新增未知攻击场景下的工业互联网恶意流量识别方法

曾凡一，苘大鹏，许晨，韩帅，王焕然，周雪，李欣纯，杨武
（哈尔滨工程大学计算机科学与技术学院，黑龙江 哈尔滨 150009）

摘　要：针对工业互联网中新增未知攻击所引发的流量数据分布偏移问题，提出了一种基于邻域过滤和稳定学习的恶意流量识别方法，旨在增强现有图神经网络模型在识别已知类恶意流量时的有效性和鲁棒性。该方法首先对流量数据进行图结构建模，捕获通信行为中的拓扑关系与交互模式；然后，基于有偏采样的邻域过滤机制划分流量子图，消除通信行为间的伪同质性；最后，应用图表示学习和稳定学习策略，结合自适应样本加权与协同损失优化方法，实现高维流量特征的统计独立性。2个基准数据集上的实验结果表明，相较对比方法，所提方法在新增未知攻击场景下的识别性能提升超过2.7%，展示了其在工业互联网环境下的高效性和实用性。

关键词：工业互联网；恶意流量识别；图神经网络；邻域过滤；稳定学习
中图分类号：TP393
文献标志码：A
DOI：10.11959/j.issn.1000–436x.2024093

Identification method for malicious traffic in industrial Internet under new unknown attack scenarios

ZENG Fanyi, MAN Dapeng, XU Chen, HAN Shuai, WANG Huanran, ZHOU Xue, LI Xinchun, YANG Wu
College of Computer Science and Technology, Harbin Engineering University, Harbin 150009, China

Abstract: Aiming at the problem of traffic data distribution shift caused by new unknown attacks in the industrial Internet, a malicious traffic identification method based on neighborhood filtering and stable learning was proposed to enhance the effectiveness and robustness of the existing graph neural network model in identifying known malicious traffic. Firstly, the graph structure of the traffic data was modeled to capture the topological relationship and interaction mode in communication behavior. Secondly, the traffic subgraph was divided based on the neighborhood filtering mechanism of biased sampling to eliminate the pseudo-homogeneity between communication behaviors. Finally, the statistical independence of high-dimensional traffic features was realized by applying graph representation learning and stable learning strategies, combined with adaptive sample weighting and collaborative loss optimization methods. The experimental results on two benchmark datasets show that compared with the baseline method, the recognition performance of the proposed method is increased by more than 2.7% in the new unknown attack scenario, which shows its high efficiency and practicability in the industrial Internet environment.

Keywords: industrial Internet, malicious traffic identification, graph neural network, neighborhood filtering, stable learning

收稿日期：2023–11–30；**修回日期：**2024–04–08
通信作者：苘大鹏，mandapeng@hrbeu.edu.cn
基金项目：国家重点研发计划基金资助项目（No.2021YFB3101403）；国家自然科学基金资助项目（No.U2003206, No.U20B2048, No.U21B2019, No.U22A2036, No.62272127）；黑龙江省自然科学基金资助项目（No.TD2022F001）
Foundation Items: The National Key Research and Development Program of China (No.2021YFB3101403), The National Natural Science Foundation of China (No.U2003206, No.U20B2048, No.U21B2019, No.U22A2036, No.62272127), The Natural Science Foundation of Heilongjiang Province (No.TD2022F001)

0 引言

随着工业4.0的快速发展[1]，工业互联网已成为企业和生产制造业的核心支柱[2]，同时也成为网络攻击者的新焦点。恶意网络行为，包括各种网络攻击和非法访问，可能导致资料泄露、设备损坏甚至生产中断，对工业互联网安全构成了严重威胁[3]。因此，恶意流量识别已成为保障工业互联网安全的重要手段。

以往，恶意流量检测与识别主要依赖基于签名或规则的方法[4]。然而，由于网络攻击行为的多样性，以及网络规模的不断增大，手动设计和维护这些签名或规则变得非常困难，此类方法已无法满足实际需要[5]。与传统的基于特征工程的恶意流量识别方法不同，基于深度学习的识别模型能够从大量流量数据中学习网络攻击的高级特征，更好地适应日益复杂的网络威胁行为[6]。

近年来，图神经网络（GNN, graph neural network）因其优秀的结构数据表示学习能力，在恶意流量识别任务中得到持续探索[6-8]。研究人员尝试利用图表示学习模型捕获网络流量中更为复杂的拓扑结构与交互模式，并取得了一定突破。Zhou等[9]提出利用图神经网络模型来分析网络通信行为，以检测僵尸网络。Boyaci等[10]提出一种基于GNN的错误数据注入攻击检测器，旨在保护智能电网计量设备测量数据的完整性。Lo等[6]将GNN应用在物联网入侵检测技术研究中，并首次尝试改进图神经网络模型GraphSAGE（graph sample and aggregate）以适应网络流量图上的边分类任务，最终的分类性能显示了GNN在入侵检测任务中的潜力。Duan等[11]提出一种基于半监督学习的动态线图神经网络的入侵检测方法，首次考虑将网络的拓扑信息和各IP对之间的信息交互过程结合在GNN模型中，从而提高了检测性能。如今，基于GNN的恶意流量检测与识别方法已被认为是应对网络安全威胁的一种有效手段[2]。

现有的恶意流量识别模型大多基于监督学习的基本假设[12-14]，即训练数据和测试数据是独立同分布（IID, independentidentically distributed）的。然而，在实际的应用场景中，尤其是在工业互联网恶意流量识别的场景中，由于噪声随机性、设备配置更新、良性/恶意网络通信行为随时空变化而发生改变等原因，数据分布偏移情况难以避免。网络流量属于流式数据，原则上，实际测试数据分布未知，尤其是在新型未知攻击频增的情况下，这对基于GNN的模型在应对数据分布偏移时的识别能力提出巨大挑战。典型的基于GNN的恶意流量识别模型通过自身的消息传递及特征聚合机制学习流量的高级表征，进一步完成下游的识别任务，典型的基于GNN的网络流量表征学习过程如图1所示，最终学习到的流量边表征将受到局部图结构及邻域内信息的影响。

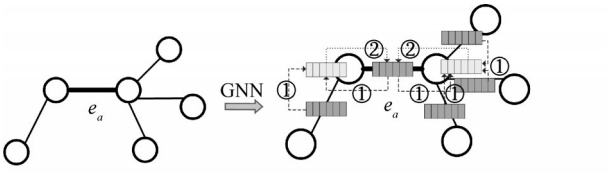

图1 典型的基于GNN的网络流量表征学习过程

然而，在新增未知攻击场景下，目标边的局部结构及邻域内信息可能会产生变化，导致数据分布产生偏移。训练阶段和测试阶段的网络流量如图2所示。训练集中的流量边e_a与测试集中的流量边e_q为同一类型，模型在训练阶段已能有效识别与e_a同一类型的恶意流量。然而，当测试数据中出现新增未知攻击时，模型对与e_a同一类型的恶意流量的识别能力严重下降。基于GNN的恶意流量识别模型[6]的泛化性能如表1所示，经实验验证，该模型在测试数据中出现新增未知类攻击的情况下，对已知类攻击的识别总体准确率下降不低于22%。这是因为连接同一节点的流量边e_q与流量边e_p在现实的网络环境中可能来自不同类别的网络通信行为，当应用现有的GNN模型学习流量边e_q的表征时，不可避免地会在消息传递和特征聚合的过程中融入e_q与e_p间的伪同质性特征，而这种特征分布是模型在训练阶段未曾学习过的数据分布，因此会出现性能严重下降的情况。

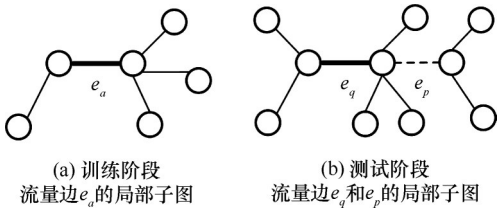

图2 训练阶段和测试阶段的网络流量

表 1　基于 GNN 的恶意流量识别模型的泛化性能

数据集	训练与测试数据分布一致		测试数据中有新增未知类攻击	
	准确率	F_1	准确率	F_1
60 万个样本	0.98	0.97	0.76	0.64
6 万个样本	0.86	0.86	0.63	0.59

针对上述问题，本文面向基于图神经网络的识别模型，提出一种基于邻域过滤和稳定学习的恶意流量识别方法，主要工作如下。

1) 将网络流量数据建模为以边为中心的图结构数据，利用可扩展的 GraphSAGE 学习流量表征，通过邻域过滤机制划分流量子图，消除流量间原始通信行为的伪同质性，提升流量边嵌入向量与标签之间的因果一致性。

2) 应用稳定学习思想，通过自适应的样本重采样方法学习观测样本权重，进一步消除流量高维表征间的非线性虚假关联，结合加权最小二乘算法进行优化，获得加权分布后流量表征的统计独立性。

3) 在公开数据集上的实验结果表明，与现有先进方法相比，所提方法在新增未知网络攻击行为导致数据分布偏移的情况下，能够保持相对稳定的恶意流量识别能力，可提升已知类恶意流量识别性能超 2.7%，同时可有效增强模型在数据不充分场景下的已知类恶意流量识别能力。

1　方法设计

为了有效应对未知数据分布偏移对基于 GNN 的恶意流量识别模型带来的挑战，本文所提出的新增未知攻击场景下的工业互联网恶意流量识别方法总体架构如图 3 所示，主要包括 4 个模块，网络流量图构建模块、基于有偏采样的网络流量子图划分模块、基于 GNN 的流量表示学习模块和基于稳定学习思想的协同损失优化模块。首先，通过构建工业互联网流量图（IITG, industrial Internet traffic graph）来建模网络通信行为，并将恶意流量识别任务转换为图上的边分类任务；然后，通过基于有偏采样的邻域过滤机制将网络流量图划分为 n 个以目标流量边为中心的自我子图，其中 n 为流量样本总数，应用原始特征相似性过滤原则，显式去除目标边自我子图内的伪同质性；接着，将 n 个目标边子图输入 GraphSAGE，通过所设计的消息传递机制获得目标边的高级表征；最后，基于稳定学习思想的损失约束，对下游分类任务的损失与确保流量边高级表征独立性的损失协同优化，进一步去除流量边表征间的虚假统计相关性，近似获得模型的泛化能力上限，以此提升恶意流量识别模型在实际应用场景中的稳定性和泛化性。

1.1　IITG 构建

本文使用的网络流量样本为基于 NetFlow 格式的网络流量记录。使用 NetFlow 等专用工具提取网络流量特征所形成的流量记录，是在生产网络中记录网络通信行为的通用格式，同时也是恶意流量检测与上下文识别中最常见的数据格式。流量记录通常包括提供通信标识的五元组信息（源 IP、源端口、目的 IP、目的端口、协议），以及数据包总数、字节总数、流持续时间等流量统计信息。

本文通过构建 IITG 来建模网络通信行为，并将恶意流量识别任务转换为图上的边分类任务。具体来说，将 Netflow 格式的网络流量数据建模为图中的节点和边，将除四元组之外的记录信息作为边的特征 h_t，将 IP 和端口号组成的二元组 $(\mathrm{IP}, \mathrm{Port})$ 作为图上的节点，以源 IP 和源端口向目的 IP 和目的端口通信的网络流量定义节点间的关系，即图上的

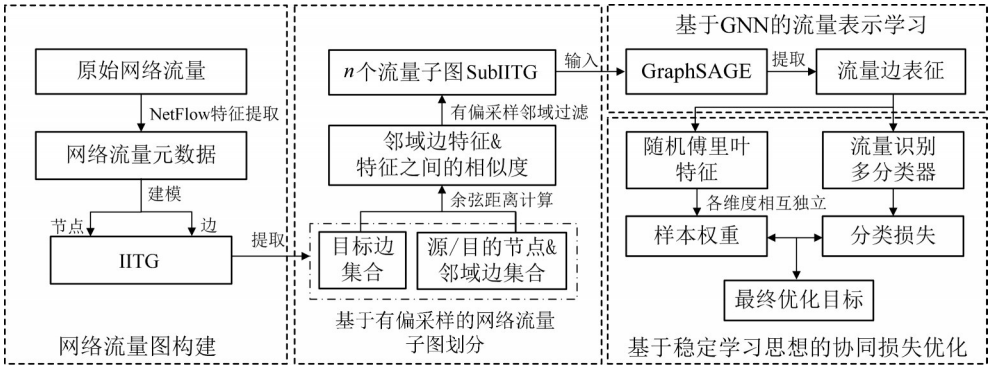

图 3　新增未知攻击场景下的工业互联网恶意流量识别方法总体架构

边。例如，当源节点（172.16.185.16:57967）与目标节点（192.168.1.156:80）交换数据时，相应的网络流量即表示两节点间的边。

具体的形式化描述如下。IITG = (V,E)是一个有向图，其中，$V=\{v_1,v_2,\cdots,v_m\}$是所有流量样本中(IP,Port)元组构建的节点集合，v_i表示图中第i个节点，m表示图中节点的总数；$E=\{e_1,e_2,\cdots,e_n\}$是所有网络流组成的边集合，$e_t=\{v_i,v_j,h_t\}$表示图中第t条边，v_{src}和v_{dst}为源节点和目的节点，h_t为边的特征，n为图中边的总数。

1.2 基于有偏采样的网络流量子图划分

新增未知网络攻击行为带来的数据分布偏移会导致现有基于图神经网络的恶意流量识别模型性能严重下降。为应对流量边邻域内伪同质性对恶意流量识别任务的影响，本文提出一种基于有偏采样的网络流量子图划分机制。具体来说，以目标边e为中心，基于图神经网络的消息传递思想，将与e连接于相同v_{src}和v_{dst}的流量边作为e的原始邻域，并通过余弦距离公式计算与源节点v_{src}相连的所有邻域边特征和e邻域边特征之间的相似度，基于相似度排序过滤e邻域内的伪同质性边，保留相似度排序在前top$k$的邻域边，并记为$e_{\text{src_neib}}$。采用同样的方式对目的节点v_{src}相连的所有邻域边进行过滤，保留相似度排序在前top$k$的邻域边，并记为$e_{\text{dst_neib}}$。最后，将以e为中心的自我子图记为$\text{SubIITG}_{\text{targ}}=\left(e,v_{\text{src}},v_{\text{dst}},E_{\text{src_neib}},E_{\text{dst_neib}}\right)$，其中$E_{\text{src_neib}}$和$E_{\text{dst_neib}}$为2个流量边集合。通过上述的有偏采样邻域过滤机制将原有的IITG划分为n个流量子图$\text{SubIITG}_{\text{targ}}$，其中targ = $1,2,\cdots,n$，所获得的流量子图集合将作为后续流量表示学习模型的输入。基于有偏采样的网络流量子图划分算法如算法1所示。

算法1 基于有偏采样的网络流量子图划分算法

输入 网络流量图IITG，目标边集合E，保留的邻域边数k

输出 流量子图集合SubIITGs = { $\text{SubIITG}_{\text{targ}}$, targ = $1,2,\cdots,n$ }

1) 初始化流量子图集合SubIITGs ← { }
2) for每一个目标边$e\in E$ do:
3) 获得源节点v_{src}和目的节点v_{dst}
4) 初始化邻域边集合$E_{\text{src_neib}}$←{}，$E_{\text{dst\_neib}}$←{}
5) for每个与v_{src}相连的邻域边$e_{\text{src_neib}}$ do:
6) 计算边特征相似度sim_{src}← Cosine Similarity$(h_{e_{\text{neib}}},h_e)$
7) if sim_{src}处于前top$k$，then
8) 添加$e_{\text{src_neib}}$到$E_{\text{src_neib}}$
9) end if
10) end for
11) for每个与v_{dst}相连的邻域边$e_{\text{src_neib}}$ do:
12) 计算边特征相似度sim_{dst}← Cosine Similarity$(h_{e_{\text{neib}}},h_e)$
13) if sim_{dst}处于前top$k$，then
14) 添加$e_{\text{dst_neib}}$到$E_{\text{dst_neib}}$
15) end if
16) end for
17) 构建子图$\text{SubIITG}_{\text{targ}}\leftarrow(e,v_{\text{src}},v_{\text{dst}},E_{\text{src_neib}},E_{\text{dst_neib}})$
18) 添加$\text{SubIITG}_{\text{targ}}$到SubIITGs
19) end for
20) 返回SubIITGs

1.3 基于GraphSAGE的流量表示学习

GraphSAGE是图神经网络模型的重要代表之一[6]，其核心思想是图上每个节点通过采样和聚合其局部邻域内节点的特征来生成目标节点的嵌入表示，进而为各种下游任务提供基础。具体的节点嵌入更新规则表达式为

$$\boldsymbol{h}_v^{(k)}=\sigma\left(\boldsymbol{W}^{(k)}\,\text{CONCAT}\left(\boldsymbol{h}_v^{(k-1)},\text{AGG}_{\{u\in N(v)\}}\left(\boldsymbol{h}_u^{(k-1)}\right)\right)\right) \tag{1}$$

其中，$\boldsymbol{h}_v^{(k)}$是节点v在第k层的嵌入，σ是一个非线性激活函数，$\boldsymbol{W}^{(k)}$是第k层权重矩阵，AGG是一个聚合函数，如均值、池化等，$N(v)$是节点v的邻居节点集合。

受文献[6]的启发，本文基于改进的GraphSAGE学习网络流量边的高级嵌入，这些嵌入能够捕获原始流量数据中的重要信息和结构属性。由于原始GraphSAGE主要用于节点嵌入表示学习，因此本文通过修改原始GraphSAGE模型以考虑边特征和子图结构，具体通过以下3个步骤完成。

1) 以边为中心的子图输入。与传统的GraphSAGE不同，本文模型接受目标边的自我子图$\text{SubIITG}_{\text{targ}}$作为输入，这使得模型能够专注于与目

标边相关的局部结构，并生成更具代表性的边嵌入。

2) 改进的消息传递和聚合函数。本文修改了消息传递函数以包含边特征。具体来说，消息函数将节点特征与边特征进行拼接，并通过一个线性层进行处理，使用平均聚合函数来聚合从邻居边传递的信息，如式(2)所示。

$$\boldsymbol{m}_{ij} = \boldsymbol{W}_{\text{msg}}\,\text{CONCAT}\,(\boldsymbol{h}_i,\boldsymbol{h}_{e_{ij}}) \tag{2}$$

其中，$\boldsymbol{m}_{ij}$是从节点v_i到节点v_j的信息，$\boldsymbol{W}_{\text{msg}}$是消息层的权重矩阵，$\boldsymbol{h}_{e_{ij}}$是边$e_{ij}$的特征。

3) 生成目标边嵌入。在完成所有层的消息传递后，对目标边e的源节点v_{src}嵌入$\boldsymbol{h}_{v_{\text{src}}}$和目的节点$v_{\text{dst}}$嵌入$\boldsymbol{h}_{v_{\text{dst}}}$进行拼接，然后将拼接后的嵌入$\boldsymbol{z}'_e$通过一个线性层和非线性激活函数来生成最终的边嵌入$\boldsymbol{z}_e$。

$$\boldsymbol{z}'_e = \text{CONCAT}\,(\boldsymbol{h}_{v_{\text{src}}},\boldsymbol{h}_{v_{\text{dst}}}) \tag{3}$$

$$\boldsymbol{z}_e = \sigma\,(\boldsymbol{W}_z\boldsymbol{z}'_e) \tag{4}$$

其中，$\boldsymbol{z}'_e$是源节点v_{src}嵌入和目的节点v_{dst}嵌入拼接后的向量，$\boldsymbol{W}_z$是线性层的权重矩阵，σ是非线性激活函数ReLU，$\boldsymbol{z}_e$是边e的最终嵌入，同时记$\boldsymbol{Z}_e$为IITG中所有流量边的嵌入集合。

1.4　基于稳定学习思想的协同损失优化

从理论上讲，当实际的恶意流量识别任务面临数据分布变化时，基于GNN的识别模型性能可能会大幅度下降，这主要是由虚假关联引起的，这些虚假关联主要来自无关表示和相关表示之间的微妙关联[15]。本文在基于有偏采样的流量子图划分模块中已经显性地去除了流量样本间明显的伪同质性，但难以完全去除流量的高维表征间潜在的虚假关联。因此，本文提出基于稳定学习思想[16]的协同损失优化模块，通过联合优化基于GraphSAGE的流量边表征编码器F、恶意流量识别任务的下游多分类器R和可学习的样本权重，学习解相关的流量边表征，以进一步消除无关表征和相关表征之间的依赖关系，提升工业互联网恶意流量识别模型的识别能力。模块主要由以下3个关键部分组成。

1) 基于随机傅里叶特征[17-20]的非线性解相关方法。假定通过GraphSAGE得到了流量边的高维表征$\boldsymbol{Z}_e$，为消除这些表征间的伪依赖，应确保每个维度之间相互独立。具体来说，期望流量边嵌入$\boldsymbol{Z}_e$各个维度间满足

$$\boldsymbol{Z}_e^{(a)} \perp\!\!\!\perp \boldsymbol{Z}_e^{(b)},\forall a,b \in [1,d\,],a \neq b \tag{5}$$

其中，$\boldsymbol{Z}_e$是d维流嵌入向量，a和b分别表示d维向量中任意不同的两维，$\perp\!\!\!\perp$表示相互独立。

为了实现上述目标，本文利用随机傅里叶特征来近似非线性核，从而能在欧几里得空间中衡量流量表征间的统计依赖性。本文定义了一个优化目标，通过最小化表征间的统计相关性，推动模型学习到更为独立的流量边表征。具体来说，首先选择P个随机傅里叶特征函数，对每个流量边表征$\boldsymbol{z}_e$进行变换，得到的新表征用于计算表征间的偏协方差矩阵。随机傅里叶特征变换为

$$f\,(\boldsymbol{Z}_e^{(a)}) := (f_1\,(\boldsymbol{Z}_e^{(a)}),f_2\,(\boldsymbol{Z}_e^{(a)}),\cdots,f_P\,(\boldsymbol{Z}_e^{(a)})) \tag{6}$$

$$g\,(\boldsymbol{Z}_e^{(b)}) := (g_1\,(\boldsymbol{Z}_e^{(b)}),g_2\,(\boldsymbol{Z}_e^{(b)}),\cdots,g_P\,(\boldsymbol{Z}_e^{(b)})) \tag{7}$$

其中，f_P和g_P是从随机傅里叶特征函数空间中选择的函数，每个函数通过不同的随机参数ω和φ生成。

随后，计算由随机傅里叶特征变换得到的新流量边表征间的偏协方差矩阵$\sum f\,(\boldsymbol{Z}_e^{(a)}),g\,(\boldsymbol{Z}_e^{(b)})$，并通过最小化该偏协方差矩阵的Frobenius范数L_{dec}来推动新流量表征间的解相关，以鼓励模型学习到不同维度上相互独立的流量边表征。具体为

$$\sum f\,(\boldsymbol{Z}_e^{(a)}),g\,(\boldsymbol{Z}_e^{(b)}) = \frac{1}{n-1}\sum_{k=1}^{n}(f\,(\boldsymbol{Z}_k^{(a)}) - \frac{1}{n}\sum_{l=1}^{n} f\,(\boldsymbol{Z}_l^{(a)}))^{\mathrm{T}}(g\,(\boldsymbol{Z}_k^{(a)}) - \frac{1}{n}\sum_{l=1}^{n} g\,(\boldsymbol{Z}_l^{(a)})) \tag{8}$$

$$L_{\text{dec}} = \sum_{1 \leqslant a < b \leqslant d}\left\| \sum f\,(\boldsymbol{Z}_e^{(a)}),g\,(\boldsymbol{Z}_e^{(b)}) \right\|_{\mathrm{F}}^{2} \tag{9}$$

其中，$\|\cdot\|_{\mathrm{F}}$表示Frobenius范数。

2) 可学习的样本权重。为了进一步消除流量边表征Z_e中的虚假关联，本文应用一种样本权重学习方法[21]。通过引入可学习的样本权重$\boldsymbol{w}$，重新定义流量边表征$\boldsymbol{Z}_e$间的偏协方差矩阵。设$\boldsymbol{w}$为n维正值向量，表示可学习到的样本权重，满足$\sum_{k=1}^{n}\boldsymbol{w}_k = n$。重新定义的偏协方差矩阵可表示为

$$\sum\nolimits^{\wedge} f\,(\boldsymbol{Z}_e^{(a)}),g\,(\boldsymbol{Z}_e^{(b)}):\boldsymbol{w} = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{w}_k f\,(\boldsymbol{Z}_k^{(a)}) - \frac{1}{n}\sum_{l=1}^{n}(\boldsymbol{w}_l f\,(\boldsymbol{Z}_l^{(a)}))^{\mathrm{T}}(\boldsymbol{w}_k g\,(\boldsymbol{Z}_k^{(a)}) - \frac{1}{n}\sum_{l=1}^{n}(\boldsymbol{w}_l g\,(\boldsymbol{Z}_l^{(a)})) \tag{10}$$

通过优化样本权重$\boldsymbol{w}$来最大程度地减少$\boldsymbol{Z}_e$间的统计依赖性，优化目标可表示为

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}\in\triangle_n}\sum_{1\leqslant a<b\leqslant d}\left\|\sum^{\wedge} f(\boldsymbol{Z}_e^{(a)}), g(\boldsymbol{Z}_e^{(b)}):\boldsymbol{w}\right\|_{\mathrm{F}}^2 \tag{11}$$

其中，$\triangle_n=\{\boldsymbol{w}\in R_+^n|\sum_{k=1}^n \boldsymbol{w}_k=n\}$。

3) 协同优化的判别性流量边表征学习。通过协同优化流量边表示学习模型参数和高维表征解相关的样本权重$\boldsymbol{w}$，学习用于识别恶意流量的判别性流量表征。在每个迭代步骤t，本文通过执行带有加权损失的反向传播来优化流量边表征编码器F和下游多分类器R。优化目标为

$$F^*,R^* = \arg\min_{F,R}\sum_{k=1}^{n}\boldsymbol{w}_k \operatorname{Loss}(R\circ F(\mathrm{SubIITG}_k),Y_k) \tag{12}$$

其中，Loss(·,·)表示交叉熵损失函数，n表示训练集中流量样本数量，$\mathrm{SubIITG}_k$和Y_k分别表示第k个以目标边为中心的流量子图和其对应的标签，$\boldsymbol{w}_k$为可学习到的流量样本k的样本权重，∘表示R作用在F后。这种方法确保了模型学习到的最终流量表征在不同的维度上是相互独立的，从而提高恶意流量识别模型的稳定性。

基于稳定学习思想的协同损失优化算法如算法2所示。

算法2 基于稳定学习思想的协同损失优化算法

输入 $\mathrm{SubIITGs}=\{\mathrm{SubIITG}_k\}_{k=1}^n$

输出 学习到的F^*,R^*

1) for e=1 to Epoch do
2) for 每个小批次$\mathrm{SubIITG}_{\mathrm{batch}}=\{\mathrm{SubIITG}_k\}_{k=1}^{\mathrm{batch}}$ do:
3) 计算$\boldsymbol{Z}_{\mathrm{e_batch}}=\{\boldsymbol{Z}_k\}_{k=1}^{\mathrm{batch}}$，$\boldsymbol{Z}_k=\mathrm{F}(\mathrm{SubIITG}_k)$
4) 初始化$\boldsymbol{w}$=(1,1,⋯,1)
5) 通过最小化式(11)优化$\boldsymbol{w}$
6) 按式(12)执行加权损失的反向传播
7) end for
8) end for

2 实验分析

2.1 实验环境

实验的服务器配置为：Intel(R) Xeon(R) Silver 4214R CPU @ 2.40 GHz，128 GB内存，NVIDIA GeForce RTX 3090显卡（24 GB显存）。训练和测试实验均在Linux bogon 4.18.0-477.21.1.el8_8.x86_64内核操作系统上完成，使用的编程语言和主要开源软件库版本为Python 3.10.12、Pytorch 2.0.1、dgl 1.1.1。

2.2 数据集

本实验采用BoT-IoT[22]数据集的扩展版本NF-BoT-IoT-v2，该数据集是基于NetFlow v9网络元数据构建的标准入侵检测特征集，包含43个具有明确业务含义的特征[23]。由于工业互联网通常包含大量的IoT设备，NF-BoT-IoT-v2数据集基于IoT网络流量生成，因此本文使用它来进行工业互联网恶意流量识别方法的研究和评估。

具体来说，NF-BoT-IoT-v2数据集包含37 763 497条数据流，其中99.64%是攻击样本，0.36%是良性样本。攻击样本包含4个主要的攻击类别，分别介绍如下。

侦察（Reconnaissance）：共计2 620 999个样本，这种攻击主要是为了收集网络主机的信息。

拒绝服务（DoS）攻击：共计16 673 183个样本，这种攻击试图通过超载计算机系统资源来阻止其对数据的访问或使用。

分布式拒绝服务（DDoS）攻击：共计18 331 847个样本，这种攻击类似于DoS，但是来自多个不同的分布式来源。

窃取（Theft）：共计2 431个样本，这种攻击旨在获取敏感数据（如窃取数据和键盘记录）。

本文探索了在出现新增未知网络攻击行为时，即使数据分布发生偏移，模型对已知类攻击流量的识别能力依旧保持良好的解决方案。为验证方案的有效性，本文从NF-BoT-IoT-v2数据集中随机抽取了良性样本和4种不同类别的网络流量样本。由于窃取攻击在该数据集中的原始样本数量较少，本文选择所有2 431个窃取攻击样本，这也符合现实世界中部分攻击行为较少且整体样本类别不均衡的实际情况。在完成数据清洗后，良性样本总计135 037个。为了更全面地评估模型性能，同时综合考量样本数量和多样性对模型性能的影响，本文建立了2个规模的数据集，在数据集NF-BoT-IoT-v2-Base（简称Base）中，保留全部135 037个良性样本；在数据集NF-BoT-IoT-v2-Mini（简称Mini）中，对良性样本随机降采样到10%，其余3类攻击样本数量均与当前数据集中良性样本数量基本一致，本文使用的数据集样本数量情况如表2所示。

表2　本文使用的数据集样本数量情况

数据集	Benign/个	DoS/个	DDoS/个	Reconnaissance/个	Theft/个
Base	135 037	127096	115 166	134 729	2 431
Mini	13 500	10 263	10 207	9 859	2 431

2.3　评价指标

本文所关注的恶意流量识别任务即分类任务，在分类任务中，TP为被模型预测为正类的正样本，TN为被模型预测为负类的负样本，FP为被模型预测为正类的负样本，FN为被模型预测为负类的正样本。参照分类任务通常采用的衡量指标，本文的恶意流量识别任务所采用的4个衡量指标分别为精确率（P）、召回率（R）、F1值（F_1）和整体准确率（Acc），具体计算式为

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{14}$$

$$F_1 = \frac{2PR}{P + R} \tag{15}$$

$$\mathrm{Acc} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{16}$$

2.4　实验设置

为了对工业互联网恶意流量识别现实场景中新增未知攻击导致的模型性能下降问题进行研究，本文设计了问题验证性实验和有效性评估实验，以在2个数据集上评估基线模型和所提方法，同时通过消融实验讨论本文所设计的各个子模块对模型识别能力的影响。

本文主要关注的数据分布偏移情况为新增未知类恶意攻击行为导致的数据分布偏移对基于图神经网络的识别模型在识别已知类攻击行为时产生的性能影响。因此，本文设计了问题验证性实验，以证明本文所关注的数据分布偏移问题属实存在且具有挑战。具体来说，原始数据集中的各个类别记为{class 0, class 1, class 2, class 3, class 4}，其中class 0为良性样本，class 1~class 4为4类攻击样本。首先，在原始数据集中随机选出某一个或几个类别的流量样本作为新增未知类攻击样本，记为Unknown，并将除Unknown之外的其余类别样本统一记为Known。然后对数据集进行划分，将Known按7∶3划分为已知类训练集Known Train和已知类测试集Known Test，在Known Test中加入Unknown样本作为分布偏移测试集Unknown Test。本文在只包含已知类样本的训练集上训练识别模型，然后在包含未知类样本的测试集上评估模型对已知类样本的识别性能。本文选择当前先进的基于图神经网络的物联网恶意流量检测模型E-GraphSAGE[6]作为基线模型（简称GSAGE），完成上述问题验证性实验。

本文主要通过2个关键模块来缓解所考虑的新增未知攻击对识别模型的性能影响，一是基于有偏采样的邻域过滤模块去除流量间明显的伪同质性；二是基于稳定学习的协同损失优化模块去除流量高维表征间的虚假统计相关性。为了验证2个关键模块提升识别模型泛化能力的有效性，本文一并进行了有效性评估实验和消融实验。

在有效性评估实验中，本文将基线方法GSAGE、对比方法E-ResGAT[24]（简称ResGAT）与所提方法在2个数据集上进行全面的性能评估，其中，对比方法ResGAT是在GSAGE基础上改进的模型。在消融实验中，本文进一步验证了所设计的2个关键模块以及各子模块内部实现方法对模型性能的作用。

2.5　问题验证性实验与结果分析

在问题验证性实验中，本文应用基线模型GSAGE分别在2个数据集上进行实验，基线模型在分布内数据和分布外数据上的性能比较如表3所示，其中新增未知类别Unknown以DDoS攻击为例。从实验结果来看，基线模型在分布外数据上的性能较已知分布数据上的性能下降0.22~0.23，在大样本量数据集Base上的识别整体准确率从0.98下降到0.76，在小样本量数据集Mini上的识别整体准确率从0.86下降到0.63。此外，在2个数据集上模型的平均F_1指标下降更为严重，分别从0.97和0.86下降到了0.64和0.59。这意味着，在实际的应用场景中，现有基于GNN的基线模型泛化能力不足，适应性有限，在面对新增未知攻击产生时，模型性能下降非常严重。主要原因在于，当新增未知类别流量与已知类别流量一并建模为网络流量图结构数据时，流量图的数据分布发生了偏移，已知类流量数据与新增未知类流量数据之间因共享同一邻域存在伪同质性关系，这种邻域内的伪同质性特征会通过图表示学习过程影响流量高维表征间的统计相关性，因此对识别模型性能造成了严重影响。

表3 基线模型在分布内数据和分布外数据上的性能比较

数据集	测试集	已知类别	P	R	F_1	Acc
Base	Known Test	Benign	1.00	1.00	1.00	0.98
		DoS	0.95	0.99	0.97	
		Recon-naissance	0.98	0.95	0.96	
		Theft	0.93	0.99	0.96	
	Unknown Test	Benign	0.99	0.95	0.97	0.76
		DoS	0.99	0.35	0.52	
		Recon-naissance	0.6	0.96	0.74	
		Theft	0.18	0.99	0.31	
Mini	Known Test	Benign	0.94	0.82	0.87	0.86
		DoS	1.00	0.82	0.90	
		Recon-naissance	0.70	0.98	0.82	
		Theft	0.99	0.76	0.86	
	Unknown Test	Benign	0.95	0.78	0.86	0.63
		DoS	0.89	0.28	0.43	
		Recon-naissance	0.76	0.71	0.73	
		Theft	0.19	0.99	0.32	

2.6 有效性评估实验与结果分析

为了有效缓解未知类攻击引发的数据分布偏移对恶意流量识别模型性能的不利影响，本文提出了一种新的恶意流量识别方法，记为 SGSL（subgraph partitioning and stable learning）。该方法旨在增强模型对已知攻击类型的识别能力，同时保持识别性能的稳定性。通过系列实验，本文对所提 SGSL 方法、基线模型 GSAGE 以及对比方法 ResGAT 进行了全面比较。实验结果证明了 SGSL 模型在对已知攻击类别识别上的有效性和鲁棒性。

对比方法在 Base 数据集上的性能比较如表 4~表 7 所示。可以看出，SGSL 对各种新增未知攻击类型的整体识别准确率达到了 0.96，显著高于 GSAGE 和 ResGAT 的平均值，分别提升了 0.36 和 0.04。此外，在 F_1 分数方面，SGSL 以 0.93 的平均值较 GSAGE 和 ResGAT 分别高出 0.29 和 0.05，这些显著提升充分说明了 SGSL 的有效性。同时，SGSL 之所以在面对未知攻击导致数据分布偏移的情况下仍能保持高准确率和 F_1 分数，是因为本文所设计的 2 个关键模块可以有效学习到流量的稳定表示，后续消融实验将对相关模块作用进行更具体的分析。

对比方法在 Mini 数据集上的性能比较如表 8~表 11 所示，进一步验证了 SGSL 方法的适应性和泛化能力。SGSL 在不同攻击类别的识别上均保持了稳定的精确率、召回率和 F_1 分数。在未知 DDoS 攻击的情况下，SGSL 取得了 0.90 的整体识别准确率，这个结果优于基线模型 GSAGE 在分布内数据上的相应指标，验证了 SGSL 在数据量不足的情况下仍可有效捕捉流量的判别性特征的能力。整体而

表4 对比方法在 Base 数据集上的性能比较（Unknown=DDoS）

方法	Benign			DoS			Reconnaissance			Theft			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.99	0.95	0.97	0.99	0.35	0.52	0.60	0.96	0.74	0.18	0.99	0.31	**0.76**	**0.63**
ResGAT	0.98	0.79	0.88	0.94	0.93	0.93	0.75	0.95	0.84	0.92	0.92	0.92	**0.88**	**0.84**
SGSL	0.98	0.98	0.98	0.96	0.98	0.97	0.96	0.94	0.95	0.92	0.98	0.94	**0.97**	**0.96**

表5 对比方法在 Base 数据集上的性能比较（Unknown=DoS）

方法	Benign			DDoS			Reconnaissance			Theft			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.82	0.95	0.88	0.98	0.99	0.98	0.23	0.96	0.37	0.89	1.00	0.94	**0.43**	**0.64**
ResGAT	0.96	0.81	0.88	0.99	0.99	0.99	0.83	0.97	0.90	0.85	0.71	0.77	**0.92**	**0.88**
SGSL	0.97	1.00	0.99	0.95	1.00	0.97	1.00	0.93	0.96	0.97	1.00	0.99	**0.97**	**0.96**

表6 对比方法在 Base 数据集上的性能比较（Unknown= Reconnaissance）

方法	Benign			DoS			DDoS			Theft			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.19	0.72	0.30	0.74	0.88	0.80	0.71	1.00	0.83	0.72	0.33	0.45	**0.41**	**0.48**
ResGAT	0.97	0.98	0.98	0.98	0.97	0.98	0.99	0.99	0.99	0.63	0.50	0.56	**0.98**	**0.88**
SGSL	0.97	1.00	0.99	1.00	0.93	0.96	0.95	1.00	0.97	0.98	0.35	0.52	**0.97**	**0.86**

表 7　对比方法在 Base 数据集上的性能比较（Unknown=Theft）

方法	Benign			DoS			Reconnaissance			DDoS			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.96	0.66	0.79	0.95	0.93	0.94	0.70	0.64	0.67	0.69	0.97	0.81	**0.80**	**0.80**
ResGAT	0.97	0.79	0.87	0.94	0.91	0.92	0.77	0.93	0.84	0.99	1.00	0.99	**0.90**	**0.91**
SGSL	0.97	0.78	0.86	0.95	0.98	0.96	0.95	0.98	0.96	0.99	0.99	0.99	**0.92**	**0.92**

表 8　对比方法在 Mini 数据集上的性能比较（Unknown=DDoS）

方法	Benign			DoS			Reconnaissance			Theft			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.95	0.78	0.86	0.89	0.28	0.43	0.76	0.71	0.73	0.19	0.99	0.32	**0.63**	**0.58**
ResGAT	0.96	0.72	0.82	0.88	0.92	0.90	0.68	0.92	0.78	0.85	0.68	0.76	**0.83**	**0.82**
SGSL	0.95	0.85	0.89	0.95	0.98	0.96	0.80	0.91	0.85	0.87	0.78	0.82	**0.90**	**0.88**

表 9　对比方法在 Mini 数据集上的性能比较（Unknown=DoS）

方法	Benign			DDoS			Reconnaissance			Theft			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.93	0.83	0.88	0.98	0.99	0.99	0.21	0.98	0.35	0.67	0.28	0.40	**0.46**	**0.52**
ResGAT	0.98	0.72	0.83	0.97	0.99	0.98	0.71	1.00	0.83	0.89	0.68	0.77	**0.87**	**0.85**
SGSL	0.86	0.80	0.83	0.99	0.99	0.99	0.78	0.90	0.84	0.87	0.69	0.77	**0.87**	**0.85**

表 10　对比方法在 Mini 数据集上的性能比较（Unknown= Reconnaissance）

方法	Benign			DoS			DDoS			Theft			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.30	0.99	0.46	0.67	0.94	0.78	0.99	0.96	0.98	0.52	0.36	0.43	**0.48**	**0.53**
ResGAT	0.93	0.93	0.93	0.93	0.94	0.94	0.99	0.99	0.99	0.83	0.80	0.82	**0.94**	**0.92**
SGSL	0.99	0.96	0.98	0.98	0.99	0.98	0.99	0.98	0.98	0.80	0.95	0.87	**0.97**	**0.95**

表 11　对比方法在 Mini 数据集上的性能比较（Unknown=Theft）

方法	Benign			DoS			Reconnaissance			DDoS			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.52	0.78	0.62	0.96	0.40	0.56	0.56	0.95	0.70	0.98	0.99	0.99	**0.66**	**0.58**
ResGAT	0.92	0.76	0.83	0.88	0.91	0.90	0.70	0.84	0.76	0.98	0.99	0.99	**0.86**	**0.87**
SGSL	0.96	0.77	0.85	0.92	0.92	0.92	0.72	0.91	0.80	0.99	0.99	0.99	**0.89**	**0.89**

言，无论是在 Base 数据集还是在 Mini 数据集上，相较于基线方法和对比方法，SGSL 在分布外数据上的识别性能提升超过 2.7%。

2.7　消融实验与结果分析

1) 关键模块消融分析。首先，对本文所提的 2 个关键模块进行消融实验，以进一步验证和分析相关模块对模型整体性能的影响。将单独融合了邻域过滤模块的基线模型记为 SSG，将单独融合了协同损失优化模块的基线模型记为 SLG，将 SSG、SLG、GSAGE 和 SGSL 在 2 个数据集上进行比较。

Mini 数据集上的模型有效性消融实验结果如表 12 所示。在融合了基于有偏采样的子图划分模块后，SSG 的识别整体准确率由 0.63 提升到 0.86，这是因为该模块可以显式去除流量间因共享同一邻域导致的目标边邻域内的伪同质性，使得基于图表示学习的流量表征编码器在为目标边聚合邻域特征同一邻域时，避免聚合异类流量特征；在融合了协同损失优化模块后，SLG 的识别整体准确率的提升也较为显著，由 0.63 升至 0.84，这是由于新增未知攻击所产生的数据分布偏移使得最终学习到的流量表征在统计上存在虚假相关性，协同损失优化模块可有效去除这种虚假关联，使模型可以依赖稳定的判别性特征来预测样本标签。从上述结果来看，在面对新增未知攻击导致的数据分布偏移问题时，2 个

表12 Mini数据集上模型有效性消融实验结果

方法	Benign			DoS			Reconnaissance			Theft			Acc	F_1
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1		
GSAGE	0.95	0.78	0.86	0.89	0.28	0.43	0.76	0.71	0.73	0.19	0.99	0.32	**0.63**	**0.58**
SSG	0.96	0.84	0.90	0.94	0.97	0.96	0.80	0.87	0.83	0.77	0.93	0.84	**0.86**	**0.88**
SLG	0.99	0.79	0.88	0.99	0.78	0.87	0.64	0.99	0.78	0.97	0.80	0.88	**0.84**	**0.85**
SGSL	0.95	0.85	0.89	0.95	0.98	0.96	0.80	0.91	0.85	0.87	0.78	0.82	**0.90**	**0.88**

关键的子模块均对模型的识别性能有显著提升效果，同时整体方案SGSL的识别性能最佳，整体识别准确率由0.63升至0.90，这充分表明，在通过显式去除流量边邻域内明显的伪同质性基础上，进一步去除高维流量表征在统计上潜在的虚假关联是非常有必要的，2个模块在协同使用时可以更有效地增强模型对分布外流量识别任务的适应能力。

2) 初始特征消融分析。然后，本文对网络流量图构建模块所使用的初始特征进行消融分析，目的是评估基于NetFlow格式的流量特征对当前任务的影响。将流量（边）特征初始化为无信息含义的随机数，其余设置保持不变，并将应用此特征设置的GSAGE模型记为GSAGE-NFF，将应用此特征设置的SGSL模型记为SGSL-NFF。上述4个模型在2个数据集上的初始特征消融实验结果对比如图4所示，将流量（边）特征替换为随机数之后的基线模型GSAGE-NFF与本文所提方法SGSL-NFF在Base数据集上的准确率分别降低到0.51和0.75，在Mini数据集上的准确率只能分别达到0.42和0.57。这一方面说明基于NetFlow的流量特征对当前任务是必要的；另一方面也显示出在结合了本文提出的关键模块后，基于图神经网络的解决方案具有了更强的空间特征学习能力，SGSL-NFF在没有初始流量特征的情况下，识别恶意流量的准确率接近带有初始流量特征的GSAGE。

3) 子图划分模块消融分析。最后，本文对子图划分模块的有偏采样邻域过滤机制进行消融分析，目的是为本文方法在实际应用中的可行性提供合理优化方案。具体来说，本文在子图划分模块中采用的邻域边相似度计算是基于精准相似度计算（PSC, precise similarity calculation）方法实现的。需要说明的是，该方法同时也是高开销的，其计算复杂度近似$O(n^2)$，其中n为流量边总数，这在需要对超大规模样本进行分类的情况下，计算开销可能难以负担。因此，本文一并实现了另一种基于局部敏感哈希（LSH, locally sensitive hashing）的快速邻域过滤方法[25]。该方法是在权衡了精确率与开销之后对精确率的妥协之策，其计算复杂度可降低到近似$O(n)$。在消融实验中，将单独融合了LSH的基线方法记为GSAGE-LSH，将单独融合了PSC的基线方法记为GSAGE-PSC，将子图划分模块方法替换为LSH的SGSL模型记为SGSL-LSH，需要注意的是，GSAGE-PSC即上文的SSG。以Unknown=DDoS为例，上述3种方法与本文方法SGSL在2个数据集上的子图划分模块消融实验结果对比如图5所示。

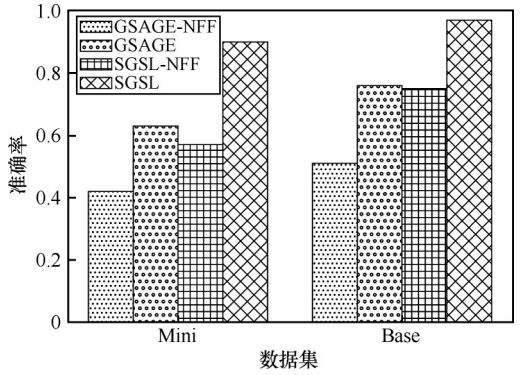

图4 初始特征消融实验结果对比

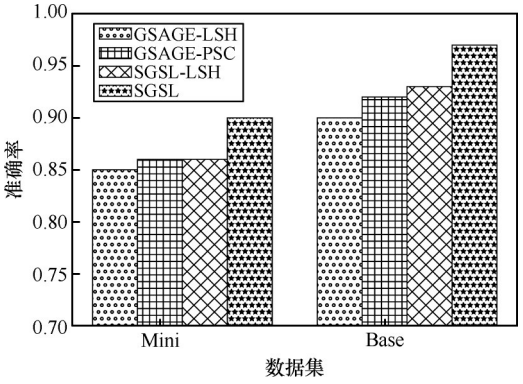

图5 子图划分模块消融实验结果对比

通过实验结果分析可知，无论是Base数据集还是Mini数据集，基于PSC实现的GSAGE模型在恶意流量识别准确率方面都更胜一筹；基于LSH

的 SGSL-LSH 虽然在 Base 数据集上识别准确率为 0.92，略低于 SGSL，但考虑其更低的计算复杂度，在需要处理海量网络流量的现实场景中，或许可以作为一种更实用的替代方案。此外，在本文所提模型的基础上，可额外集成更多性能优化技术，如批处理、并行化、分布式计算（如 Apache Spark）以及深度图学习（DGL）图计算加速库等策略，以进一步提升模型在实际应用中的处理性能。

3　结束语

本文聚焦于工业互联网安全的核心问题，面向恶意流量识别任务，通过深入分析图神经网络在恶意流量识别中的应用，基于实验发现在新增未知类别攻击场景下，现有方法识别已知类恶意流量的性能严重下降。为了解决这一问题，本文提出了一种新的恶意流量识别方法，通过邻域过滤机制将网络流量数据转换为以边为中心的子图结构数据，并利用图神经网络模型学习流量表征，进一步通过表征独立性的协同目标优化，增强最终流量表征的判别性。实验结果显示，与现有的基于 GNN 的恶意流量识别模型相比，本文方法在面对新增未知网络攻击行为导致的流量数据分布偏移情况下，能够保持相对稳定的恶意流量识别能力，并实现了对已知类恶意流量识别性能的显著提升。特别是在小样本数据集上，本文方法的整体识别准确率达到了 90% 以上，已明显超过了现有模型在分布内数据上的识别性能。此外，通过消融实验，本文进一步验证了所设计的邻域过滤模块和协同损失优化模块对识别模型性能提升的重要作用。这些结果不仅验证了本文方法的有效性，也为实际场景下的恶意流量识别问题提供了新的思路和解决方案。

参考文献：

[1] 蔡岳平, 李栋, 许驰, 等. 面向工业互联网的 5G-U 与时间敏感网络融合架构与技术[J]. 通信学报, 2021, 42(10): 43-54.
CAI Y P, LI D, XU C, et al. Integrating 5G-U with time-sensitive networking for industrial Internet: architectures and technologies[J]. Journal on Communications, 2021, 42(10): 43-54.

[2] 黄韬, 汪硕, 黄玉栋, 等. 确定性网络研究综述[J]. 通信学报, 2019, 40(6): 160-176.
HUANG T, WANG S, HUANG Y D, et al. Survey of the deterministic network[J]. Journal on Communications, 2019, 40(6): 160-176.

[3] NUAIMI M, FOURATI L C, HAMED B. Intelligent approaches toward intrusion detection systems for industrial Internet of things: a systematic comprehensive review[J]. Journal of Network and Computer Applications, 2023, 215:103637.

[4] FU C P, LI Q, XU K. Detecting unknown encrypted malicious traffic in real time via flow interaction graph analysis[J]. arXiv Preprint, arXiv: 2301.13686, 2023.

[5] WALI K N, ALSHEHRI MOHAMMED S, KHAN MUAZZAM A, et al. A hybrid deep learning-based intrusion detection system for IoT networks[J]. Mathematical Biosciences and Engineering: MBE, 2023, 20(8): 13491-13520.

[6] LO W W, LAYEGHY S, SARHAN M, et al. E-GraphSAGE: a graph neural network based intrusion detection system for IoT[J]. arXiv Preprint, arXiv: 2103.16329, 2021.

[7] ALWASEL B, ALDRIBI A, ALRESHOODI M, et al. Leveraging graph-based representations to enhance machine learning performance in IIoT network security and attack detection[J]. Applied Sciences, 2023, 13(13): 7774.

[8] CARLETTI V, FOGGIA P, VENTO M. Detecting abnormal communication patterns in IoT networks using graph neural networks[C]//Proceedings of the Graph-Based Representations in Pattern Recognition. New York: ACM Press, 2023: 127-138.

[9] ZHOU J W, XU Z Y, RUSH A M, et al. Automating botnet detection with graph neural networks[J]. arXiv Preprint, arXiv: 2003.06344, 2020.

[10] BOYACI O, UMUNNAKWE A, SAHU A, et al. Graph neural networks based detection of stealth false data injection attacks in smart grids[J]. IEEE Systems Journal, 2022, 16(2): 2946-2957.

[11] DUAN G H, LV H W, WANG H Q, et al. Application of a dynamic line graph neural network for intrusion detection with semisupervised learning [J]. IEEE Transactions on Information Forensics and Security, 2022, 18: 699-714.

[12] KUANG K, CUI P, ATHEY S, et al. Stable prediction across unknown environments[C]// Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. New York: ACM Press, 2018: 1617-1626.

[13] LIA P Y, YAN J, SELLIER J M, et al. TADA: a transferable domain-adversarial training for smart grid intrusion detection based on ensemble divergence metrics and spatiotemporal features[J]. Energies, 2022, 15(23): 8778.

[14] KHEDDAR H, HIMEUR Y, AWAD A I. Deep transfer learning for intrusion detection in industrial control networks: a comprehensive review[J]. arXiv Preprint, arXiv:2304.10550, 2023.

[15] ARJOVSKY M, BOTTOU L, GULRAJANI I, et al. Invariant risk minimization[J]. arXiv Preprint, arXiv: 1907.02893, 2019.

[16] ZHANG X X, CUI P, XU R Z, et al. Deep stable learning for out-of-distribution generalization[J]. arXiv Preprint, arXiv: 2104.07876, 2021.

[17] REN M Y, ZENG W Y, YANG B, et al. Learning to reweight examples for robust deep learning[J]. arXiv Preprint, arXiv: 1803.09050, 2018.

[18] RAHIMI A, RECHT B. Random features for large-scale kernel machines [C]//Proceedings of the 2007 Conference onNeural Information Processing Systems. New York: ACM Press, 2007: 1177-1184.

[19] LI Z, TON J F, OGLIC D, et al. Towards A unified analysis of random Fourier featuresJ]. arXiv Preprint, arXiv: 1806.09178, 2018.

[20] LI H Y, WANG X, ZHANG Z W, et al. OOD-GNN: out-of-distribution generalized graph neural network[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 7328-7340.

[21] WU Z R, XIONG Y J, YU S X, et al. Unsupervised feature learning via non-parametric instance discrimination[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscat-

away: IEEE Press, 2018: 3733-3742.
[22] KORONIOTIS N, MOUSTAFA N, SITNIKOVA E, et al. Towards the development of realistic botnet dataset in the Internet of things for network forensic analytics: bot-IoT dataset[J]. arXiv Preprint, arXiv: 1811.00701, 2018.
[23] SARHAN M, LAYEGHY S, PORTMANN M. Towards a standard feature set for network intrusion detection system datasets[J]. Mobile Networks and Applications, 2022, 27(1): 357-370.
[24] CHANG L Y, BRANCO P. Graph-based solutions with residuals for intrusion detection: the modified E-GraphSAGE and E-ResGAT algorithms[J]. arXiv Preprint, arXiv: 2111.13597, 2021.
[25] DATAR M, IMMORLICA N, INDYK P, et al. Locality-sensitive hashing scheme based on p-stable distributions[C]//Proceedings of the Twentieth Annual Symposium on Computational Geometry. New York: ACM Press, 2004: 253-262.

[作者简介]

曾凡一（1993– ），女，蒙古族，辽宁昌图人，哈尔滨工程大学博士生，主要研究方向为网络入侵检测、加密恶意流量分析。

苘大鹏（1980– ），男，辽宁抚顺人，博士，哈尔滨工程大学教授、博士生导师，主要研究方向为网络流量安全监测、新型网络与人工智能安全。

许晨（1996– ），男，山东菏泽人，博士，哈尔滨工程大学讲师、硕士生导师，主要研究方向为自然语言处理、人工智能安全。

韩帅（1991– ），女，黑龙江哈尔滨人，博士，哈尔滨工程大学讲师、硕士生导师，主要研究方向为大数据管理与安全等。

王焕然（1988– ），男，黑龙江哈尔滨人，博士，哈尔滨工程大学讲师、硕士生导师，主要研究方向为图表示学习、深度学习模型可解释性等。

周雪（1994– ），女，黑龙江牡丹江人，哈尔滨工程大学博士生，主要研究方向为网络安全、人工智能安全。

李欣纯（1999– ），女，黑龙江齐齐哈尔人，哈尔滨工程大学博士生，主要研究方向为数据安全与隐私保护。

杨武（1974– ），男，辽宁宽甸人，博士，哈尔滨工程大学教授、博士生导师，主要研究方向为网络与信息安全、人工智能应用及安全。