future internet

MDPI

Article

A Self-Supervised Learning Model for Unknown Internet Traffic Identification Based on Surge Period

Dawei Wei [1], Feifei Shi [1] and Sahraoui Dhelim [2,*]

[1] School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
[2] School of Computer Science, University College Dublin, Belfield, D04 V1W8 Dublin, Ireland
* Correspondence: sahraoui.dhelim@hotmail.com

Abstract: The identification of Internet protocols provides a significant basis for keeping Internet security and improving Internet Quality of Service (QoS). However, the overwhelming developments and updating of Internet technologies and protocols have led to large volumes of unknown Internet traffic, which threaten the safety of the network environment a lot. Since most of the unknown Internet traffic does not have any labels, it is difficult to adopt deep learning directly. Additionally, the feature accuracy and identification model also impact the identification accuracy a lot. In this paper, we propose a surge period-based feature extraction method that helps remove the negative influence of background traffic in network sessions and acquire as many traffic flow features as possible. In addition, we also establish an identification model of unknown Internet traffic based on JigClu, the self-supervised learning approach to training unlabeled datasets. It finally combines with the clustering method and realizes the further identification of unknown Internet traffic. The model has been demonstrated with an accuracy of no less than 74% in identifying unknown Internet traffic with the public dataset ISCXVPN2016 under different scenarios. The work provides a novel solution for unknown Internet traffic identification, which is the most difficult task in identifying Internet traffic. We believe it is a great leap in Internet traffic identification and is of great significance to maintaining the security of the network environment.

Keywords: unknown Internet traffic identification; self-supervised learning; surge period; clustering

check for updates

Citation: Wei, D.; Shi, F.; Dhelim, S. A Self-Supervised Learning Model for Unknown Internet Traffic Identification Based on Surge Period. *Future Internet* **2022**, *14*, 289. https://doi.org/10.3390/fi14100289

Academic Editor: Mario Di Mauro

Received: 23 September 2022
Accepted: 6 October 2022
Published: 10 October 2022

1. Introduction

With the overwhelming advances in mobile communications, coupled with the demanding requirements of intelligent applications, such as smart cities and industrial Internet of Things, an increasing number of smart mobile devices have been connected to the Internet [1]. They establish connections, transmit and exchange information with network operators through network access devices, and have led to large volumes of Internet traffic. Especially with the bad influence of COVID-19 in recent years, most businesses and cooperation have been transferred to online forms, meaning Internet traffic has increased at a faster rate.

At the same time, the boom in Internet technologies prompts the emergence of various Internet applications and services, which result in the diversification of unknown Internet protocols. In addition, the continuous updating of Internet protocols brought by application upgrade also generate many unknown protocols. Notably, unknown Internet protocols seriously threaten the security of the network environment and have become the focus of current Internet security research. It is extremely important to carry out unknown Internet traffic identification, make full use of Internet resources, and guarantee the security and stability of the Internet environment [2].

As far as we know, Internet protocols are different in terms of format, coding, length, etc., based on which it is possible to identify different protocols and further learn about the behaviors of users within a certain range. This is significant for ensuring users' privacy and maintaining the security of the Internet environment. Thus far, Internet traffic identification has been widely used in areas such as anomaly detection, attack denial, and resource allocation [3].

The common techniques of Internet traffic identification mainly include port, packet payload, user behavior, and flow-based measurements [4]. The port-based Internet traffic identification could be regarded as one of the most traditional methods. It mainly identifies the fixed port number used by different applications to distinguish the protocol types with a fast identification speed. However, this approach is already obsolete since most applications adopt dynamic ports nowadays [5]. The packet-based identification mainly refers to identifying Internet traffic depending on packet payload, namely Deep Packet Inspection; this approach requires protocol templates predefined by experts, which is almost impossible for unknown Internet traffic [6]. Moreover, the packet-based analysis puts forward strict requirements for computational abilities. The behavior-based Internet traffic identification concentrates more on analyzing users' behavior [7]. It is more applicable to scenarios that are closer to users and cannot handle complicated situations where the Internet traffic comes from different end users and applications.

Compared with the methods mentioned above, the flow-based one is another quite popular method for realizing Internet traffic identification, which depends on the general description of flow characteristics from different levels [8]. For example, it may concentrate on the packet-level characteristics of packet size distribution, packet arrival time interval distribution, etc., or the session-level features of the transmission of multiple sessions for a request, the number of session bytes, and the session duration. The flow-based identification method does not need to pay attention to the content of the data packets but achieves protocol classification and identification by extracting and analyzing flow features [9]. Therefore, it is more suitable for identifying encrypted Internet traffic and unknown ones.

Flow-based Internet traffic identification is usually divided into two key steps one is feature selection and extraction, and the other is model design and training. The flow features are usually extracted by methods of machine learning, such as statistical analysis and deep learning. For example, flow features generated by statistical analysis, or the so-called statistical features of Internet traffic, usually refer to the mean, variance, and sum [10]. Notably, there are hundreds of traffic flow statistical characteristics, based on which the algorithms need to select and filter the best feature combinations for further identification. That means the feature selection based on expert knowledge requires a lot of manual work or an extremely complex feature selection process, which largely limits the versatility of the algorithm. However, features obtained by deep learning could be automatically trained and extracted via algorithms such as CNN, deep auto-encoder, etc. Compared with statistical features, the deep learning features are closer to the properties of the traffic flow itself, and more importantly, it does not need excessive human intervention when extracting and selecting features. The feature extraction method based on deep learning has been widely used in dealing with the identification of unknown Internet traffic.

However, the feature extraction method with deep learning usually has strict requirements on the length and format of the input, while various network sessions may have different lengths [11]. Hence, it needs to sample and fill in the original traffic data first for a unified format. For example, if we adopt the first 1024 bytes of the payload as a set of data, the content that exceeds 1024 bytes needs to be deleted, and the insufficient part should be filled in with zeroes. In addition, most Internet traffic is background traffic, keepalive packets, etc., which have a negative impact on the extraction of traffic flow features [12,13]. Therefore, when adopting the method based on flow features to identify unknown Internet traffic, it is necessary to improve the accuracy of the features, as well as the training models, to realize the further identification of unknown Internet traffic.

To solve the problems mentioned above, we establish a self-supervised learning model for unknown Internet traffic identification based on the surge period. First, we propose a feature extraction method inspired by the idea of surge period and improve the accuracy of original features of unknown Internet traffic. Then, we establish a feature pre-training model based on self-supervised learning and generate the feature matrix for further traffic identification. Finally, we combine the pre-trained features with a clustering algorithm to realize the identification of unknown Internet traffic. The method is demonstrated with relatively high accuracy on the public dataset ISCXVPN2016. The main contributions of this paper are as follows:

- Propose an original feature extraction method based on the surge period to improve the accuracy of original features extracted from the raw Internet traffic data.
- Establish a JigClu-based feature pre-training model for unknown Internet traffic, which solves the challenge of no labels in unknown Internet traffic identification.
- Present a self-supervised learning identification model for unknown Internet traffic combined with JigClu clustering and prove its performance on the public dataset ISCXVPN2016.

The rest of this paper is arranged as follows. Section 2 introduces the background and related works, especially the popular methods of unknown Internet traffic identification. Section 3 presents the self-supervised learning model for unknown Internet traffic identification based on characteristics of the surge period. Results and discussion are presented in Section 4. Section 5 concludes the paper.

2. Literature Review

In general, the main difference between common Internet traffic identification and unknown ones is whether there are protocol labels to train models. Herein, the common Internet traffic refers to the known Internet traffic. For common Internet traffic with protocol labels, it is possible to extract features through statistical algorithms or deep learning methodsand further realize the classification and identification with labels. However, for unknown Internet traffic, it is hard to train models directly with deep learning as there are no labels at all.

Thus far, researchers have begun to identify unknown Internet traffic based on flow characteristics, such as the statistical features of the length of each packet, the number of packets, and packet size, or features learned via deep learning methods. These features could be trained with algorithms of clustering, semi-supervised learning, transfer learning, etc., to realize the final Internet traffic identification as expected [14]. Features hold important roles in achieving unknown Internet traffic classification and identification, on which we elaborate the related works and the possible algorithms adopted during the identification.

2.1. The Flow-Based Unknown Internet Traffic Identification with Statistical Features

The flow-based unknown Internet traffic identification with statistical features is a hot research topic [15]. As far as we know, the statistical features of Internet traffic mainly include the number of packets, the length of each packet, packet size, arrival time, etc. Based on the statistical features, researchers adopt various approaches, such as clustering, deep auto-encoder, etc., to achieve unknown Internet traffic identification.

The clustering algorithm is currently one of the most common methods for dealing with statistical flow characteristics and realizing the identification of unknown Internet traffic [16]. For example, Zhang adopts a semi-supervised clustering method to deal with unknown Internet traffic identification with statistical flow features [17]. It collects a small amount of labeled traffic, clusters the labeled and unlabeled data together, and further marks the clustering results with extended tags. This method finally adopts 20 session-level statistical features and realizes the identification of up to three types of unknown traffic, with the highest accuracy of about 65%.

In addition to semi-supervised clustering, deep clustering also holds an important role in training statistical characteristics and identifying Internet traffic. For example, Wang analyzes features of data length, control keywords, and address information to establish the eigenvector matrix of unknown traffic and adopts deep clustering to realize the identification [18]. Zhang extracts the bottleneck features from traffic flow and realizes unknown traffic identification by combining deep auto-encoder with constrained clustering [19].

In a word, the statistical features of traffic flow show advantages in helping identify unknown Internet traffic and could adapt to the changing requirements of user privacy protection, continuous updates of protocols, etc. However, it is usually difficult to filter and select efficient statistical features for model training and analysis. In the meantime, the processing of statistical features requires a lot of computing resources, which is somewhat tough to deploy in large networks. Hence, some studies try to use deep learning to automatically extract features for unknown Internet traffic identification.

2.2. The Flow-Based Unknown Internet Traffic Identification Based on Deep Learning Features

With the continuous maturity of deep learning algorithms, many people pay attention to automatically extracting Internet features by deep learning and further achieving accurate identification of large-scale Internet traffic [20]. With the help of deep learning, it is possible to calculate and find the hidden features in the original Internet traffic by back-propagation algorithms [21]. More importantly, it could effectively solve the problem of manual design and screening of statistical features in traditional algorithms. Thus far, deep learning has become the mainstream in identifying unknown Internet traffic.

The convolutional neural network (CNN) is one of the deep learning algorithms widely adopted in unknown Internet traffic classification. It can automatically extract features in the original Internet traffic through different convolution kernels and realize classification and identification. For example, Ma proposes a deep learning method based on CNN to realize unknown traffic identification [20]. It uses a payload of up to 1024 bytes as a data unit and maps it into a 32 × 32 feature matrix of unknown Internet traffic. The work adopts the first 10 kinds of protocols for known protocol types and the other 3 Internet protocols as the simulation data for unknown protocol types. Moreover, Yang also constructs a transfer learning approach based on a deep adaption network, which consists of multiple convolutional layers and a fully connected layer [22]. It trains the CNN model based on features sampled from the Internet traffic and then extends them to the labeled and unlabeled samples. The method is finally demonstrated with a relatively improved performance based on two public datasets of QUIC and Ariel. Wang [21] adopts an Artificial Neural Network (ANN) to extract unknown Internet traffic features and realizes the final identification with an accuracy of 67.16% with a probability condition of 0.8.

In addition, deep auto-encoder is also popular in dealing with unknown Internet traffic identification with no labels. For example, Zhao proposes an identification method of unknown Internet traffic based on embedding and deep auto-encoder [23]. He extracts the subsequence of packet payloads with n-grams as features, based on which the deep auto-encoder and deep clustering are adopted to realize unknown traffic identification. Hu [24] also focuses on unknown Internet traffic identification in an open-collection environment. In his experiment, he constructs a model based on CNN and a transformer encoder and selects five types of unknown Internet traffic manually, with a final identification precision of around 70%. Roselin also uses a deep auto-encoder to automatically extract the traffic features and realize the unknown Internet traffic identification [14]. The original traffic is converted into 20 × 20 data blocks as an input of the deep auto-encoder, based on which the unknown traffic identification is realized through the BIRCH clustering algorithm. However, the method also has shortcomings, especially in dealing with background traffic, which greatly impacts the recognition accuracy.

Compared with statistical features, the features trained and learned by deep learning depend more on the structures and properties of the traffic data itself. It could largely reduce the human labor involved in designing, selecting, and filtering the most appropriate statistical features automatically. Hence, deep learning methods will have more possibilities in terms of accuracy and efficiency when dealing with the identification of emerging unknown Internet traffic.

3. Methods

As we discussed above, flow-based Internet traffic identification exactly provides a method of feature extraction via machine learning and realizes further Internet traffic identification. When using machine learning to identify Internet traffic, the process could be divided into two key steps; one is feature selection and extraction, the other is the model design and training. Similarly, for unknown Internet traffic, the extraction of features, as well as the training of identification models, are the keys to ensuring the performance of the classification and identification.

However, there are some challenges that need to be overcome in unknown Internet traffic identification. In the aspect of feature extraction, although features learned by deep learning methods outperform the statistical ones a lot, the demanding requirements of neural network input, as well as the bad impacts of background traffic, largely decrease the accuracy of features extracted from original traffic. It is significant to remove the influence of background traffic and obtain many traffic features, to improve the accuracy of features in unknown Internet traffic. In the aspect of establishing the identification model, deep learning algorithms, especially CNNs, are widely used in identifying large volumes of Internet traffic. When training such models, a certain amount of labeled data is usually required. Additionally, the models would adopt loss functions to calculate the gap between predicted results and actual ones and use the optimizer to continuously optimize the model parameters. However, it seems difficult to directly adopt CNNs for unknown Internet traffic identification since most of them have no labels at all. In this case, a new method of training unlabeled datasets needs to be explored.

In this paper, we propose a novel self-supervised learning method for unknown Internet traffic identification. First, we design a feature extraction approach based on the idea of the surge period to improve the accuracy of features extracted from raw unknown Internet traffic. Then, we adopt JigClu, one of the self-supervised learning methods, to establish a pre-trained model to assist further identification with clustering algorithms, which would overcome the shortcomings of no labels in identifying unknown Internet traffic. The main process of the method is shown in Figure 1.

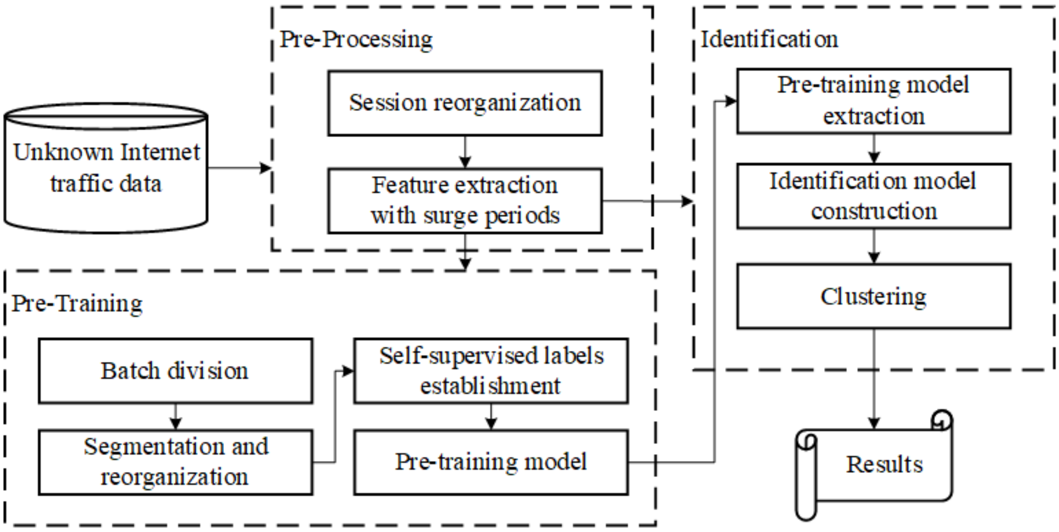

Pre-Processing
Session reorganization
Feature extraction with surge periods
Unknown Internet traffic data
Identification
Pre-training model extraction
Identification model construction
Clustering
Pre-Training
Batch division
Segmentation and reorganization
Self-supervised labels establishment
Pre-training model
Results

Figure 1. The main process of our proposed method of unknown Internet traffic identification.

In Figure 1, we divide the process of our proposed method into three stages, which are pre-processing, pre-training, and identification. When obtaining large volumes of unknown Internet traffic data, we first reorganize the traffic sessions at the pre-processing stage and extract original features based on surge periods to obtain as many valuable features as possible. Afterward, the extracted features enter into the pre-training stage, where the self-supervised labels are established with the help of JigClu, and the pre-trained feature matrix is generated. Finally, it would be combined with the clustering method to realize the ultimate identification of unknown Internet traffic. We will elaborate on each part of our innovative contributions in the following sections, especially the surge period-based feature extraction method and the self-supervised learning-based identification model.

3.1. The Surge Period-Based Feature Extraction Method in Unknown Internet Traffic Identification

In this paper, we propose a novel surge period-based method of feature extraction for unknown Internet traffic identification.

As we mentioned above, machine learning has strict requirements on the length and format of the input. Hence it is necessary to contain more Internet traffic features as input when identifying unknown Internet traffic. We all know that the communication of Internet protocols can be regarded as a combination of different network behaviors (such as user authentication, data request, and data upload). In other words, different network behaviors contain different characteristics of Internet protocols. Hence it is significant to obtain more network behaviors in the input of machine learning to acquire more Internet protocol features.

The feature extraction method we propose follows the idea of the surge period, which was initially proposed by Shi [25]. As the name implies, the surge period refers to the time interval of surging Internet traffic. It is defined by the surge time of Internet traffic density or bandwidth utilization and usually marks the traffic part where the network channel is busy uploading or downloading packets. Additionally, the time interval between any two adjacent data packets, except for the first one and the last one, should not be greater than the period of the predetermined time window size.

Figure 2 outlines the flow diagrams of the surge period-based feature extraction method, as well as the traditional one without surge periods.

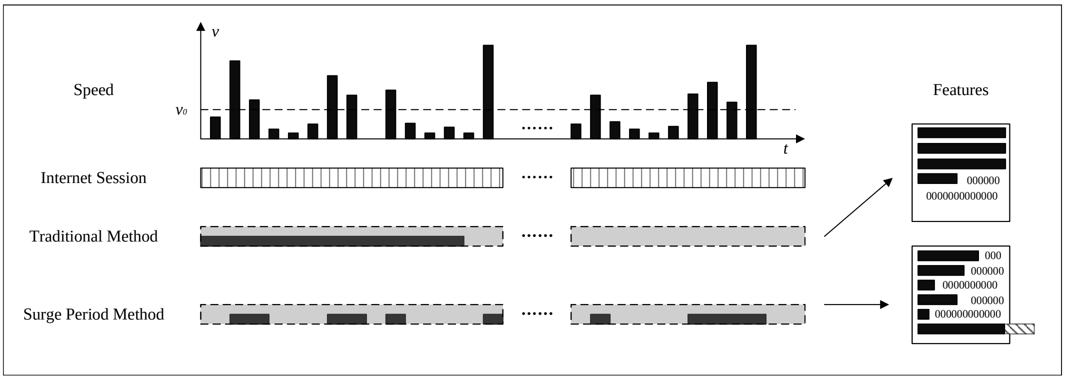


Figure 2. The flow diagram of the surge period-based feature extraction method and the traditional one without surge period considerations.

The surge period-based feature extraction method identifies the background traffic through the instantaneous speed of network sessions. Moreover, each surge period refers to the beginning of new network behavior, which allows the algorithm to further extract the flow characteristics. In this way, it could eliminate the impact of background traffic and meanwhile contain rich network behaviors, which help to extract a large number of network traffic features for further identification. For example, if a surge period contains n kinds of behaviors, the amount of feature information of each network behavior x is calculated as Equation (1):

$$info\{x\} = -\log P_x = \log \frac{1}{P_x} \tag{1}$$

Hence the total features $info\{X\}$ included in each surge period can be calculated using Equation (2):

$$\begin{aligned} info\{X\} &\approx -\log\left(\prod_{i\in[0,\ldots,n_0]} P(x_i)\right) \\ &= -\sum_{i\in[0,\ldots,n_0]} \log P(x_i) \\ &= \sum_{i\in[0,\ldots,n_0]} info\{x_i\} \end{aligned} \tag{2}$$

where n_0 is the number of network behaviors in a surge period.

Algorithm 1 introduces the surge period-based feature extraction method in detail. First, it reorganizes the Internet traffic according to a 5-tuple of network sessions and finds up-streams and down-streams of each network session. Meanwhile, it needs to detect the timestamp and size sequence of various data packets. Afterward, the method would calculate the speed of bit flow in each session with the period of t_0 and regard these with a speed greater than v_0 as a surge period. Following that, it extracts the packet length of the first n surge periods in chronological order and pads with zeros if there are not enough surge periods. Each surge period is characterized by the length of the first k packets, and similarly, pads with zeros if the data length of the surge period is insufficient. In the end, all data packets in all surge periods are traversed and extracted to generate the original features for subsequent pre-training and identification.

Algorithm 1 The surge period-based feature extraction method of unknown Internet traffic identification

```
Require: Internet traffic flow, the number of surge periods: $n$, the number of data packets:
    $k$, bit stream speed in network session: $v_0$
Ensure: Traffic features
Remove data packets of non TCP/UDP protocols
for Each data packet in TCP/UDP protocols do
    Read in the information of five tuples
    Reorganize the traffic flows according to five tuples
    Exchange the source port, destination port, source address and destination address
    in the tuple to find the up-stream or down-stream of the same session
end for
for Each network session do
    Sort data packets in its up-stream and down-stream in a chronological order
    for The beginning time $t$ of each network session to $t + t_0$ do
        Calculate the sum of packet length $sumlen$ for all data packets in the time unit
        if $sumlen > v_0$ then
            Mark $t + t_0$ as a surge period
            Extract $k$ packet length sequentially in the surge period as features
            if The number of already counted surge periods >$n$ then
                Break
            else
                Look for the next surge period
            end if
        end if
    end for
end for
Return features
```

Compared with other traditional feature extraction methods, the surge period-based one does not select a fixed network session as the feature extraction unit, but regards a dense period with a high flow rate as a surge period, and further analyzes and extracts the flow characteristics of surge periods. This method shows great advantages in efficiently extracting features of unknown Internet traffic. On the one hand, it avoids the influence of noise, such as background traffic and improves the accuracy of features. On the other hand, it also greatly expands the number of flow features that can be extracted, laying a solid foundation for the efficient identification of unknown traffic.

3.2. The Self-Supervised Learning-Based Identification Model for Unknown Internet Traffic

To deal with the large volumes of unlabeled data, scholars and researchers are devoted to exploring novel methods. Self-supervised learning is such an emerging approach to pre-training models with large amounts of unlabeled data. It aims to design proxy tasks to mine the characteristics of the data itself and adopts them as supervisory information for unlabeled data to improve the performance of feature extraction. In this way, the self-supervised learning method could help effectively improve the training speed and enhance the accuracy and stability of models to a great extent. Thus far, self-supervised learning has been widely adopted in areas of speech recognition, computer vision, etc. The unknown Internet traffic identification also encounters the challenge of no labels, and self-supervised learning exactly provides an effective solution for realizing the identification.

In this paper, we establish a self-supervised learning-based identification model for unknown Internet traffic. As can be seen in Figure 3, the model is composed of the JigClu-based feature pre-training model and the clustering-based identification method.

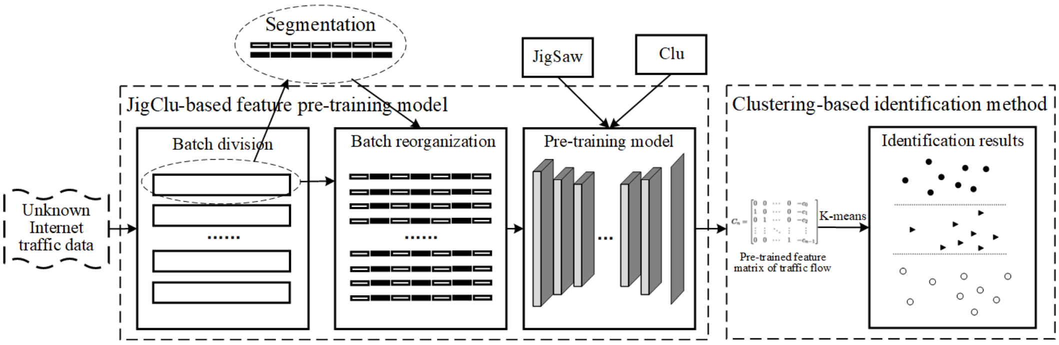

Segmentation
JigSaw
Clu
JigClu-based feature pre-training model
Batch division
Batch reorganization
Pre-training model
Unknown Internet traffic data
......
......
...
Clustering-based identification method
Identification results
$C_n = \begin{bmatrix} 0 & 0 & \cdots & 0 & -c_0 \\ 1 & 0 & \cdots & 0 & -c_1 \\ 0 & 1 & \cdots & 0 & -c_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & -c_{n-1} \end{bmatrix}$
K-means
Pre-trained feature matrix of traffic flow

Figure 3. The self-supervised learning-based identification model for unknown Internet traffic.

3.2.1. The JigClu-Based Flow Feature Pre-Training Model

As far as we know, JigClu is an efficient self-supervised learning approach. It aims to enable the model to restore a batch of scrambled data to the original state through self-supervised learning training. The main procedures are outlined as follows:

- Batch division and reorganization: This is the prerequisite step before further procedures. It allows the division of traffic data into batches with the same length of b. For the data in each batch, the algorithm will segment and reorganize the fragments from the same batch to form a "new" batch of data.
  In our proposed model, we divide the input data with a feature size of (n, k) into $b \times 4$ data blocks with the same size. The fragments in the same batch are recombined to form input with a size of (b, n, k). This input is new data with the same size as the original data while the content is disordered. At the same time, we record the scrambled fragments as labels to calculate the training loss functions.
- Pre-training model: This step mainly focuses on establishing and optimizing the pre-training model based on the "new" batch of data. The pre-training model adopts a deep residual network (ResNet) to extract features, which is composed of multiple residual units. In each residual unit, there are multiple convolutional layers and shortcut connections.
  As shown in Figure 4, we assume that the input of the residual unit is x, and its expected output is $H(x)$, which is a complex potential map. If we want to learn such a complicated model, the training would be more difficult. With the help of shortcut

connections, ResNet's learning goal could be calculated as $F(x) := H(x) - x$, and it helps solve the gradient disappearance problem of deep neural networks. The ResNet will finally generate two kinds of output, used to determine the original data block to which the included data belongs and the location to which the data belongs. The output is transformed into $(b \times 2 \times 2)$ vectors, represented as $L = L_1, L_2, \ldots L_{4b}$, which will be further used for data positioning and reorganization operations.

- JigSaw: For the data positioning module, its goal is to find the original location of multiple data blocks in the data. For example, if the input data are spliced into four data blocks, the first block of the spliced data is from the first small block of some original data, then the tag corresponding to this block could be marked as 0. Similarly, the second one is marked with the tag 1, and so on. Here, we regard the process of data positioning as a problem of multi-classification and adopt the function of CrossEntropy to calculate the loss of JigSaw. The details can be seen in Equation (3).

$$loss_{Jigsaw} = -\sum_{i=1}^{M} y_i \log(p_i) \tag{3}$$

where p_i refers to the predicated distribution and y_i is the actual distribution. M represents the number of types.

- Clu: The purpose of the data reorganization module is to find as many data blocks from the same original data. For the input of each data batch, it is necessary to determine which small data block comes from the same source to shorten the distance between multiple data blocks from the same original data and to lengthen the distance between different original data. Here, we regard it as a clustering problem, and the distance is calculated as Equations (4) and (5), where C_i is a collection of data from the same cluster.

$$loss_{i,j} = -\log \frac{\exp(\cos(l_i, l_j)/\tau)}{\sum_{k=1}^{4b} L_{k \neq i} \exp(\cos(l_i, l_k)/\tau)} \tag{4}$$

$$loss_{clu} = \frac{1}{4b} \sum_{i} \left(\frac{1}{3} \sum_{j \in C_i} loss_{i,j} \right) \tag{5}$$

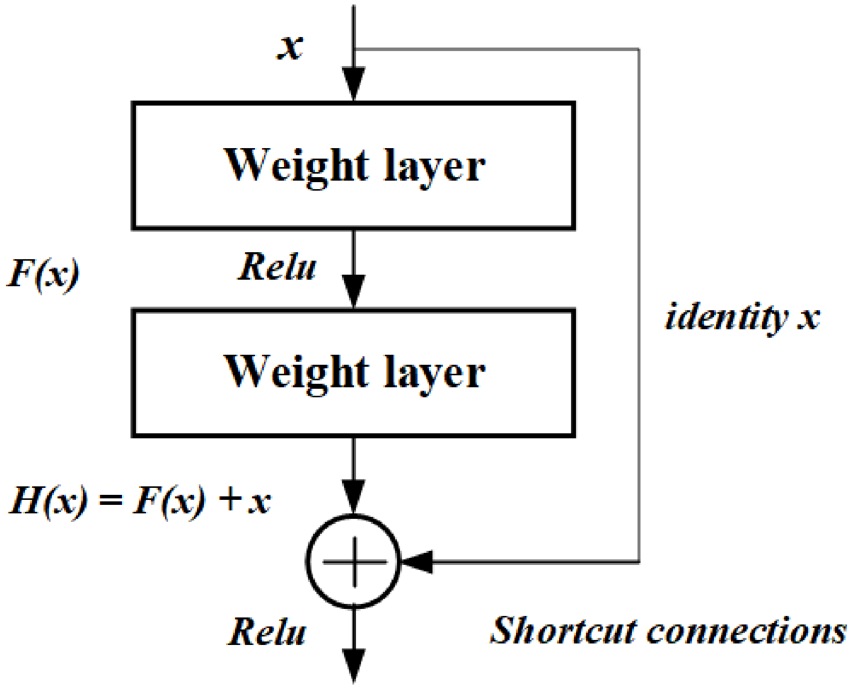


Figure 4. The pre-training model in the algorithm of JigClu.

By adopting the self-supervised learning algorithm of JigClu, the original features acquired at the pre-processing stage would be pre-trained and learned. It further generates the pre-trained feature matrix that can be used for further identification.

To be noted, the JigSaw and Clu steps in the JigClu method would not appear in the final identification model. They are only adopted here by calculating loss functions and, in turn, help optimize the model parameters. Before the final procedure of identification, we need to remove the modules of JigSaw and Clu, keep the remaining part in JigClu as the feature pre-training model, and further combine it with the clustering algorithm to realize the identification.

3.2.2. The Clustering-Based Identification Method

After we have established the JigClu-based pre-training model and generated the pre-trained feature matrix of Internet traffic, we use the traditional clustering method to realize the final identification of Internet traffic. By loading the feature pre-training model and connecting an unsupervised clustering model after the output, Internet traffic from the same type could be identified to the same cluster.

To the best of our knowledge, the clustering method is to divide datasets into different classes or clusters according to selected features. Its purpose is to make data in the same cluster as similar as possible and data in different clusters as different as possible.

In our method, we adopt K-means to realize further traffic identification. Among all clustering methods, K-means is an efficient approach that has been widely used in various classification and identification tasks. It calculates the distance between data samples and each cluster center and assigns them to the nearest one. Afterward, the algorithm needs to recalculate the center of each cluster and redistribute each data sample until the termination condition is satisfied.

Herein, since the features obtained by the pre-training model are continuous, we choose the Euclidean distance to calculate the similarity between data samples, which is shown as in Equation (6).

$$dist(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} = \|x - y\|_2 \tag{6}$$

where x and y are pieces of data in the dataset, and x_i and y_i are features pre-trained via the JigClu method.

The purpose of K-means is to find the best solutions that can minimize the square error of the datasets, which is:

$$E = \sum_{s=1}^{k} \sum_{x \in C_s} (x - \overline{X_s})^2 \tag{7}$$

where $\overline{X_s}$ is the average vector of the cluster of C_s. In other words, it is the center of cluster C_s, and is defined as in Equation (8). $|C_s|$ refers to the number of data samples in the cluster.

$$\overline{X_s} = \frac{1}{|C_s|} \sum_{x \in C_s} x \tag{8}$$

4. Results and Discussions

4.1. Dataset

In this paper, we adopt the publicly available ISCX VPN-nonVPN dataset (ISCXVPN2016) to demonstrate the performance of our proposed method [26]. The dataset was established by the Canadian Institute for Cybersecurity by defining sets of tasks. It is one of the most used datasets in the field of Internet traffic identification and has been widely applied in VPN traffic identification, encrypted protocol classification, abnormal traffic detection, etc. This paper adopts it as the dataset for unknown Internet traffic identification to evaluate the performance of our proposed model. Moreover, the use of public datasets also allows other researchers to make comparisons and improvements based on our experiments to prompt progress in the area of unknown Internet traffic identification. The dataset contains 21 different Internet protocols of 7 types, which are web, email, chat, streaming, file transfer, VoIP, and P2P. In our experiment, we divide the dataset into three scenarios that are different in the types and quantities of Internet protocols. Table 1 outlines the details of the dataset.

Table 1. The detailed information of the ISCXVPN2016.

Class	Totalconv	Totalsurge	TotalP
aim_chat	419	459	4766
email	7518	4417	45,757
facebook_audio	79,955	80,924	1,858,229
facebook_chat	507	1851	10,882
facebook_video	431	11,603	739,891
ftps	849	12,716	7,872,864
gmailchat	450	2134	12,240
hangouts_audio	79,444	80,935	1,858,075
hangouts_chat	436	1771	13,202
hangouts_video	1533	19,612	1,990,642
icq_chat	437	624	4119
netflix	380	2089	299,263
sftp	11,690	5543	1,700,514
skype_audio	38,820	32,839	882,213
skype_chat	8828	4253	70,183
skype_file	57,317	12,401	1,365,641
skype_video	599	16,261	1,520,980
spotify	301	918	41,180
vimeo	452	7299	146,375
voipbuster	2963	36,481	842,535
youtube	925	1330	252,071
Total	294,254	336,460	21,531,622

To demonstrate the robustness and effectiveness of our proposed model, we adopt three scenarios to make the verification. The types of Internet traffic included in each scenario are as follows:

- Scenario A: facebook_audio, aim_chat, facebook_video, facebook_chat, email.
- Scenario B: hangouts_chat, netflix, gmail_chat, hangouts_video, ftps, hangouts_audio, icq_chat.
- Scenario C: skype_chat, sftp, scp, skype_file, skype_video, youtube, vimeo, skype_audio, spotify, voipbuster.

4.2. Experiment Performance

We introduce the parameters of our workspace as follows: Intel® Xeon® Silver 4110 CPU @ 2.10 GHz, 64.0 GB RAM @ 2666 MHz, 480 GB SSD, and NVIDIA TITAN Xp.

In our established model, the inputs are TCP/UDP sessions of different lengths. The data packets in each session are shaped in chronological order, and the surge period-based features are extracted to generate a visual gray image with a size of 20 × 20. Partial feature images of network sessions are shown in Figure 5.

After obtaining the original features of unknown Internet traffic based on surge periods, we train them with the JigClu-based feature pre-training model for 100 epochs and adopt NADAM as the optimizer of the pre-training model. We conduct multiple rounds of tests for parameter optimization with the method of the grid search, and the optimal parameter settings are shown in Table 2.

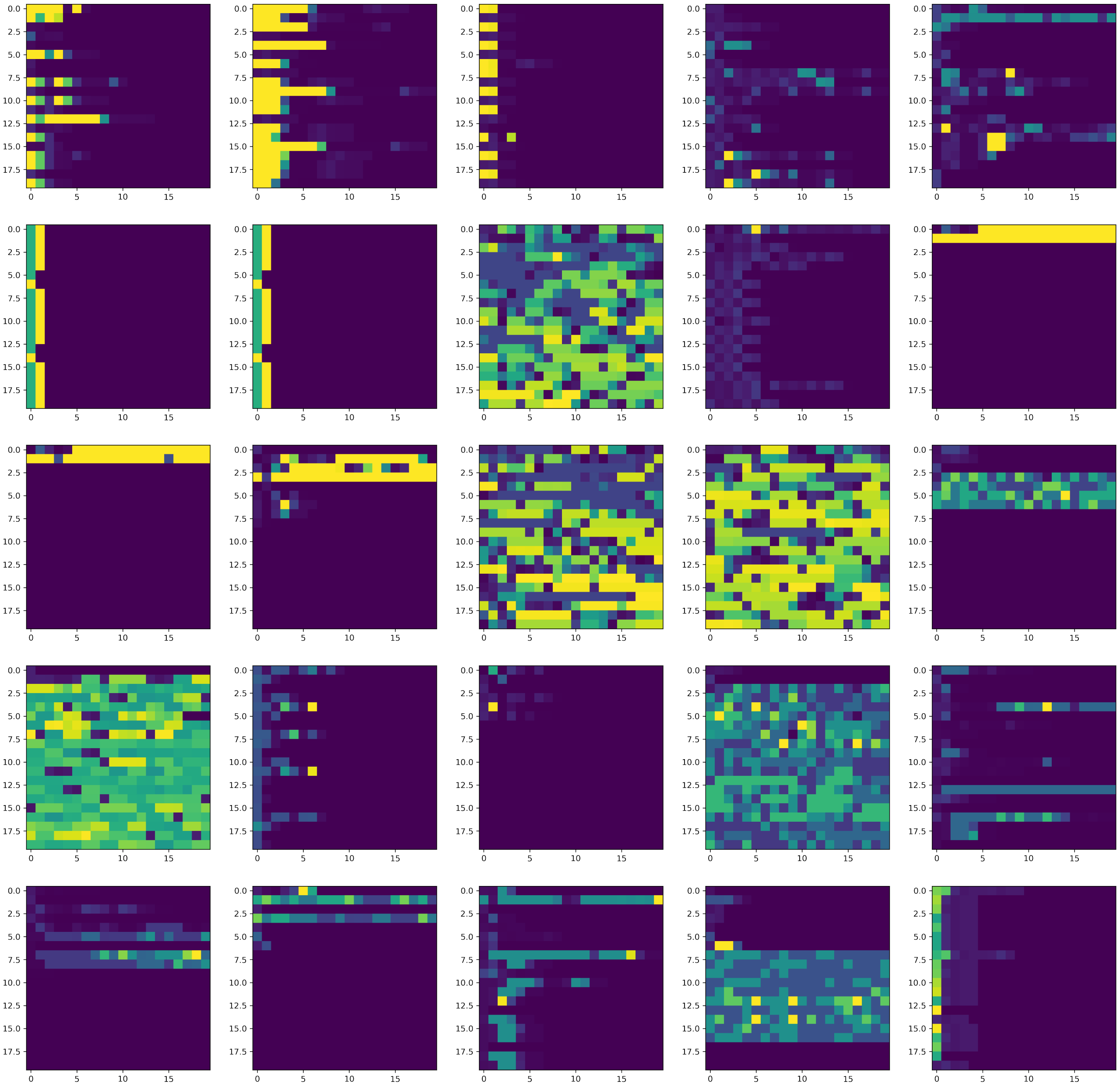


Figure 5. Partial images of surge periods-based original features extracted from network sessions.

Table 2. The optimal parameter settings in our experiment.

Parameter	Value	Meaning
temperature	0.7	Gain coefficient of superCluLoss
batch_size	128	Number of samples in each batch
epochs	200	Training rounds
learning_rate	0.0001	The learning rate
inputshape	(20,20)	Size of features extracted from the session

Figure 6 presents the changes in the loss function when training the JigClu-based feature pre-training model. In the training process, the loss function gradually reduces until the convergence speed slows down after 20 cycles and tends to converge after 50 cycles. It demonstrates the model with a relatively good training speed.

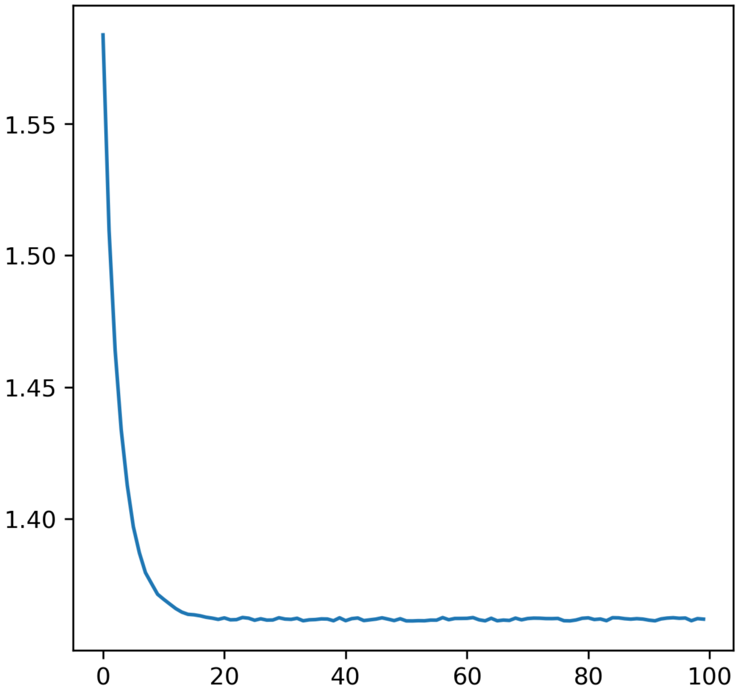


Figure 6. The loss function of the JigClu-based feature pre-training model.

After the pre-training model has been established, we remove the data positioning and reorganization modules and connect the features learned from unknown Internet traffic with the following clustering methods. Here, we conduct the final classification with K-means by setting different k clusters under the three different scenarios.

In order to evaluate the performance of our proposed method in identifying unknown Internet traffic, we adopt a confusion matrix to calculate the common indicators of precision (P), recall (R), and F_1 score.

In our experiment, we introduce the confusion matrix of true positive (TP), true negative (TN), false positive (FP), and false negative (FN). TP refers to assigning two identical protocol traffic data to the same cluster, while TN represents assigning two different protocols traffic data to different clusters. Additionally, FP means wrongly identifying traffic data of different protocols as the same cluster, and FN is to assign traffic data in identical protocols to different clusters. Accordingly, the P, R, and F_1 score, are calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F_1 = \frac{2PR}{P + R} \tag{11}$$

As shown in Table 3, the P of the proposed self-supervised learning-based identification model could reach above 74% when identifying unknown Internet traffic, with the best performance of 86% under scenario A. Although the highest precision does not exceed 90%, it is still an outstanding achievement for the identification of unknown Internet protocols without labels. The data in bold in the table represent the best performance results under the listed conditions.The performance results demonstrate the proposed model could extract the flow-based characteristics and further realize the identification of unknown Internet traffic. We can also find that with the increase in the number of clusters, the performance indicators such as P, R, and F_1 score are generally on the rise, while the rising speed is gradually slowing down. It shows that when the k_clusters reach a certain value, the evaluation indicators of our model tend to converge.

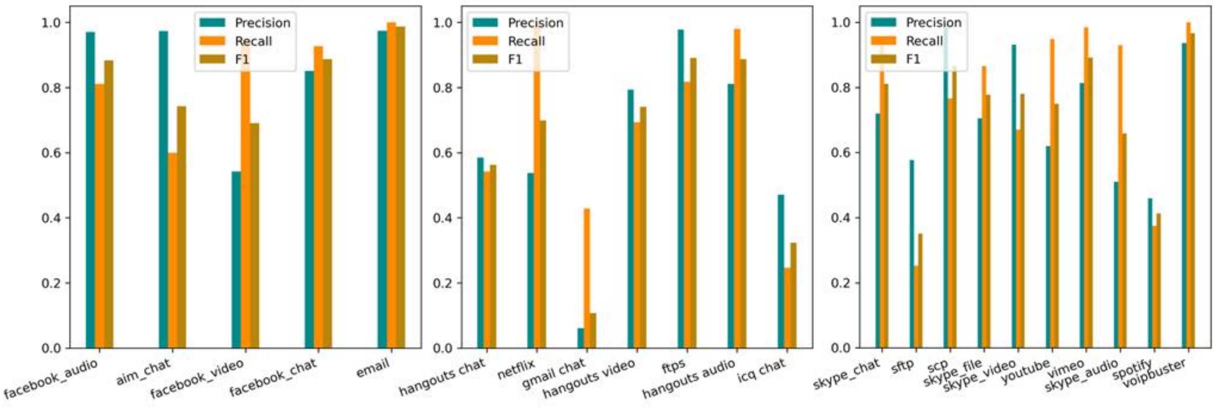
When comparing the performance among different scenarios, it can be seen that the identification performance of our model would slightly decrease with the increasing types of unknown Internet protocols, but the overall performance is still good.

Table 3. The performance of the proposed self-supervised learning-based identification model for unknown Internet traffic.

Scenario	k_clusters	F_1_score	P	R
A	10	0.634750567	0.707801418	0.707801418
A	20	0.759527691	0.782370821	0.782370821
A	30	0.774174454	0.797163121	0.797163121
A	40	0.791222723	0.81337386	0.81337386
A	50	0.796625203	0.818439716	0.818439716
A	60	0.807701369	0.826545086	0.826545086
A	70	0.82940477	0.840932118	0.840932118
A	80	0.856617997	0.860182371	0.860182371
A	90	**0.856617997**	0.860182371	0.860182371
A	100	0.856576486	**0.860992908**	**0.860992908**
B	10	0.631232209	0.702567865	0.702567865
B	20	0.656807596	0.719148936	0.719148936
B	30	0.663706266	0.724578136	0.724578136
B	40	0.677842173	0.731327953	0.731327953
B	50	0.698839937	0.739985326	0.739985326
B	60	0.713056204	0.7540719	0.7540719
B	70	0.723907056	0.760674982	0.760674982
B	80	0.73798451	0.769332355	0.769332355
B	90	0.751859908	0.777696258	0.777696258
B	100	**0.769164772**	**0.786353632**	**0.786353632**
C	10	0.361244359	0.437019101	0.437019101
C	20	0.531886447	0.543395211	0.543395211
C	30	0.578844264	0.581113801	0.581113801
C	40	0.591782961	0.594834544	0.594834544
C	50	0.611349266	0.61393597	0.61393597
C	60	0.668733435	0.664191552	0.664191552
C	70	0.681865684	0.679096045	0.679096045
C	80	0.710997885	0.71401668	0.71401668
C	90	0.697500494	0.708528383	0.708528383
C	100	**0.741288262**	**0.743287598**	**0.743287598**

To further illustrate the effectiveness of our proposed method, we have carried out a detailed analysis of the identification performance of each Internet protocol under different scenarios. Figure 7 shows the results. To be noted, all results are obtained under the optimal configuration state. As can be seen in Figure 7, the model shows different identification performances for different Internet protocols. It has a relatively good identification performance for most protocols, and some have reached a precision of over 90%. However, the performance is not so satisfactory for protocols such as the sftp in scenario C.

We further analyze the identification results of sftp in scenario C. The final identification results are shown as follows: sftp (58%), scp (38%), and skype_file (4%). We find that nearly 40% of sftp protocols are regarded mistakenly as scp protocols. This is mainly because both sftp and scp are encrypted file transfer protocols. Compared with the scp protocol, the sftp protocol has less data, which also affects the identification performance. It implies that our model could effectively distinguish different types of unknown network protocols while the performance in identifying different Internet protocols of the same type needs to be strengthened.



(a) Scenario A **(b) Scenario B** **(c) Scenario C**

Figure 7. The identification performance of each Internet protocol in different scenarios.

5. Conclusions

In this paper, we propose a novel method of unknown Internet traffic identification based on self-supervised learning. We first present a fresh feature extraction method inspired by the idea of the surge period, which largely improves the accuracy of original features extracted from the Internet traffic data. Additionally, to solve the problem of unknown traffic without labels, we establish a JigClu-based feature pre-training model that helps establish self-supervised labels and generates a pre-trained feature matrix for further identification. It finally combines with K-means and realizes the unknown Internet traffic identification. The method has been demonstrated with the public dataset ISCXVPN2,and has shown an accuracy of no less than 74%.

This work initially adopts self-supervised learning to achieve unknown Internet traffic identification. It gives insights into identifying unlabeled unknown Internet data with self-supervised learning and is regarded as a meaningful exploration. More work in-depth is needed to be discussed in the future.

Author Contributions: Conceptualization, D.W. and S.D.; methodology, D.W.; validation, D.W. and F.S.; writing—original draft preparation, D.W. and F.S.; writing—review and editing, D.W. and S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to express their appreciation to the Canadian Institute for Cybersecurity for their publicly available dataset ISCXVPN2016.

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

References

1. Dhelim, S.; Aung, N.; Kechadi, T.; Ning, H.; Chen, L.; Lakas, A. Trust2Vec: Large-Scale IoT Trust Management System based on Signed Network Embeddings. *IEEE Internet Things J.* **2022**. [CrossRef]
2. Azamuddin, W.M.H.; Hassan, R.; Aman, A.H.M.; Hasan, M.K.; Al-Khaleefa, A.S. Quality of service (Qos) management for local area network (LAN) using traffic policy technique to secure congestion. *Computers* **2020**, *9*, 39. [CrossRef]
3. Nguyen, T.T.; Armitage, G. A survey of techniques for internet traffic classification using machine learning. *IEEE Commun. Surv. Tutor.* **2008**, *10*, 56–76. [CrossRef]
4. Callado, A.; Kamienski, C.; Szabó, G.; Gero, B.P.; Kelner, J.; Fernandes, S.; Sadok, D. A survey on internet traffic identification. *IEEE Commun. Surv. Tutor.* **2009**, *11*, 37–52. [CrossRef]
5. Bakhshi, T.; Ghita, B. On Internet Traffic Classification: A Two-Phased Machine Learning Approach. *J. Comput. Netw. Commun.* **2016**, *2016*, 21. [CrossRef]

6. Bujlow, T.; Carela-Espanol, V.; Barlet-Ros, P. Independent comparison of popular DPI tools for traffic classification. *Comput. Netw.* **2015**, *76*, 75–89. [CrossRef]
7. Zeng, X.; Chen, X.; Shao, G.; He, T.; Han, Z.; Wen, Y.; Wang, Q. Flow context and host behavior based shadowsocks's traffic identification. *IEEE Access* **2019**, *7*, 41017–41032. [CrossRef]
8. Mohd, A.; Nor, D. Towards a Flow-based Internet Traffic Classification for Bandwidth Optimization. *Int. J. Comput. Sci. Secur.* **2009**, *3*, 146–153.
9. Soysal, M.; Schmidt, E.G. Machine learning algorithms for accurate flow-based network traffic classification: Evaluation and comparison. *Perform. Eval.* **2010**, *67*, 451–467. [CrossRef]
10. Lashkari, A.H.; Draper-Gil, G.; Mamun, M.S.I.; Ghorbani, A.A. Characterization of tor traffic using time based features. In Proceedings of the ICISSp, Porto, Portugal, 19–21 February 2017; pp. 253–262.
11. Salman, O.; Elhajj, I.H.; Kayssi, A.; Chehab, A. Data representation for CNN based internet traffic classification: A comparative study. *Multimed. Tools Appl.* **2021**, *80*, 16951–16977. [CrossRef]
12. Sirinam, P.; Imani, M.; Juarez, M.; Wright, M. Deep fingerprinting: Undermining website fingerprinting defenses with deep learning. In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, Toronto, ON, Canada, 15–19 October 2018; pp. 1928–1943.
13. Soos, G.; Ficzere, D.; Varga, P. Towards traffic identification and modeling for 5g application use-cases. *Electronics* **2020**, *9*, 640. [CrossRef]
14. Roselin, A.G.; Nanda, P.; Nepal, S.; He, X. Intelligent anomaly detection for large network traffic with Optimized Deep Clustering (ODC) algorithm. *IEEE Access* **2021**, *9*, 47243–47251. [CrossRef]
15. Peng, L.; Yang, B.; Chen, Y.; Chen, Z. Effectiveness of statistical features for early stage internet traffic identification. *Int. J. Parallel Program.* **2016**, *44*, 181–197. [CrossRef]
16. Erman, J.; Arlitt, M.; Mahanti, A. Traffic classification using clustering algorithms. In Proceedings of the 2006 SIGCOMM Workshop on Mining Network Data, Pisa, Italy, 15 September 2006; pp. 281–286.
17. Zhang, J.; Chen, C.; Xiang, Y.; Zhou, W.; Vasilakos, A.V. An effective network traffic classification method with unknown flow detection. *IEEE Trans. Netw. Serv. Manag.* **2013**, *10*, 133–147. [CrossRef]
18. Wang, W.; Bai, B.; Wang, Y.; Hei, X.; Zhang, L. Bitstream protocol classification mechanism based on feature extraction. In Proceedings of the 2019 International Conference on Networking and Network Applications (NaNA), Daegu, Korea, 10–13 October 2019; pp. 241–246.
19. Zhang, Y.; Zhao, S.; Sang, Y. Towards unknown traffic identification using deep auto-encoder and constrained clustering. In Proceedings of the International Conference on Computational Science, Faro, Portugal, 12–14 June 2019; pp. 309–322.
20. Ma, R.; Qin, S. Identification of unknown protocol traffic based on deep learning. In Proceedings of the 2017 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu , China, 12–16 December 2017; pp. 1195–1198.
21. Wang, Z. The applications of deep learning on traffic identification. *BlackHat USA* **2015**, *24*, 1–10.
22. Yang, Z.; Lin, W. Unknown traffic identification based on deep adaptation networks. In Proceedings of the 2020 IEEE 45th LCN Symposium on Emerging Topics in Networking (LCN Symposium), Sydney, Australia, 16–19 November 2020; pp. 10–18.
23. Zhao, S.; Zhang, Y.; Sang, Y. Towards unknown traffic identification via embeddings and deep autoencoders. In Proceedings of the 2019 26th International Conference on Telecommunications (ICT), Hanoi, Vietnam, 8–10 April 2019; pp. 85–89.
24. Hu, X.; Gu, C.; Chen, Y.; Chen, X.; Wei, F. OpenCBD: A Network-Encrypted Unknown Traffic Identification Scheme Based on Open-Set Recognition. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 1746373. [CrossRef]
25. Shi, Y.; Biswas, S. Website fingerprinting using traffic analysis of dynamic webpages. In Proceedings of the 2014 IEEE Global Communications Conference, Phoenix, AZ, USA, 15–17 April 2014; pp. 557–563. [CrossRef]
26. Draper-Gil, G.; Lashkari, A.H.; Mamun, M.S.I.; Ghorbani, A.A. Characterization of encrypted and vpn traffic using time-related. In Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP), Rome, Italy, 19–21 February 2016; pp. 407–414.