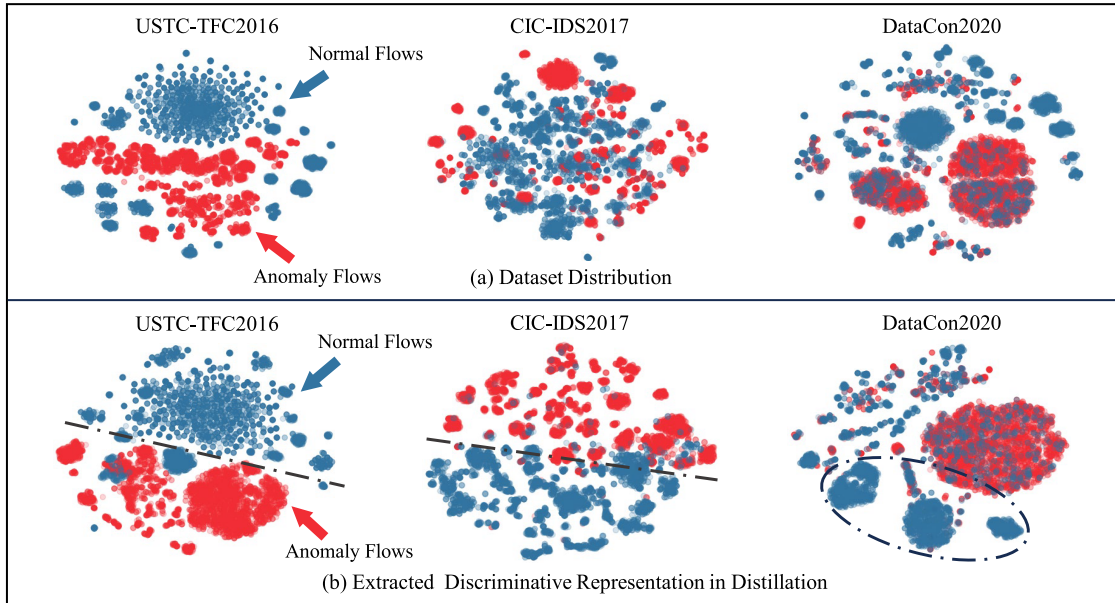


Fig. 7. The t-SNE visualization comparison for reconstruction-based and distillation-based paradigm among three datasets. (a) presents the dataset distribution both in reconstruction-based paradigm and distillation-based paradigm; (b) presents the extracted discriminative representation in distillation-based paradigm.

TABLE IV

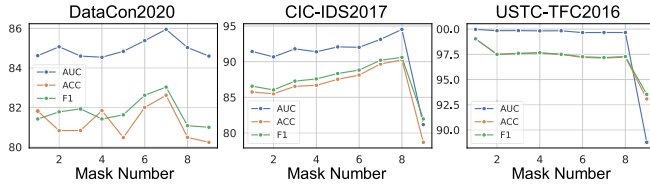
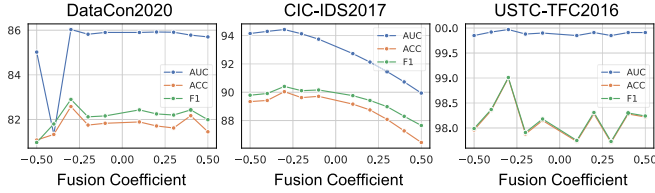
ABLATION STUDIES ON MASKING STRATEGY, ANOMALY SCORING AND ARCHITECTURE DESIGN, WITH THE RESULTS SHOWN IN % FORMAT. THE ABBREVIATIONS ARE EXPLAINED AS FOLLOWS MT: MASKING TRAINING, RMT: RANDOM MASKING TRAINING, PA: PACKET-LEVEL ANOMALY AWARENESS, FA: FLOW-LEVEL ANOMALY AWARENESS, ViT: VISION TRANSFORMER, RES: WIDERESNET50, LGA: LOCAL-GLOBAL WINDOW ATTENTION, AND VPN FT: VPN DATASET FINE-TUNING

Method	DataCon2020			CIC-IDS2017			USTC-TFC2016		
	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
w/o MT	84.92	81.07	81.90	93.26	87.04	87.19	99.91	98.17	98.18
w/ RMT	84.79	80.74	81.61	92.29	85.25	86.04	99.91	98.08	98.11
w/o PA	<u>85.93</u>	81.91	82.40	93.28	<u>89.45</u>	<u>89.99</u>	99.86	97.79	97.82
w/o FA	84.97	80.75	81.10	83.76	79.87	82.69	<u>99.91</u>	<u>98.31</u>	<u>98.32</u>
w/ ViT-Res	83.65	81.86	81.31	90.70	88.35	89.16	99.88	97.84	97.85
w/ Res-Res	83.52	80.39	81.57	89.24	84.15	85.94	99.84	97.80	97.92
w/ LGA-LGA	79.25	80.84	81.97	60.60	62.79	72.46	93.75	90.90	91.59
w/ VPN FT	85.43	<u>82.28</u>	<u>82.61</u>	92.42	89.03	89.27	99.87	98.17	98.24
<b>ConMD</b>	<b>86.03</b>	<b>82.58</b>	<b>82.90</b>	<b>94.43</b>	<b>90.04</b>	<b>90.41</b>	<b>99.97</b>	<b>99.00</b>	<b>99.01</b>

2) *Effect of Multi-View Anomaly Scoring*: The removal of flow-level scoring (w/o FA) results in substantial performance degradation, underscoring the critical role of flow-level anomaly awareness in network flows, especially on DataCon2020 and CIC-IDS2017 datasets. In addition, we observe a more notable decline in performance upon removal of packet-level scoring (w/o PA) compared to flow-level scoring (w/o FA) on USTC-TFC2016 dataset. These results confirm that these two scoring mechanisms possess different anomaly awareness capabilities. Flow-level awareness focuses on contextual anomalies, while packet-level awareness highlights anomalies in the masked packet region. By integrating these perspectives, our model combines their strengths and achieves superior results. Notably, although packet-level scoring (w/o FA) exhibits poor performance on the CIC-IDS2017 dataset, it still contributes to enhancing the overall

performance of flow-level scoring (w/o PA) through our fusion approach.

3) *Effect of Various Student-Teacher Architecture*: ViT-based (w/ ViT-Res) student backbones exhibit only a marginal improvement over WideResNet50 (w/ Res-Res), indicating the limited effectiveness in capturing network characteristics. In contrast, our local-global window attention mechanisms yields significant gains by capturing both intra-packet and inter-packet interactions at packet and flow levels. Furthermore, we also explored the feasibility of local-global attention modules as a teacher network (w/ LGA-LGA). Unfortunately, Table IV reveals that the LGA-LGA framework suffers from a significant defect. As discussed in study [25], student network tends to over-generalize and output similar feature representations to those of the teacher network, regardless of normal and anomalous samples. The identical local-global attention in


 Fig. 8. Experience on the masked packet number  $N$ .

 Fig. 9. Experience on the coefficient  $\beta$  for anomaly awareness fusion.

both student and teacher networks may exacerbate this issue, leading to non-discriminative representations.

4) *Effect of Various Fine-Tuning Dataset:* We replace the fine-tuning dataset with ISCX-VPN2016 [59] (w/ VPN FT), which contains encrypted but non-malicious traffic. This substitution avoids introducing attack information leakage while preserving traffic characteristics for fair fine-tuning. The results indicate that the replacement produces nearly consistent performance across all three datasets. This shows both the general applicability of using encrypted traffic for fine-tuning and the overall effectiveness of the distillation framework.

### E. Hyperparameter Analysis

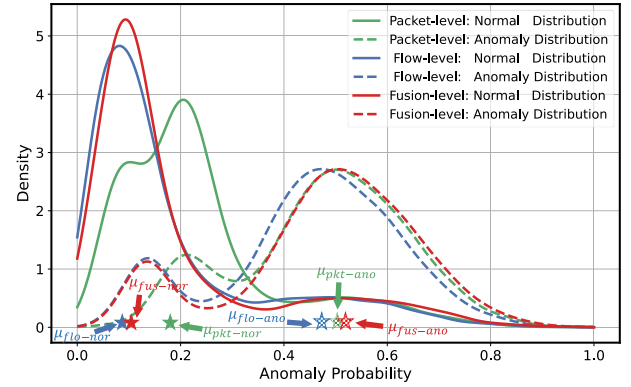
1) *The Amount of Packets With the Optimal Masking  $N$ :* The number of packets  $N$  selected for masking is closely related to the information strength contained within each packet, as shown in Eq. (4). In Fig. 8, the best performances were achieved with  $N = 7, 8$ , and 1 (i.e., the extra masking  $n = 1$ ) on three datasets, respectively. This result indicates that an insufficient number of masked packets fails to support effective demasking training, leading to suboptimal context enhancement for the student backbone. Conversely, an excessive number of masked packets within a flow adversely affects modeling, potentially causing the demasking training to collapse. From a statistical view, the lack of sufficient informative tokens renders the recovery of numerous masked or missing packets infeasible. Correspondingly, masking all packets (i.e., setting  $N$  to the maximum 9) leads to a substantial decrease in detection performance across all test datasets. Based on our experiment results, optimal context-enhanced recovery is achieved when the number of masked packets  $N$  exceeds the number of weakly informative packets  $N_{weak}$  by one.

2) *The Optimal Anomaly Scoring Fusion Coefficients  $\beta$ :* We searched for fusion parameter  $\beta$  within the range from  $-0.5$  to  $0.5$ , as detailed in Fig. 9 (where 0 denotes flow-level anomaly awareness branch in Table IV). The best results are achieved at  $\beta = -0.3$  for feature fusion. Align with findings in Table IV, this search result confirms that flow-level anomaly awareness plays a more crucial role in detecting anomalous network

TABLE V

EXPERIENCE ON THE FINE-TUNING EPOCH (AUC RESULT REPORTED WITH % FORMAT). EPC DENOTES THE NUMBER OF EPOCHS

Dataset	Epc-0	Epc-1	Epc-2	Epc-3	Epc-4	Epc-5
DataCon2020	79.24	<b>86.03</b>	64.09	61.54	63.23	62.22
CIC-IDS2017	87.48	<b>94.43</b>	67.12	65.85	69.06	63.26
USTC-TFC2016	99.18	<b>99.97</b>	97.29	96.67	95.77	95.56


 Fig. 10. The density distribution of tested samples among anomaly score branches on DataCon2020.  $\mu_*$  denotes the mean of each density distribution.

traffic. While packet-level awareness can detect packet-level anomalies by incorrectly repairing masked anomaly packets, traffic flows are essentially sequential data. Anomalies in traffic flows are mainly manifested through the interactions between packets (i.e., contextual anomalies).

3) *Fine-Tuning Scope of Teacher Network:* There is a potential risk that traffic data may diverge from the natural image distribution [18]. To mitigate this issue, we fine-tuned the teacher network, which was initially pretrained on ImageNet. As illustrated in Table V, it is crucial to strike a balance between leveraging pretrained knowledge and adapting the network to traffic data distributions. Epc 0 represents the model as pretrained on ImageNet without any fine-tuning. Our experiments indicate that even a single epoch of fine-tuning leads to an increase in AUC value, indicating that the fine-tuned teacher network is more effective in guiding the student network on flow data. However, the results also reveal that AUC values begin to decline after two epochs of fine-tuning. This suggests that excessive fine-tuning can lead to teacher network overfitting on the traffic data, causing the network to lose its capacity to extract diverse features effectively.

### F. Multi-View Anomaly Scoring Analysis

The previous quantitative experiments have demonstrated that the fused multi-view anomaly awareness can further improve detection performance. To provide in-depth insights, we qualitatively visualize the fusion process to better understand its working principles, as illustrated in Fig. 10, 11 and 12.

Fig. 10 and 11 suggest a lower anomaly score for flow-level (blue color) than packet-level (green color) in normal samples. This result confirms that flow-level awareness better understands network flow patterns and effectively quantifies anomalous distributional deviations. In contrast, anomaly



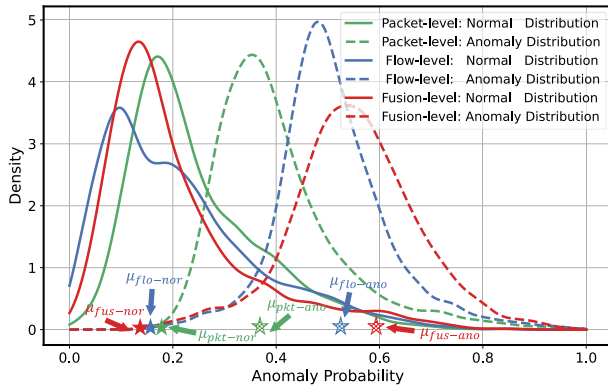


Fig. 11. The density distribution of tested samples among anomaly score branches on CIC-IDS2017.  $\mu_*$  denotes the mean of each density distribution.

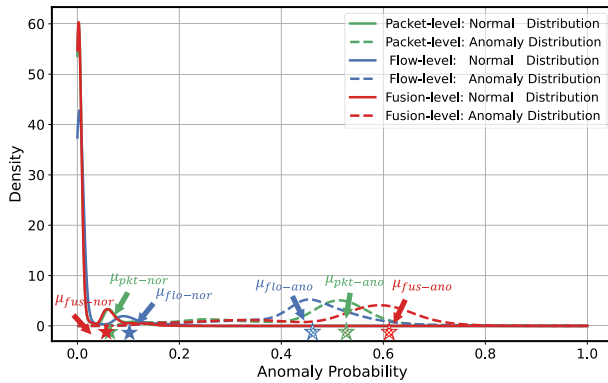


Fig. 12. The density distribution of tested samples among anomaly score branches on USTC-TFC2016.

score distribution of packet-level anomaly awareness (green color) exists significant prediction confusion (i.e., normal and anomaly distributions have the closest mean values). Despite these non-ideal results, packet-level awareness can still contribute to flow-level detection. Fusion-level awareness (red color) integrates multiple perspectives' strengths to highlight anomalous deviations, amplifying deviations of anomaly samples (i.e., the mean value of anomaly distribution significantly shifts to the right). This approach provides greater differentiation and reduced overlap between the normal and anomalous distribution regions, thereby enabling more accurate detection.

On USTC-TFC2016 dataset, shown in Fig. 12, packet-level awareness leads to more pronounced deviations compared to flow-level awareness, demonstrating a reverse phenomenon. This indicates that flow-level anomaly awareness is not always the dominant factor. Packet-level awareness on USTC-TFC2016 is significantly important. However, compared to single perspectives, a key observation is that the fusion-level distribution is more distinctly differentiated. This further validates the rationale for multi-view awareness. Our fusion scheme proves to be an effective and promising approach for anomaly evaluation in traffic data.

### G. Efficiency Analysis

Model efficiency is critical for practical deployment. Following prior work [60], we compare the computational overhead among distillation-based, reconstruction-based, and

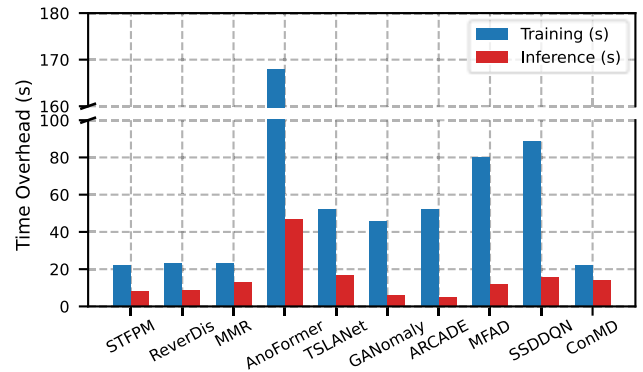


Fig. 13. Overhead comparison in training and inference time.

conventional network classification methods. All methods are uniformly conducted on an Ubuntu 18.04 system with an NVIDIA GeForce RTX 3090 GPU. We report the training and inference time per epoch (in seconds) for a complete efficiency analysis, as shown in Fig. 13.

For training overhead, network traffic anomaly detection methods such as GANomaly, ARCADE, and MFAD incur approximately twice the time of ConMD. This is because their use of additional adversarial training significantly increases the training overhead. Transformer-based models, including AnoFormer and TSLANet, also exhibit high computational costs due to the dimensional explosion in attention mechanisms. In contrast, ConMD processes traffic as image patches, which significantly alleviates this issue. Furthermore, when compared with distillation-based methods such as STFCPM, ReverDis, and MMR, ConMD achieves the most optimal training overhead, highlighting the efficiency of our designed context-aware modules and context-enhanced training mechanisms. Interestingly, even when compared with the common network traffic classification method SSDDQN [60], ConMD consistently maintains optimal training overhead.

For inference overhead, although ConMD is not the most efficient, it remains highly competitive and holds considerable potential. The additional cost primarily arises from the multi-view anomaly scoring mechanism. This design greatly enhances detection performance and, at the same time, achieves a balanced trade-off between accuracy and efficiency. Furthermore, it also provides a solid foundation for future optimization. In efficiency-prioritized scenarios, these branches can be deployed in parallel to further enhance inference speed.

## VI. CONCLUSION

In this work, we introduced ConMD, a novel knowledge distillation framework designed to address the confused modeling problem in zero-positive network traffic anomaly detection. Our approach excels in two key aspects: (1) During training, ConMD effectively captures contextual information through a context-aware backbone and a context-enhanced training strategy, facilitating a deeper understanding of normal network behavior; and (2) During inference, ConMD fuses packet- and flow-level representations through a multi-view scoring mechanism for a more comprehensive anomaly assessment. Extensive quantitative and qualitative experiments validate

the superiority of our distillation framework in fostering a more differentiated modeling pipeline and enhancing detection capabilities compared to existing methods.

In future work, we plan to explore more lightweight distillation architectures to further improve inference efficiency. We also intend to evaluate ConMD in broader and more dynamic online network environments, particularly in scenarios involving complex or real-time zero-day attacks. Our long-term vision is to extend ConMD toward real-time, zero-positive anomaly detection, achieving efficient and practical implementation in real-world systems.

## REFERENCES

- [1] Y. Zhong, Z. Wang, X. Shi, J. Yang, and K. Li, "RFG-HELAD: A robust fine-grained network traffic anomaly detection model based on heterogeneous ensemble learning," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 5895–5910, 2024.
- [2] R. Xie et al., "Rosetta: Enabling robust TLS encrypted traffic classification in diverse network environments with TCP-aware traffic augmentation," in *Proc. ACM Turing Award Celebration Conf. China*, Jul. 2023, pp. 131–132.
- [3] Z. Cheng et al., "Kairos: Practical intrusion detection and investigation using whole-system provenance," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2024, pp. 3533–3551.
- [4] Y. Sharon, D. Berend, Y. Liu, A. Shabtai, and Y. Elovici, "TANTRA: Timing-based adversarial network traffic reshaping attack," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3225–3237, 2022.
- [5] M. Shen, J. Zhang, L. Zhu, K. Xu, and X. Du, "Accurate decentralized application identification via encrypted traffic analysis using graph neural networks," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 2367–2380, 2021.
- [6] X. Han, S. Liu, J. Liu, B. Jiang, Z. Lu, and B. Liu, "ECNet: Robust malicious network traffic detection with multi-view feature and confidence mechanism," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 6871–6885, 2024.
- [7] W. Li et al., "Prism: Real-time privacy protection against temporal network traffic analyzers," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 2524–2537, 2023.
- [8] E. Li, Z. Shang, O. Gungor, and T. Rosing, "SAFE: Self-supervised anomaly detection framework for intrusion detection," 2025, *arXiv:2502.07119*.
- [9] Y. A. Farrukh, S. Wali, I. Khan, and N. D. Bastian, "SeNet-I: An approach for detecting network intrusions through serialized network traffic images," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 107169.
- [10] W. Wang, M. Zhu, X. Zeng, X. Ye, and Y. Sheng, "Malware traffic classification using convolutional neural network for representation learning," in *Proc. Int. Conf. Inf. Netw. (ICOIN)*, Da Nang, Vietnam, Jan. 2017, pp. 712–717.
- [11] R. Zhao et al., "A novel self-supervised framework based on masked autoencoder for traffic classification," *IEEE/ACM Trans. Netw.*, vol. 32, no. 3, pp. 2012–2025, Jun. 2024.
- [12] K. Fauvel, F. Chen, and D. Rossi, "A lightweight, efficient and explainable-by-design convolutional neural network for internet traffic classification," in *Proc. 29th ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Aug. 2023, pp. 4013–4023.
- [13] X. Lian, C. Cao, Y. Liu, X. Xu, Y. Zheng, and F. Zhou, "Facing anomalies head-on: Network traffic anomaly detection via uncertainty-inspired inter-sample differences," in *Proc. ACM Web Conf.*, Apr. 2025, pp. 3908–3917.
- [14] X. Lian, Y. Liu, S. Wang, and Y. Zheng, "Payload level anomaly network traffic detection via semi-supervised contrastive learning," in *Proc. IEEE 23rd Int. Conf. Trust, Secur. Privacy Comput. Commun. (TrustCom)*, Dec. 2024, pp. 2559–2566.
- [15] Z. Zhao, Z. Li, Z. Song, W. Li, and F. Zhang, "Trident: A universal framework for fine-grained and class-incremental unknown traffic detection," in *Proc. ACM Web Conf.*, May 2024, pp. 1608–1619.
- [16] X. Lian, Y. Zheng, Z. Dang, C. Peng, and X. Gao, "Semi-supervised anomaly traffic detection via multi-frequency reconstruction," *Pattern Recognit.*, vol. 161, May 2025, Art. no. 111215.
- [17] W. T. Lunardi, M. A. Lopez, and J.-P. Giacalone, "ARCADE: Adversarially regularized convolutional autoencoder for network anomaly detection," *IEEE Trans. Netw. Service Manage.*, vol. 20, no. 2, pp. 1305–1318, Jun. 2023.
- [18] Y. Zheng, X. Lian, Z. Dang, C. Peng, C. Yang, and J. Ma, "A semi-supervised anomaly network traffic detection framework via multimodal traffic information fusion," in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2023, pp. 4455–4459.
- [19] C. DataCon. (2020). *Datacon: Open Dataset Datacon2020-Encrypted Malicious Traffic Dataset Direction Open Dataset*. [Online]. Available: <https://datacon.qianxin.com/opendata>
- [20] X. Yao, C. Zhang, R. Li, J. Sun, and Z. Liu, "One-for-all: Proposal masked cross-class anomaly detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 4792–4800.
- [21] Z. You et al., "A unified model for multi-class anomaly detection," in *Proc. NeurIPS*, 2022, pp. 4571–4584.
- [22] H. Deng and X. Li, "Anomaly detection via reverse distillation from one-class embedding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 9727–9736.
- [23] Y. Cai, D. Liang, D. Luo, X. He, X. Yang, and X. Bai, "A discrepancy aware framework for robust anomaly detection," *IEEE Trans. Ind. Informat.*, vol. 20, no. 3, pp. 3986–3995, Mar. 2024.
- [24] R. Zhao et al., "Yet another traffic classifier: A masked autoencoder based traffic transformer with multi-level flow representation," in *Proc. AAAI*, vol. 37, 2023, pp. 5420–5427.
- [25] X. Zhang, S. Li, X. Li, P. Huang, J. Shan, and T. Chen, "DeSTSeg: Segmentation guided denoising student-teacher for anomaly detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3914–3923.
- [26] Y. Qing et al., "Low-quality training data only? A robust framework for detecting encrypted malicious network traffic," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2024.
- [27] Z. Zhao, Z. Liu, H. Chen, F. Zhang, Z. Song, and Z. Li, "Effective DDoS mitigation via ML-driven in-network traffic shaping," *IEEE Trans. Depend. Secure Comput.*, vol. 21, no. 4, pp. 4271–4289, Jul. 2024.
- [28] Y. Liang, Y. Xie, S. Tang, S. Yu, X. Liu, and J. Hu, "Network traffic content identification based on time-scale signal modeling," *IEEE Trans. Depend. Secure Comput.*, vol. 20, pp. 2607–2624, 2022.
- [29] C. Yao, Y. Yang, K. Yin, and J. Yang, "Traffic anomaly detection in wireless sensor networks based on principal component analysis and deep convolution neural network," *IEEE Access*, vol. 10, pp. 103136–103149, 2022.
- [30] A. Parameswari, R. Ganeshan, V. Ragavi, and M. Shereesha, "Hybrid rat swarm hunter prey optimization trained deep learning for network intrusion detection using CNN features," *Comput. Secur.*, vol. 139, Apr. 2024, Art. no. 103656.
- [31] 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," *Comput. Netw.*, vol. 190, May 2021, Art. no. 107974.
- [32] G. AlMahadin et al., "VANET network traffic anomaly detection using GRU-based deep learning model," *IEEE Trans. Consum. Electron.*, vol. 70, no. 1, pp. 4548–4555, Feb. 2024.
- [33] R. Devendiran and A. V. Turukmane, "Dugat-LSTM: Deep learning based network intrusion detection system using chaotic optimization strategy," *Expert Syst. Appl.*, vol. 245, Jul. 2024, Art. no. 123027.
- [34] H. Nguyen and R. Kashef, "TS-IDS: Traffic-aware self-supervised learning for IoT network intrusion detection," *Knowl.-Based Syst.*, vol. 279, Nov. 2023, Art. no. 110966.
- [35] M. Zhong, M. Lin, and Z. He, "Dynamic multi-scale topological representation for enhancing network intrusion detection," *Comput. Secur.*, vol. 135, Dec. 2023, Art. no. 103516.
- [36] M. Zhong, M. Lin, C. Zhang, and Z. Xu, "A survey on graph neural networks for intrusion detection systems: Methods, trends and challenges," *Comput. Secur.*, vol. 141, Jun. 2024, Art. no. 103821.
- [37] Q. Zhou, L. Wang, H. Zhu, T. Lu, and V. S. Sheng, "WF-transformer: Learning temporal features for accurate anonymous traffic identification by using transformer networks," *IEEE Trans. Inf. Forensics Security*, vol. 19, pp. 30–43, 2024.
- [38] P. Dodia, M. AlSabah, O. Alrawi, and T. Wang, "Exposing the rat in the tunnel: Using traffic analysis for tor-based malware detection," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2022, pp. 875–889.
- [39] E. Caville, W. W. Lo, S. Layeghy, and M. Portmann, "Anomal-E: A self-supervised network intrusion detection system based on graph neural networks," *Knowl.-Based Syst.*, vol. 258, Dec. 2022, Art. no. 110030.

- [40] S. Akcay, A. Atapour-Abarghouei, and T. P. Breckon, "GANomaly: Semi-supervised anomaly detection via adversarial training," in *Proc. 14th Asian Conf. Comput. Vis.* Perth, WA, Australia: Springer, Dec. 2019, pp. 622–637.
- [41] L. Zhang, X. Xie, K. Xiao, W. Bai, K. Liu, and P. Dong, "MANomaly: Mutual adversarial networks for semi-supervised anomaly detection," *Inf. Sci.*, vol. 611, pp. 65–80, Sep. 2022.
- [42] Z. Wang, R. Zhou, S. Yang, D. He, and S. Chan, "A novel lightweight IoT intrusion detection model based on self-knowledge distillation," *IEEE Internet Things J.*, vol. 12, no. 11, pp. 16912–16930, Jun. 2025.
- [43] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Uninformed students: Student-teacher anomaly detection with discriminative latent embeddings," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4182–4191.
- [44] G. Wang, S. Han, E. Ding, and D. Huang, "Student-teacher feature pyramid matching for anomaly detection," in *Proc. Brit. Mach. Vis. Conf.*, 2021, p. 306.
- [45] M. Rudolph, T. Wehrbein, B. Rosenhahn, and B. Wandt, "Asymmetric student-teacher networks for industrial anomaly detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 2591–2601.
- [46] G. Tong, Q. Li, and Y. Song, "Two-stage reverse knowledge distillation incorporated and self-supervised masking strategy for industrial anomaly detection," *Knowl.-Based Syst.*, vol. 273, Aug. 2023, Art. no. 110611.
- [47] Z. Zhang, Z. Zhao, X. Zhang, C. Sun, and X. Chen, "Industrial anomaly detection with domain shift: A real-world dataset and masked multi-scale reconstruction," *Comput. Ind.*, vol. 151, Oct. 2023, Art. no. 103990.
- [48] Z. Cheng, Y. Liu, T. Zhong, K. Zhang, F. Zhou, and P. S. Yu, "Disentangling inter - and intra-cascades dynamics for information diffusion prediction," *IEEE Trans. Knowl. Data Eng.*, vol. 37, no. 8, pp. 4548–4563, Aug. 2025.
- [49] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *Proc. Brit. Mach. Vis. Conf.*, 2016, pp. 87.1–87.12.
- [50] A. Habibi Lashkari, G. Draper Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of tor traffic using time based features," in *Proc. 3rd Int. Conf. Inf. Syst. Secur. Privacy*, 2017, pp. 253–262.
- [51] A. Hatamizadeh, H. Yin, J. Kautz, P. Molchanov, and M. Pavlo, "Global context vision transformers," in *Proc. ICML*, 2022, pp. 12633–12646.
- [52] M. Tan and Q. V. Le, "EfficientNetv2: Smaller models and faster training," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10096–10106.
- [53] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
- [54] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy (ICISSP)*, 2018, pp. 108–116.
- [55] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly transformer: Time series anomaly detection with association discrepancy," in *Proc. ICLR*, 2022.
- [56] E. Eldele, M. Ragab, Z. Chen, M. Wu, and X. Li, "TSLANet: Rethinking transformers for time series representation learning," in *Proc. ICML*, 2024, pp. 12409–12428.
- [57] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proc. 8th IEEE Int. Conf. Data Min.*, Dec. 2008, pp. 413–422.
- [58] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural Comput.*, vol. 13, no. 7, pp. 1443–1471, Jul. 2001.
- [59] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related features," in *Proc. 2nd Int. Conf. Inf. Syst. Secur. Privacy*, 2016, pp. 407–414.
- [60] S. Dong, Y. Xia, and T. Peng, "Network abnormal traffic detection model based on semi-supervised deep reinforcement learning," *IEEE Trans. Netw. Service Manage.*, vol. 18, no. 4, pp. 4197–4212, Dec. 2021.