计算机科学
COMPUTER SCIENCE

基于Transformer的域自适应物联网流量入侵检测方法

朱枫, 叶宗国, 李鹏, 徐鹤

引用本文

朱枫, 叶宗国, 李鹏, 徐鹤. 基于Transformer的域自适应物联网流量入侵检测方法[J]. 计算机科学, 2026, 53(3): 443-452.

ZHU Feng, YE Zongguo, LI Peng, XU He. Transformer-based Domain Adaptation Method for IoT Traffic Intrusion Detection [J]. Computer Science, 2026, 53(3): 443-452.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200167

基于 Transformer 的域自适应物联网流量入侵检测方法

朱　枫[1]　叶宗国[1]　李　鹏[1,2]　徐　鹤[1,2]
1 南京邮电大学计算机学院　南京 210023
2 江苏省无线传感网络高技术研究重点实验室　南京 210023
(zhufeng@njupt.edu.cn)

摘　要　随着物联网(Internet of Things,IoT)设备的普及,使用入侵检测来保护 IoT 设备免受恶意攻击至关重要。但是,IoT 的数据稀缺性限制了传统入侵检测方法的效果。同时,现有基于域自适应的入侵检测方法的对齐方式粗糙,忽略了内在语义属性的转移,降低了特征的可区分性。为解决上述问题,提出了一种基于 Transformer 的域自适应物联网入侵检测(Transformer-Based Domain-Adaptive IoT Intrusion Detection,TDAIID)模型,从域间、类间和样本间 3 个层次对齐互联网入侵(Network Intrusion,NI)域和物联网入侵(Internet of Things Intrusion,II)域。交叉注意力机制聚焦于 NI 源域和 II 目标域中相同类别样本之间的相似特征,实现样本级别的域特征对齐;多重几何语义对齐从域级和类级两个角度进行语义对齐,有助于交叉注意力机制学习更丰富、更准确的源 NI 域知识。此外,为了充分挖掘未标记 II 目标域的潜力,从几何角度提出了一种动态中心感知伪标签算法,用于提高伪标签标记的准确性,有效降低错误分配伪标签造成的负迁移。在多个常用入侵检测数据集上的综合实验表明,TDAIID 模型的性能优于当前先进的基线模型。

关键词:域自适应;物联网;入侵检测;交叉注意力;迁移学习
中图分类号　TP391

Transformer-based Domain Adaptation Method for IoT Traffic Intrusion Detection

ZHU Feng[1],YE Zongguo[1],LI Peng[1,2] and XU He[1,2]
1 College of Computer Science,Nanjing University of Posts and Telecommunications,Nanjing 210023,China
2 Jiangsu High Technology Research Key Laboratory for Wireless Sensor Networks,Nanjing 210023,China

Abstract　With the proliferation of IoT devices,intrusion detection systems(IDS) are essential to safeguard IoT networks from malicious attacks. However,the scarcity of IoT-specific data limits the effectiveness of traditional methods,while existing domain adaptation approaches often rely on coarse alignment,overlooking intrinsic semantic properties and lowering feature discriminability. To address these issues,this paper proposes a semi-supervised domain adaptation model,named TDAIID. This model aligns NI domain and II domain at domain,class,and sample levels. The cross-attention mechanism ensures fine-grained feature alignment by focusing on similarities between same-class samples in the source and target domains. Multiple geometric semantic alignment is semantically aligned from both domain-level and class-level perspectives,facilitating the cross-attention mechanism in learning richer and more accurate knowledge from the source NI domain. To fully exploit unlabeled target data,a dynamic center-aware pseudo-labeling algorithm is proposed to improve pseudo-label accuracy and mitigate negative transfer caused by mislabeling. Experiments on several widely-used intrusion detection datasets demonstrate that the TDAIID model outperforms state-of-the-art baseline methods,showcasing its superior performance on IoT intrusion detection.

Keywords　Domain adaptation,Internet of Things,Intrusion detection,Cross-attention,Transfer learning

1　引言

随着物联网(Internet of Things,IoT)设备的应用越来越普遍[1],医疗保健、公共交通等多个领域已经逐步向智能化转变。然而,物联网基础设施通常由资源有限的设备组成,这些设备的供应商很少进行安全维护,使得恶意攻击者可以利用物联网的安全漏洞进行入侵,从而破坏底层的物联网基础设施及其支持的应用程序。因此,建立强大的入侵检测系统(Intrusion Detection System,IDS)对物联网的安全防护至关重要[2]。

到稿日期:2024-12-23　返修日期:2025-03-07
基金项目:国家自然科学基金(61902196,62102196);江苏省科技支撑计划(BE2019740);江苏省六大人才高峰高层次人才项目(RJFW-111)
This work was supported by the National Natural Science Foundation of China(61902196,62102196),Scientific and Technological Support Project of Jiangsu Province(BE2019740) and Six Talent Peaks Project of Jiangsu Province(RJFW-111).
通信作者:李鹏(lipeng@njupt.edu.cn)

近年来，基于规则和机器学习(Machine Learning，ML)技术的物联网流量入侵检测系统越来越受到关注。例如，Xie等[3]提出物联网入侵检测特征筛选算法，并将其应用于决策树、随机森林等多种机器学习模型上，实验结果表明，该算法有效提高了入侵检测性能。然而，这些方法或需要建立复杂的规则库，或依赖于大量标记完整的数据，这对于物联网流量入侵检测来说尤为困难，因为物联网设备生成的流量数据通常涉及用户隐私问题[4]，且专业知识和手动注释的成本较高，它们限制了物联网入侵检测数据的公开发布，以及物联网场景下基于规则和ML方法的有效性。

针对传统ML方法的缺陷，可以借助迁移学习技术来提升模型的学习效果。然而，在利用迁移学习进行物联网流量数据分类时，由于不同物联网设备、场景以及攻击行为具有差异，收集的流量数据也存在较大的不同。因此，互联网入侵(NI)域的流量样本或基于NI域样本训练的模型不能直接应用于物联网入侵(II)域中。域自适应[5](Domain Adaption，DA)作为迁移学习的一类方法，能够从有完全标记数据的源域获得知识，将知识转移到具有少量未标记数据的不同但相似的目标域。通过从知识丰富的源域转移知识，可以促进在类似但数据稀缺的目标领域的学习。然而在域自适应问题中，源域和目标域虽然共享相同的标签空间，但这并不能保证两者的数据相似性，因此仍可能出现跨域差异较大的问题，直接迁移的效果可能不理想。现有的领域对齐方法主要集中在减少跨领域的全局差异，主要包括基于差异的对齐方法和基于对抗的对齐方法[6]。许多学者通过将两域的数据样本映射到另一个空间，然后使用分布对齐的方法来减小源域和目标域之间的分布差异，增加两域的相似性。这种对齐通常通过缩小联合分布、边际分布和条件分布之间的距离来实现。在网络入侵检测领域，互联网和物联网之间共享几种常见的攻击类型，通过将两个域映射到一个共同的特征子空间，这些域自适应方法可以传递丰富的网络入侵知识，从而有助于物联网领域中的入侵检测。例如，Ly等[7]利用两个自动编码器作为源域和目标域的特征提取器，并最小化其瓶颈层之间的最大均值差异(Maximum-Mean Discrepancy，MMD)以实现知识转移。Peng等[8]通过寻找最优变换矩阵，适配源域与目标域之间的条件概率和边缘概率，实现源域与目标域间的特征迁移。然而，这些基于域自适应的入侵检测模型通常只进行粗粒度的对齐，它们通过强制性手段将源域和目标域对齐到一个共同的特征子空间中，却忽略了对内在语义属性的转移，可能会导致来自不同类别的样本混淆在一起，从而降低学习到的特征的可区分性。

针对现有域自适应方法存在的问题，受已有工作的启发，本文提出一种基于Transformer的域自适应物联网流量入侵检测模型(TDAIID)。该模型针对大量完全标记的NI源域流量数据和少部分有标记的II域流量数据，从域级、类别级和样本级3个层次对齐源域和目标域，从丰富的NI源域学习知识，最终实现对物联网流量的入侵检测。本文的主要贡献总结如下：

1)提出的TDAIID模型首次将交叉注意力机制应用于流量数据中的知识迁移。其不仅通过多重几何语义对齐从域和类间两个角度对齐源域和目标域，还通过交叉注意力机制聚焦于源域和目标域中相同攻击类型样本间的相似特征，帮助实现更精准的跨域特征对齐和融合。

2)提出一种动态中心感知算法，从模型预测结果和几何相似性两方面综合考虑给无标记II目标域数据分配伪标签，提高伪标签分配准确率，减少来自不同类别样本的干扰，更准确地构造类别相同的源域和目标域数据对，供模型进行特征知识的学习。

3)在4个常用的入侵检测领域数据集上进行实验，结果表明了TDAIID模型的有效性，其在多个评价指标上均超过了当前先进的基线模型[1)]。

2 相关工作

2.1 传统的入侵检测方法

传统的入侵检测采用基于签名的检测方法，通过维护一组恶意攻击的签名或规则，并将传入的网络流量与这些预定义的攻击模式进行匹配来检测攻击。Christian等[9]提出通过定期扫描连接的物联网设备中的预定义恶意行为来主动检测攻击，扫描规则基于漏洞报告信息自动更新。Douglas等[10]提出了一种通过有效位模式匹配的轻量级深度数据包异常检测策略，通过分析数据包中有效载荷内容的模式，并利用n-gram匹配算法迅速构建特征表示。

随着ML技术快速发展，ML开始被广泛应用于物联网场景下的入侵检测。Valerian等[11]提出了一种基于联邦学习的入侵检测框架，使用多层感知机和自编码器神经网络架构对物联网设备进行异常检测和分类。Mojtaba等[12]、Sarumathi等[13]和Li等[14]都利用基于ML的方法，分别构建了随机森林、多核支持向量机和卷积神经网络来执行物联网设备的入侵检测，并取得了不错的效果。

然而，上述方法或依赖于全面和不断更新的规则库，需要复杂的专业知识来构建，成本很高；或依赖于完全标记的训练数据集。此外，资源能力不足和相关数据隐私的问题，导致物联网数据稀缺且难以获得。因此，研究基于DA的入侵检测方法是有必要的，它可以解决物联网场景下的数据稀缺和数据标注问题。

2.2 域自适应及其在入侵检测中的应用

DA能从知识丰富的领域转移知识，以促进对类似但知识稀缺的目标领域的学习。其中源域和目标域呈现异质性，例如，互联网和物联网领域的入侵数据可能来源于不同类型的设备，这些设备在不同的环境下工作，遵循不同的分布等。Xie等[15]提出了一种联合对齐模型，通过Wasserstein距离最小化和对抗学习执行全局域对齐，还通过最小化两个分类器产生的概率输出之间的距离来实现知识转移。Yao等[16]提出了一种区分性分布对齐方法，结合了交叉熵损失和平方损失这两种损失函数，以提高对齐期间数据的可区分性。

在基于DA的入侵检测方面，Ning等[17]提出了一种半监

1) https://github.com/yessya1999/TDAIID

督模型，通过复用源域部分特征提取层参数和 MMD 特征对齐，从一个小规模的源域转移知识，以促进目标域的入侵检测。Hu 等[18]提出了一种结合注意力机制的深度卷积域自适应网络，利用局部 MMD 最小化来对齐相似的源域和目标域，并利用注意力机制来防止收敛时间过长。为了解决物联网的数据稀缺问题，Wu 等[19]提出了一种自适应双向推荐和自我改进网络，通过推荐系统的双向推荐兴趣匹配在共享特征空间中对齐 NI 域和 II 域，同时参考推荐系统的决策，分配更准确的伪标签。虽然基于 DA 的方法已被应用于入侵检测领域，但是这些方法大多通过距离度量方式，从域整体和类别之间粗粒度对齐源域和目标域，未能综合考虑从域间、类间和样本间等多个角度实现细粒度的语义对齐。此外，这些方法均没有考虑综合使用自注意力机制和交叉注意力机制来促进攻击知识的学习和传递。

3　模型介绍

3.1　模型整体架构

3.1.1　模型定义

TDAIID 模型在半监督环境下工作，将具有 n_s 条流量数据的 NI 源域定义为 $\mathcal{D}_s=\{\mathcal{X}_s,\mathcal{Y}_s\}=\{(x_i^s,y_i^s)\}_{i=1}^{n_s}$，其中 y_i^s 是 x_i^s 的对应标签，标签类别个数为 K。II 目标域仅有少量数据被标记，定义如式(1)所示：

$$\begin{aligned}&\mathcal{D}_{\mathrm{TL}}=\{\mathcal{X}_{\mathrm{TL}},\mathcal{Y}_{\mathrm{TL}}\}=\{(x_{\mathrm{TL}i},y_{\mathrm{TL}i})\}\\&\mathcal{D}_{\mathrm{TU}}=\{\mathcal{X}_{\mathrm{TU}}\}=\{(x_{\mathrm{TU}j})\},\mathcal{D}_{\mathrm{T}}=\mathcal{D}_{\mathrm{TL}}\cup\mathcal{D}_{\mathrm{TU}}\\&y_{\mathrm{TL}i}\in[1,K],i\in[1,n_{\mathrm{TL}}],j\in[1,n_{\mathrm{TU}}]\\&n_t=n_{\mathrm{TL}}+n_{\mathrm{TU}},n_{\mathrm{TL}}\ll n_{\mathrm{TU}}\end{aligned}\tag{1}$$

其中，$\mathcal{D}_{\mathrm{TL}}$ 是有标签的 II 目标域数据构成的集合，$\mathcal{D}_{\mathrm{TU}}$ 是无标签 II 目标域数据构成的集合；n_t 是整个 II 目标域数据数量，考虑到 $\mathcal{D}_{\mathrm{T}}$ 数据较为稀缺，故 n_t 小于 n_s。同时，考虑到对目标域数据进行标记代价高昂，故假设 II 目标域中有标记的数据量 n_{TL} 远少于无标记的数据量 n_{TU}。两个域的样本分别从两个不同但相关的概率分布 P 和 Q 中提取，即 $P\neq Q$。本文提出了一个统一的域自适应框架，该框架可以学习语义信息的域相似表示，以便在 II 目标域上进行泛化。本文所使用的符号及其相应解释如表 1 所列。

表 1　符号及其含义

Table 1　Symbols and their interpretations

符号	符号含义
$\mathcal{D}_s$	NI 源域
$\mathcal{D}_{\mathrm{T}}$	II 目标域
$\mathcal{D}_{\mathrm{TL}}$	少量被标记 II 目标域
$\mathcal{D}_{\mathrm{TU}}$	大量无标记 II 目标域
n_t	$\mathcal{D}_{\mathrm{T}}$ 的数据数量
n_{TL}	$\mathcal{D}_{\mathrm{TL}}$ 的数据数量
n_{TU}	$\mathcal{D}_{\mathrm{TU}}$ 的数据数量
$\mathcal{X}_s$	$\mathcal{D}_s$ 数据样本
$\mathcal{X}_t$	$\mathcal{D}_{\mathrm{T}}$ 数据样本
$\mathcal{X}_{\mathrm{TL}}$	$\mathcal{D}_{\mathrm{TL}}$ 数据样本
$\mathcal{X}_{\mathrm{TU}}$	$\mathcal{D}_{\mathrm{TU}}$ 数据样本
Y_s	$\mathcal{X}_s$ 的真实标签
Y_{TL}	$\mathcal{X}_{\mathrm{TL}}$ 的真实标签
$\hat{Y}_{\mathrm{TU}}$	$\mathcal{X}_{\mathrm{TU}}$ 分配的伪标签

3.1.2　整体框架

本文提出的基于 Transformer 架构的域自适应物联网入侵检测模型 TDAIID 的框架如图 1 所示。

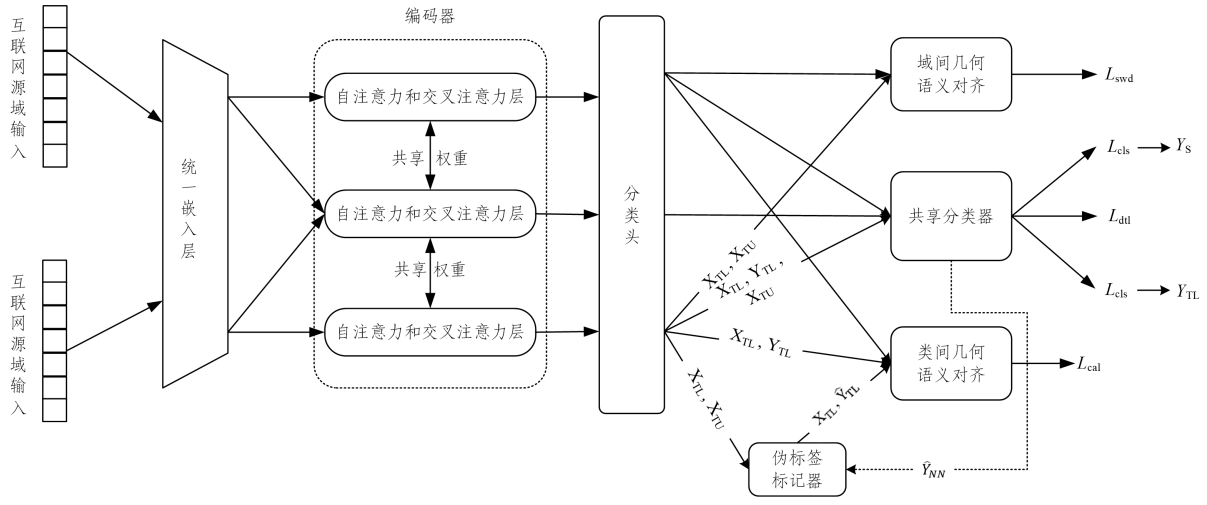

图 1　TDAIID 模型结构

Fig. 1　Model structure of TDAIID

首先，统一嵌入层将分类字段进行 One-Hot 编码，与数值字段结合后传递给嵌入层，将每条记录映射为连续的特征向量，并将源 $\mathcal{D}_s$ 与 $\mathcal{D}_{\mathrm{T}}$ 的流量输入映射到同一共享的公共特征子空间，转换成 Transformer 编码器能够处理的格式。编码器层由 3 个共享权重的 Transformer 注意力模块组成，分别命名为源分支、目标分支和源-目标分支。$\mathcal{D}_s$ 与 $\mathcal{D}_{\mathrm{T}}$ 的数据分别输入源分支和目标分支中，通过自注意力模块学习各自领域的特有特征；而在源-目标分支中，通过交叉注意力模块学习 $\mathcal{D}_s$ 与 $\mathcal{D}_{\mathrm{T}}$ 中不同样本之间的相似特征，从而促进知识迁移。

根据 Liam 等[20]的研究，IDS 分类任务通常只与最后一个流的类别相关，因此只取编码器层的最后一个输出向量，即来自最后一个流的特征向量，作为分类头的输入。由于 $\mathcal{D}_s$ 与 $\mathcal{D}_{\mathrm{T}}$ 共享相同的分类类别，所以 3 个分支使用相同的分类器进行训练，从而为模型提供监督信号。

在 $\mathcal{D}_s$ 与 $\mathcal{D}_{\mathrm{T}}$ 之间的域间隙过大的情况下，只使用交叉注

意力机制进行域自适应可能对齐效果不够理想，不同类别样本区分度不够。本文通过多重几何对齐的方式，从全局域和局部类别角度对齐 NI 和 II 域。在全局域层面，该方法使用切片 Wasserstein 距离[15]作为域间距离度量，整体对齐$\mathcal{D}_s$与$\mathcal{D}_T$，学习域级相似语义特征。在局部类别层面，使用类别质心对齐方法，最小化不同领域间相同类别实例之间的差距，学习具有区分度的类级相似语义特征。此外，为了充分挖掘未标记 II 目标域的潜力，本文从几何角度提出了一种动态中心感知伪标签标记的方法。该方法能够有效降低错误分配伪标签带来的负迁移，提高伪标签的准确性。

3.1.3 总体优化目标

$\mathcal{D}_s$和$\mathcal{D}_{TL}$的真实标签与共享分类器产生的预测结果通过交叉熵损失函数$\mathcal{L}_{ce}$计算得到监督损失函数$\mathcal{L}_{cls}$：

$$\mathcal{L}_{cls}=\frac{1}{n_s}\sum_{x_i\in\mathcal{X}_s,y_i\in\mathcal{Y}_s}\mathcal{L}_{ce}(C(f(x_i)),y_i)+\frac{1}{n_{TL}}\sum_{x_i\in\mathcal{X}_{TL},y_i\in\mathcal{Y}_{TL}}\mathcal{L}_{ce}(C(f(x_i)),y_i) \tag{2}$$

最终，TDAIID 模型的整体损失函数 Loss 的计算方式如式(3)所示：

$$Loss=\min_{E,C}(\mathcal{L}_{cls}+\rho\mathcal{L}_{swd}+\delta\mathcal{L}_{cal}+\tau\mathcal{L}_{dtl}) \tag{3}$$

其中，$\mathcal{L}_{swd}$和$\mathcal{L}_{cal}$分别为域级和类级语义对齐的损失函数，详见 3.3 节；$\mathcal{L}_{dtl}$作为蒸馏损失函数，用于实现源-目标分支到目标分支的知识迁移，详见 3.2 节；ρ，δ 和 τ 是控制相应模块权重的超参数，使用 Adam 梯度下降以端到端的方式训练整个 TDAIID 模型。

3.2 模型编码器层设计

3.2.1 自注意力机制

针对流数据的自注意力机制是 TDAIID 模型编码器中的核心组件，能够有效捕捉网络会话中不同网络流之间的相互依赖关系。通过多个独立的自注意力头，模型可以在不同的子空间中并行地关注流量数据的不同特征，增强对流数据的表征能力。

假设输入为网络流序列 $X=[x_1,x_2,\cdots,x_n]$，每个网络流由多个特征组成。将这些特征嵌入一个矩阵中，再加上位置编码 PE，最终自注意力层的输入表示为 $\boldsymbol{Z}\in\mathbb{R}^{n\times d}$，其中 n 是时间步数，d 表示每个嵌入向量的维度。

注意力机制对嵌入特征 $\boldsymbol{Z}$ 进行线性变换，将其映射到 $\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 向量中，如式(4)所示：

$$\begin{aligned}\boldsymbol{Q}&=\boldsymbol{W}_q\boldsymbol{Z}\\\boldsymbol{K}&=\boldsymbol{W}_k\boldsymbol{Z}\\\boldsymbol{V}&=\boldsymbol{W}_v\boldsymbol{Z}\end{aligned} \tag{4}$$

其中，$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 分别表示查询矩阵、键矩阵和数值矩阵；$\mathbf{W}_q$，$\boldsymbol{W}_k$，$\boldsymbol{W}_v$分别是$\boldsymbol{Q}$，$\boldsymbol{K}$，$\boldsymbol{V}$ 的权重矩阵，在训练过程中随机初始化并更新。通过计算 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 之间的相似性，得到注意力分数。利用 softmax 函数进行归一化，得到注意力权重。然后将注意力权重应用到 $\boldsymbol{V}$ 上，得到最终的注意力输出，如式(5)所示：

$$Attn_{self}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

其中，d_k是缩放因子，用于避免在高维空间中点积计算结果过大造成数值不稳定。

3.2.2 交叉注意力机制

交叉注意力模块源自于自注意力模块，但它的输入由源域和目标域两组不同的流量数据组成。对于源域流量嵌入 $\boldsymbol{Z}_s$，生成查询矩阵 $\boldsymbol{Q}_s$；对于目标域流量 $\boldsymbol{Z}_t$，生成键矩阵 $\boldsymbol{K}_t$和数值矩阵 $\boldsymbol{V}_t$。交叉注意力机制通过计算 $\boldsymbol{Q}_s$和$\boldsymbol{K}_t$之间的相似度，生成注意力权重，然后利用注意力权重对目标域的值矩阵 $\boldsymbol{V}_t$ 进行加权求和，得到交叉注意力的输出，如式(6)所示：

$$Attn_{cross}(\boldsymbol{Q}_s,\boldsymbol{K}_t,\boldsymbol{V}_t)=\mathrm{softmax}\left(\frac{\boldsymbol{Q}_s\boldsymbol{K}_t^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}_t \tag{6}$$

其中，d_k是缩放因子；$\boldsymbol{Q}_s\boldsymbol{K}_t^{\mathrm{T}}$是源域和目标域特征向量之间的点积，用于计算相似性。不同的注意力权重反映了 $\boldsymbol{Q}_s$和 $\boldsymbol{K}_t$之间的相似性，来自目标域流量中更相似的特征将获得更大的权重，并对输出做出更大的贡献。

目前，许多研究人员[21-22]将交叉注意力用于特征融合。特别是在多模态任务中，他们应用交叉注意力模块来聚合和对齐来自两种模式的信息。鉴于交叉注意力在特征融合和对齐中的强大功能，本文使用交叉注意力机制来解决流量数据的域自适应问题，从而在知识迁移过程中实现更有效的特征融合和分类。通过在 NI 域和 II 域之间建立样本级别的交互关系，可以关注分布间的局部对应关系，捕获更细粒度的分布特征，而不仅仅依赖域间和类别间几何对齐，进而提高对齐精度。不同类别样本之间通过交叉注意力进行特征融合和对齐会弱化不同类别之间的界限，容易学习到噪声。只有当$\mathcal{X}_s$与$\mathcal{X}_t$属于同一类别时，交叉注意力机制学习到的特征对知识迁移才是真正有帮助的。因此，本文利用动态中心感知伪标签算法，尽可能给$\mathcal{D}_{TU}$数据分配准确的伪标签。结合$\mathcal{D}_{TL}$数据的真实标签，使整个$\mathcal{D}_T$都有标签。在数据预处理阶段，将$\mathcal{D}_s$与$\mathcal{D}_T$的相同类别样本凑成一个数据对，输入 TDAIID 模型进行训练。

3.2.3 编码器层结构

传统的 Transformer 模型最初联合使用深层的编码器和解码器。然而，当用于非顺序任务时(例如，入侵检测分类任务基于流输入序列生成单个分类输出)，解码器层可以被移除并用分类头替换，从而创建完全由编码器构建的模型。所以，TDAIID 模型只使用 Transformer 编码器模块，包括嵌入层、自注意力和交叉注意力层和分类头，如图 1 所示。根据 Liam 等[20]的研究结果，浅层 Transformer 编码器与深层 Transformer 编码器在入侵检测任务上性能相当。同时，当浅层编码器用于特定任务时，由于参数量少，模型只需要少量数据来训练，通常不容易发生过拟合。因此，整个编码器模块只使用两层自注意力和交叉注意力层，没有使用深层编码器。

编码器层中的自注意力和交叉注意力层结构如图 2 所示，源分支和目标分支使用多头自注意力机制，源-目标分支使用多头交叉注意力机制。源-目标分支的第一层输入来自其他两个分支。在第 N 层，交叉注意力模块的 $\boldsymbol{Q}$ 来自源分支第 N 层的$\boldsymbol{Q}_s$，而 $\boldsymbol{K}$ 与 $\boldsymbol{V}$ 来自目标分支的$\boldsymbol{K}_t$与$\boldsymbol{V}_t$。然后，交叉注意力模块输出对齐融合后的特征，这些特征与第 $N-1$ 层的输出$\boldsymbol{Z}_{st}^{(N-1)}$相加并归一化得到$\hat{\boldsymbol{Z}}_{st}^{(N)}$。$\hat{\boldsymbol{Z}}_{st}^{(N)}$经过前馈神经网络和层归一化得到第 N 层的输出$\boldsymbol{Z}_s^{(N)}$，如式(7)所示：

$$\hat{\boldsymbol{Z}}_{st}^{(N)}=LayerNorm(\boldsymbol{Z}_{st}^{(N-1)}+Attn_{\text{cross}}(\boldsymbol{Q}_s,\boldsymbol{K}_t,\boldsymbol{V}_t)) \tag{7}$$

由于交叉注意力机制给$\mathcal{D}_s$和$\mathcal{D}_T$中相似特征和不相似特征赋予不同的注意力权重，因此源-目标分支的特征不仅融合对齐了两个域的分布，对输入对中的噪声也具有一定的鲁棒性。因此，本文使用源-目标分支的输出来指导目标分支的训练，有利于提高目标域的分类准确率。其中，源-目标分支和目标分支分别为教师和学生。将源-目标分支中分类器的概率分布视为一个软标签，用蒸馏损失函数进一步监督目标分支，其中q_k和p_k分别是来自源-目标分支和目标分支的类别k的概率，如式(8)所示：

$$L_{\text{dtl}}=\sum_k q_k\log p_k \tag{8}$$

在推理过程中，只使用目标分支。输入是待测试的目标域流量数据，只有目标分支被触发，共享分类器的输出被用作最终的预测标签。

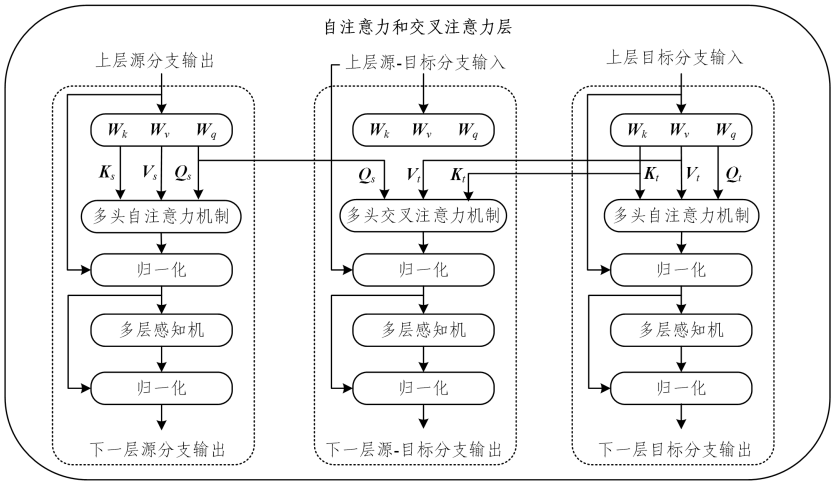

图 2　自注意力和交叉注意力层结构

Fig. 2　Structure of self-attention and cross-attention layer

3.3　多重几何语义对齐

考虑到 NI 域和 II 域之间的域间隙过大时，只使用交叉注意力机制进行域自适应可能对齐效果不够理想。通过进一步的多重几何语义对齐，分别从全局域之间的粗粒度语义对齐和局部类别之间的质心级细粒度对齐，缩小$\mathcal{D}_s$与$\mathcal{D}_T$间的语义差异。通过语义对齐，相同类别的$\mathcal{X}_s$与$\mathcal{X}_t$间的相似特征更加明显，不同类别的$\mathcal{X}_s$与$\mathcal{X}_t$间的差异更大，有利于交叉注意力机制的知识迁移。

域间粗粒度语义对齐方法的有效性在很大程度上取决于距离度量的可靠性。常用的距离度量方法有 MMD、KL-散度和 Wasserstein 距离(Wasserstein Distance，WD)等。MMD 方法依赖核函数的选择，难以适应高维特征空间；KL-散度在处理极端分布差异时表现不稳定，尤其是分布无交集时。而 WD 利用数据的几何特性，通过最优传输自然衡量分布差异，能更好地捕捉复杂流量特征间的关系，已被广泛应用于域自适应中[23]。对于 NI 域和 II 域这类高维异构数据，MMD 和 KL-散度均不适合，而 WD 能更精确地对齐分布，实现跨域特征的有效对齐。本文使用基于 WD 改进的切片 Wasserstein 距离(Sliced Wasserstein Distance，SWD)作为度量方式，通过 Radon 变换将高维分布估计分解为更容易估计的众多一维分布，使之比原始 WD 更容易计算。将μ和v的相应概率密度函数表示为F_μ和F_v，则F_μ与F_v间的 SWD 定义如式(9)所示：

$$SWD(F_\mu,F_v)=\sum_{m=1}^{M}\sum_{i=1}^{N}c(S(\mathcal{R}\circ F_\mu(\cdot,\alpha_m))_i,S(\mathcal{R}\circ F_v(\cdot,\alpha_m))_i) \tag{9}$$

其中，c是成本函数；$\mathcal{R}\circ F(\cdot,\alpha)$表示 Radon 变换，将函数$F_\mu$映射到超平面上的积分的集合。利用投影的有限总和简化求解 SWD：随机抽取从α_1到α_m的M个随机方向(超平面)上的投影，通过一维排序函数S对N个输入样本的投影$\mathcal{R}\circ F(\cdot,\alpha)$进行排序，随后求解一维最优运输问题并累加，得到 SWD 的近似值。最终，使用 SWD 实现跨域分布对齐，计算式如式(10)所示：

$$\mathcal{L}_{\text{swd}}(\mathcal{X}_s,\mathcal{X}_t)=\sum_{m=1}^{M}\sum_{i=1}^{N}c(S(\langle f(x_s),\alpha_m\rangle)_i,S(\langle f(x_t),\alpha_m\rangle)_i) \tag{10}$$

其中，$f(x_s)$和$f(x_t)$分别表示源域和目标域数据通过编码器输出的特征表示。

仅执行全局级的域对齐不足以实现细粒度的语义一致性，类别区分度不足。特别是当$\mathcal{D}_s$与$\mathcal{D}_T$差距过大时，粗粒度的全局对齐可能使得两个域中相同类别的数据训练后映射到不同的类别，损害语义转移的有效性。对于不同领域的分布对齐，使用质心进行类别间对齐是一种计算高效且易于实现的方法，能够在一定程度上缓解不同域相同类别之间分布上的差异。同时，在伪标签算法以及 II 目标域少量标记数据的帮助下，本文通过局部多重对齐的方式，综合考虑有标记和无标记的 II 目标域质心，相比于单纯的 NI 域和 II 域质心对齐，更能增进分布的相似度，实现类别间对齐。源域质心$\mu_S^{(k)}$、目标域质心$\mu_{\text{TL}}^{(k)}$和目标域联合质心$\mu_{\text{TT}}^{(k)}$的定义如式(11)所示：

$$\begin{aligned}\mu_S^{(k)}&=\frac{1}{|\mathcal{X}_s^{(k)}|}\sum_{x_i\in\mathcal{X}_s^{(k)}}f(x_i)\\ \mu_{\text{TL}}^{(k)}&=\frac{1}{|\mathcal{X}_{\text{TL}}^{(k)}|}\sum_{x_i\in\mathcal{X}_{\text{TL}}^{(k)}}f(x_i)\\ \mu_{\text{TT}}^{(k)}&=\frac{1}{|\mathcal{X}_{\text{TL}}^{(k)}\cup\mathcal{X}_{\text{TU}}^{(k)}|}\sum_{x_i\in\mathcal{X}_{\text{TL}}^{(k)}\cup\mathcal{X}_{\text{TU}}^{(k)}}f(x_i)\end{aligned} \tag{11}$$

通过目标域联合质心的构建，可以充分利用伪标签标记的目标域，同时结合有标记数据，一定程度上能容忍错误标记的干扰。此外，通过最小化不同域、不同类别质心之间的 L_2 距离，使同一类别更加相似，不同类别之间差距更大，实现细粒度的语义对齐。最终，类别语义对齐损失函数 L_{cal} 如式(12)所示：

$$L_{\text{cal}}=\sum_{k=1}^{K}(\|\mu_{\text{S}}^{(k)}-\mu_{\text{TL}}^{(k)}\|_2^2+\|\mu_{\text{S}}^{(k)}-\mu_{\text{TT}}^{(k)}\|_2^2)-\alpha\sum_{k=1}^{K}\sum_{l\neq k}\frac{1}{\|\mu_{\text{S}}^{(k)}-\mu_{\text{S}}^{(l)}\|_2^2+\|\mu_{\text{TL}}^{(k)}-\mu_{\text{TL}}^{(l)}\|_2^2} \tag{12}$$

其中，α 是权重系数，用来平衡类间分离性与类内紧凑性；$\|\mu_{\text{S}}^{(k)}-\mu_{\text{TL}}^{(k)}\|_2^2$ 表示 D_s 与 D_{TL} 中同类数据特征中心距离；$\|\mu_{\text{S}}^{(k)}-\mu_{\text{TT}}^{(k)}\|_2^2$ 表示 D_s 与 D_t 中同类数据特征中心距离；$\|\mu_{\text{S}}^{(k)}-\mu_{\text{S}}^{(l)}\|_2^2+\|\mu_{\text{TL}}^{(k)}-\mu_{\text{TL}}^{(l)}\|_2^2$ 表示 D_s 与 D_{TL} 中非同类数据特征中心距离。

3.4 动态中心感知伪标签算法

由于 $\mathcal{X}_s$ 与 $\mathcal{X}_t$ 之间的数据特征分布不同，直接使用模型分类器得到的输出结果为 $\mathcal{X}_{\text{TU}}$，分配伪标签存在风险[24]。错误的伪标签一旦被用于模型训练，会使模型更加倾向于错误的预测，从而形成恶性循环。这种误差在迭代过程中不断累积，使得模型的准确性进一步降低。例如，目标域第一和第三类别相似，分类器输出结果为[0.6,0.3,0.4,0.2,0.1]，原本属于第三类的目标样本可能会被强制标记为[1.0,0.0,0.0,0.0]。为了减轻这种影响，本文借助 $\mathcal{D}_{\text{TL}}$ 数据以及通过 $\mathcal{D}_{\text{TL}}$ 数据预训练的模型，提出了一种动态中心感知伪标签算法。

首先，实验结果表明，模型如果仅使用 $\mathcal{X}_{\text{TL}}$ 进行有监督训练，直接对 $\mathcal{X}_{\text{TU}}$ 进行分类，会出现过拟合的情况，对 $\mathcal{X}_{\text{TU}}$ 进行分类的准确率明显降低。因此，本文使用 $\mathcal{X}_{\text{TL}}$ 训练的模型作为预训练模型，将权重参数迁移到 TDAIID 模型中。相比随机初始化模型权重，TDAIID 不仅能够利用 $\mathcal{D}_{\text{TL}}$ 来提高每次训练迭代对 $\mathcal{X}_{\text{TU}}$ 分类的准确率，而且能够减少得到模型最优解的迭代次数。

其次，将 TDAIID 模型对 $\mathcal{X}_{\text{TU}}$ 的类别预测概率以及 $\mathcal{X}_{\text{TL}}$ 的真实类别作为权值，通过类似于加权 K-Means 聚类的方法得到整个目标域 $\mathcal{D}_{\text{T}}$ 中每个类的质心 C_k，如式(13)所示：

$$C_k=\frac{\sum_{x_{\text{TL}},y\in D_{\text{TL}},x_{\text{TU}}\in D_{\text{TU}}}(\delta_{\text{TU}}^{k},y_{\text{TL}})\{f(x_{\text{TL}}),f(x_{\text{TU}})\}}{\sum_{x_{\text{TL}},y_{\text{TL}}\in D_{\text{TL}},x_{\text{TU}}\in D_{\text{TU}}}(\delta_{\text{TU}}^{k},y_{\text{TL}})} \tag{13}$$

其中，δ_{TU}^{k} 表示 $\mathcal{X}_{\text{TU}}$ 通过共享分类器得到的在类别 k 上的概率，$f(x_{\text{TL}})$ 和 $f(x_{\text{TU}})$ 分别表示 $\mathcal{X}_{\text{TL}}$ 和 $\mathcal{X}_{\text{TU}}$ 通过编码器得到的特征提取结果。

在获得每个类别的质心之后，可以将每个 $\mathcal{X}_{\text{TU}}$ 数据分配到有最高余弦相似度的质心所在类别，如式(14)所示：

$$y_{\text{TU}}^{(\text{PL})}=\underset{k}{\text{argmax}}\{CS(f(x_{\text{TU}}),C_k)\},x_{\text{TU}}\in D_{\text{TU}} \tag{14}$$

其中，$CS()$ 是余弦相似度，用于计算每个 $\mathcal{X}_{\text{TU}}$ 距离的每个类别中心点的距离。

当 TDAIID 模型预测结果 $\hat{Y}_{\text{NN}}$ 和动态中心感知伪标签算法得到的伪标签 $y_{\text{TU}}^{(\text{PL})}$ 达成共识时，才分配真正的伪标签 $\hat{Y}_{\text{TU}}$。使用最终得到的 $\hat{Y}_{\text{TU}}$ 构建输入模型的相同类别的源-目标数据对。在数据预处理阶段，对于每个源-目标数据对，如果 $\hat{Y}_{\text{TU}}$ 或 Y_{TL} 的类别与 Y_s 一致，则保留这一对用于训练，否则作为噪声丢弃。因随着迭代训练的进行，模型分类准确率逐渐提升，故通过动态中心感知伪标签算法，每 5 轮动态更新一次 $\hat{Y}_{\text{TU}}$。

4 实验与分析

4.1 实验数据集

本文使用了 4 个广泛使用的入侵检测数据集进行验证，包括两个 NI 数据集(UNSW-NB15[25] 和 CICIDS 2018[26])，以及两个 II 数据集(UNSW-BOTIOT[27] 和 UNSW-TONIOT[28])。

1)NI 数据集(UNSW-NB15)。UNSW-NB15(N)数据集于 2015 年通过 IXIA PerfectStorm 工具生成，旨在解决此前 IDS 数据集中的冗余记录或缺失值等问题。它包含良性网络行为以及 9 种攻击类型，如 DoS 攻击、侦查攻击等。

2)NI 数据集(CICIDS 2018)。CICIDS 2018(C)数据集发布于 2017 年，是最新的网络流量数据集之一，使用 CICFlowMeter 收集数据。该数据集包含良性数据和 7 种常见的入侵攻击类型，反映了当前网络攻击的趋势，包括 DoS、DDoS、暴力破解攻击等。

3)II 数据集(UNSW-BOTIOT)。UNSW-BOTIOT(B)数据集创建于 2017 年，专注于现实中的 II 场景。该数据集涵盖 5 种物联网场景，包括气象站、智能冰箱和智能恒温器等，使用了物联网常用的轻量级通信协议 MQTT，包含 4 种常见的物联网攻击类别，如 DoS 攻击和信息窃取攻击等。本文还在半监督设置下，按照 1∶5，1∶10 和 1∶50 的比例调整 n_{TL}∶n_{TU} 的比率，确保未标记的目标 II 数据量远高于标记数据。

4)II 数据集(UNSW-TONIOT)。UNSW-TONIOT(T)数据集发布于 2021 年，涵盖当前物联网设备使用的协议、标准和技术，进一步扩展了物联网设备的多样性和所考虑的攻击类型。实验台包含 7 种物联网传感器，如天气监测器、智能冰箱监测器、Modbus 传感器和 GPS 跟踪器。数据集涵盖了 9 种威胁，包括扫描攻击和 DoS 攻击等。同样，n_{TL}∶n_{TU} 的比率设定为1∶5，1∶10和1∶50。

互联网和物联网在协议类型、场景多样性和攻击方式上存在显著差异，给域自适应带来挑战。4 个数据集的详细信息如表 2 所列。互联网主要使用 TCP、HTTP 等协议，而物联网依赖 MQTT 等轻量级协议，导致数据特征表示以及特征分布之间存在巨大差异。物联网的场景多样，涉及智能家居、工业控制等复杂环境，攻击手段受设备和网络架构限制，表现出不同特征，即使相同的攻击类型，使用不同的攻击方式得到的特征也不完全相同。此外，物联网数据攻击样本稀缺，且标签标注困难，增加了跨域对齐的难度。

表 2　数据集的详细信息

Table 2　Detail information of the datasets

数据集类型	数据集名称	协议类型	场景	攻击类型
互联网	UNSW-NB15(N)	TCP，UDP，HTTP	单一场景	模糊攻击，分析攻击(包括端口扫描、垃圾邮件、HTML 文件渗透等)，后门攻击，DoS 攻击，利用攻击，通用攻击(针对所有分组密码的攻击)，侦查攻击，木马代码，蠕虫攻击
	CICIDS 2018(C)	TCP，HTTP，HTTPS	单一场景	暴力破解，DoS 攻击，心脏出血攻击，Web 攻击，渗透攻击，僵尸网络攻击，DDoS 攻击
物联网	UNSW-BOTIOT(B)	MQTT，HTTP	5 种场景	DoS 攻击，DDoS 攻击，侦查攻击，欺骗攻击
	UNSW-TONIOT(T)	MQTT，HTTP，UDP	7 种场景	扫描攻击，跨站脚本攻击，密码攻击，DoS 攻击，DDoS 攻击，注入攻击，后门攻击，中间人攻击，勒索软件攻击

为了便于在 NI 和 II 领域之间进行知识迁移和模型训练，本文从上述数据集中选择了 5 个共享类别，分别是良性流量、DoS 攻击、DDoS 攻击、侦查攻击和密码攻击。随着物联网设备的普及，DoS 和 DDoS 攻击尤其频繁，例如 Mirai 僵尸网络就曾导致大规模网络瘫痪；侦查攻击通常是物联网攻击的前置步骤，它通过扫描端口和分析流量收集设备信息，为后续的恶意操作铺路；密码攻击由于物联网设备普遍存在弱密码问题而更加高发，攻击者可以通过暴力破解轻松获取控制权。这些攻击手段在互联网和物联网设备中广泛存在，它们在 CICIDS 2018，UNSW-NB15，UNSW-BOTIOT 和 UNSW-TONIOT 数据集中的占比分别为 99.85%，54.2%，100% 和 77.03%[24]。因此，通过知识迁移，可以有效检测物联网领域中大多数的现代攻击。

这些数据集采用不同的特征提取工具，导致各自的特征存在异构性，这可能使源域与目标域之间的差距过大，难以实现有效的迁移。在实际应用场景中，通常需要对 PCAP 流量包进行特征提取。如果使用相同的工具来提取特征，可以消除$\mathcal{D}_S$与$\mathcal{D}_T$之间的特征异构性，促进$\mathcal{D}_S$到$\mathcal{D}_T$的迁移。因此，本文统一使用各数据集的 NetFlow 版本[29]进行实验。

4.2　实验相关参数和评价指标

本文基于 TensorFlow 框架实现了 TDAIID 模型，并在 Intel(R) Xeon(R) CPU E5-2620 v4 和 Nvidia GeForce RTX 2080 GPU 的环境下完成实验。TDAIID 模型的嵌入层使用单层全连接神经网络，嵌入维度为 64；编码层使用两层自注意力和交叉注意力网络；共享分类器的结构为两层全连接神经网络，两层维度分别是 128 和 64。根据经验以及实际实验结果对超参数进行设置，其中 $\alpha=0.1$，$\rho=0.1$，$\delta=0.01$，$\tau=0.1$。使用 Adam 梯度下降优化器训练 TDAIID 模型，学习率设置为 0.0001，迭代次数 epoch 设置为 250 轮次。本文将$\mathcal{D}_{TL}$数据按 4∶1 的比例划分为训练集和验证集，将对$\mathcal{D}_{TL}$中流量的攻击类别识别准确率和 F1 值作为评价指标来评估性能。

4.3　基线模型

本文使用 6 种先进的物联网入侵检测方法作为对比方法，来验证 TDAIID 模型的优越性。

1) MLP-T：使用多层感知机(Multilayer Perceptron，MLP)，仅使用$\mathcal{D}_{TL}$流量数据训练。

2) FlowTrans-S：使用 FlowTransformer[20] 模型，只使用$\mathcal{D}_S$流量数据进行训练。

3) FlowTrans-T：同样使用 FlowTransformer 模型，只使用$\mathcal{D}_{TL}$流量数据进行训练。

4) DDAC[16]：一种半监督域适应方法，通过对齐公共子空间中的边缘分布和条件分布，减少源域和目标域分歧；通过扩大不同类别中心之间的距离来提高类别能力；使用模型分类器结果作为伪标签参与训练。

5) MCL[30]：MCL 方法从域间、域内和样本间 3 个层面对齐源域和目标域，通过捕获不同层次上的信息进行互补，构建端到端的一致性学习框架。

6) SLA[31]：SLA 借鉴了标签校正的思想，将源域标签视为目标域分类的理想标签的噪声版本，通过动态清理源域标签噪声，使源域数据与目标域数据相匹配。

其中，MLP-T，FlowTrans-T 和 FlowTrans-S 方法仅使用$\mathcal{D}_S$或者$\mathcal{D}_{TL}$流量数据训练；DDAC，MCL 和 SLA 都是近几年具有代表性的半监督域自适应方法，针对$\mathcal{D}_T$中只有少数标记的数据，将知识从$\mathcal{D}_S$转移到$\mathcal{D}_T$，提高对$\mathcal{D}_{TL}$数据的分类识别效果。本文按照 1∶50 的比例将$\mathcal{D}_T$数据划分为$\mathcal{X}_{TL}$和$\mathcal{X}_{TU}$进行对比实验，并使用最高的结果作为最终结果。

4.4　实验结果与分析

为了更加直观地展示不同的 NI 源域$\mathcal{D}_S$和 II 目标域$\mathcal{D}_T$之间的差异，本文通过 PCA 降维技术将多维特征简化为一维特征，绘制核密度估计(Kernel Density Estimate，KDE)图来展示不同域之间 DoS 攻击类别的分布差异，如图 3 所示。图 3 中横轴表示降维后的新特征空间的某一主方向，这个方向保留了数据中最大的方差信息；纵轴表示特征分布的相对密度，即在某一横轴位置出现样本的相对可能性。从图 3 中可以看出，每个$\mathcal{D}_S$与$\mathcal{D}_T$的峰值位置不同，曲线密度存在明显偏移，说明$\mathcal{D}_S$与$\mathcal{D}_T$之间样本分布存在差异。其中，图 3(b)—图3(d)中$\mathcal{D}_S$与$\mathcal{D}_T$之间峰值差异明显且存在多峰情况，说明域之间分布差异明显且数据分布更加多样和复杂。图 3(a)中$\mathcal{D}_S$与$\mathcal{D}_T$之间主峰明显、峰值相近，说明域之间分布差异小，迁移更加容易。

为了验证 II 目标域中有标签数据和无标签数据在不同比率下 TDAIID 模型的有效性，特别是n_{TU}显著高于n_{TL}的极端情况，本文将n_{TL}∶n_{TU}的比率分别设置为 1∶5，1∶10 和 1∶50，其中 1∶50 的比率可以代表标签极其稀缺的场景。分别统计不同比率下 TDAIID 模型的准确率和 F1 值，实验结果如表 3 所列。从结果可以看出，TDAIID 模型在不同数据集、不同比率的情况下，仍然能保持较高的准确率和 F1 值。在 1∶50 的极端情况下，TDAIID 模型仍然表现出了较高的稳定性，与 1∶5的情况相比，平均准确率和平均 F1 值仅分别下降 0.71 个百分点和 0.85 个百分点。这证明了 TDAIID 模型在 II 目标域标签稀缺条件下的鲁棒性。

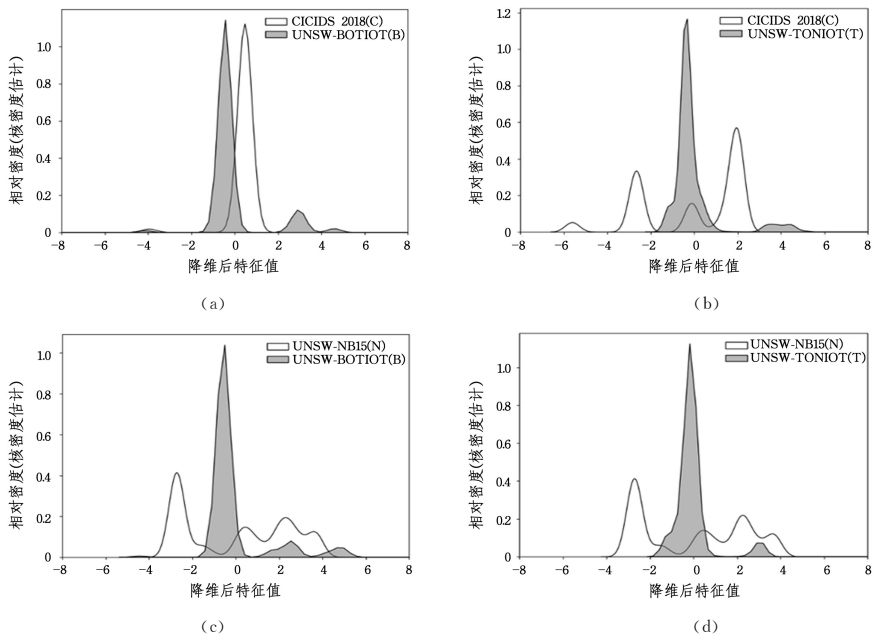

图 3 域间 DoS 攻击的特征分布差异

Fig. 3 Feature distribution differences of DoS attacks between NI domain and II domain

表 3 不同比率下 TDAIID 模型的性能

Table 3 Performance of the TDAIID model under different ratios

(%)

$n_{TL}:n_{TU}$	C→T		C→B		N→T		N→B		平均性能	
	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值
1:5	96.83	93.88	97.98	96.82	96.08	94.87	97.30	96.38	97.05	95.49
1:10	96.12	93.61	97.49	96.08	95.82	94.35	97.08	96.12	96.63	95.04
1:50	95.92	93.09	97.29	95.92	95.56	94.13	96.59	95.41	96.34	94.64

为了验证 TDAIID 模型相比于其他先进基线模型的优越性，比较不同方法在$n_{TL}:n_{TU}$比率为 1:50 场景下的性能。实验结果如表 4 所列，可以看出，TDAIID 在所有任务上都优于其他基线模型。相比于仅使用$\mathcal{X}_s$进行训练的 FlowTrans-S 模型，仅使用$\mathcal{X}_{TL}$进行训练的 MLP-T 和 FlowTrans-T 模型对$\mathcal{D}_T$的入侵检测性能有明显提升，说明$\mathcal{D}_s$与$\mathcal{D}_T$之间存在较大差异，不能将使用$\mathcal{D}_s$数据训练好的模型直接用于$\mathcal{D}_T$的入侵检测。与仅使用$\mathcal{X}_{TL}$训练的 MLP-T 和 FlowTrans-T 模型相比，TDAIID 模型的性能显著提高，说明仅使用$\mathcal{D}_{TL}$数据训练的模型发生了过拟合现象，导致整体性能下降。

表 4 1:50 比率场景下不同基线模型的性能

Table 4 Performance of different baseline models on 1:50 ratio scenario

(%)

$n_{TL}:n_{TU}$	C→T		C→B		N→T		N→B		平均性能	
	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值
MLP-T	76.71	76.04	86.88	86.87	82.15	81.93	81.85	81.38	81.90	81.56
FlowTrans-S	56.88	45.13	66.36	52.86	52.26	41.17	58.69	45.58	58.55	46.19
FlowTrans-T	79.77	71.83	89.26	86.91	80.67	73.64	88.17	83.92	84.47	79.08
DDAC	82.71	81.43	90.11	89.73	80.21	80.03	83.14	82.85	84.04	83.51
MCL	82.90	82.75	94.00	93.90	78.99	78.36	92.17	92.19	87.02	86.80
SLA	92.50	92.44	94.95	94.95	90.58	90.45	92.77	92.79	92.70	92.66
TDAIID(Ours)	**95.92**	**93.09**	**97.29**	**95.92**	**95.56**	**94.13**	**96.59**	**95.41**	**96.34**	**94.64**

如表 4 所列，在 4 个物联网入侵检测任务上，TDAIID 的性能均高于其他半监督域自适应基线模型(DDAC，MLC，SLA)，比性能次优的 SLA 模型在平均准确率和平均 F1 值上分别提高了 3.64 个百分点和 1.98 个百分点。原因在于 DDAC，MCL 和 SLA 方法或直接使用模型的预测结果为X_{TU}分配伪标签，或仅考虑从几何角度为X_{TU}分配伪标签，忽视了模型本身的预测结果，使伪标签标记准确率受到影响，进而产生噪声干扰。此外，这 3 种方法考虑均未使用交叉注意力机制，无法从样本级角度细粒度对齐和学习$\mathcal{X}_s$和$\mathcal{X}_t$之间的相似特征。因此，这些方法在分布差异较大的$\mathcal{D}_s$与$\mathcal{D}_T$之间的迁移效果更差，如 C→T，N→T 和 N→T 的场景。4.5 节的消融实验结果也验证了以上提升对入侵检测性能的作用。

4.5 消融实验

在进行了整体的性能评估之后，为了验证 TDAIID 中每个模块对于提高模型检测能力的效果，本文基于相同实验条

件，通过移除完整 TDAIID 模型中的单个模块来进行消融实验，结果如表 5 所列。其中，No Domain 表示无全局域间对齐模块；No Category 表示无局部类别间对齐模块；No Cross Atten 表示无交叉注意力机制对齐模块；No Center Aware 表示仅通过模型预测分配为标签，不使用动态中心感知伪标签算法；All 表示完整的 TDAIID 模型。

从实验结果可以看出，完整模型的入侵检测性能优于所有的消融模型，这验证了各个模块对于入侵检测性能的提升都是有帮助的。具体而言，在移除交叉注意力机制模块后，模型的性能下降最明显，平均准确率和平均 F1 值分别下降2.04 个百分点和 2.80 个百分点，这验证了交叉注意力机制对于语义特征对齐和融合的有效性。其次，在移除动态中心感知伪标签算法后，模型平均准确率和平均 F1 值分别下降 1.22 个百分点和 1.54 个百分点，这验证了伪标签标记算法对 II 目标域中无标记数据标记准确率的提升。最后，在移除域间对齐或类间对齐后，模型平均准确率和平均 F1 值分别下降约 0.7 个百分点和 1.1 个百分点，这验证了多重几何语义对齐的必要性。

表 5　1∶50 比率下 TDAIID 中不同模块对入侵检测性能的影响

Table 5　Impact of different components in TDAIID on intrusion detection performance at 1∶50 ratio

(%)

消融模型	C→T		C→B		N→T		N→B		平均性能	
	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值	准确率	*F*1 值
No Domain	95.13	92.10	96.29	94.63	94.90	92.55	96.20	94.87	95.63	93.54
No Category	95.03	91.99	96.11	94.19	95.12	93.30	96.12	94.76	95.59	93.56
No Cross Atten	94.23	90.76	96.08	93.74	92.85	90.40	94.06	92.46	94.30	91.84
No Center Aware	94.95	91.79	96.93	95.38	93.77	92.16	94.82	93.06	95.12	93.10
All	**95.92**	**93.09**	**97.29**	**95.92**	**95.56**	**94.13**	**96.59**	**95.41**	**96.34**	**94.64**

结束语　本文通过在基于流数据的 Transformer 模型中引入交叉注意力机制，提出了一种基于 Transformer 的域自适应物联网入侵检测模型 TDAIID。该模型通过联合利用 3 个级别的语义对齐机制，学习具有细粒度知识和高区分度的域不变特征表示，提升 II 目标域入侵检测性能；通过提出的基于几何距离的动态中心感知伪标签算法，提高伪标签分配的准确度。本文在 4 个常用的入侵检测领域数据集上进行多组实验，实验结果表明了 TDAIID 模型的先进性。

由于本文的场景设定 II 目标域有少量有标记标签，因此有一定的局限性。未来计划在 II 目标域完全无标签的场景下，探索更加鲁棒的无监督域自适应方法，为数据稀缺的物联网环境提供更有效的解决方案。

参 考 文 献

[1] LU Z, XU H, PAN J. Study on Intrusion Detection in IoTs Environment Based on GAN&CNN[J]. Chinese Journal of Sensors and Actuators, 2025, 38(10): 1853-1861.

[2] ZHAO J, JIANG W. IoT Intrusion Detection Model Integrating Improved TCN and DRSN[J]. Journal of Chinese Computer Systems, 2025, 46(2): 474-481.

[3] XIE Y, LIU L. RFLE Algorithm Based on the Internet of Things Intrusion Detection Model[J]. Journal of Air & Space Early Warning Research, 2025, 39(3): 203-208.

[4] ELHADJ B, THOMAS W, WALAA H. A Critical Review of Practices and Challenges in Intrusion Detection Systems for IoT: Toward Universal and Resilient Systems[J]. IEEE Communications Surveys & Tutorials, 2018, 20(4): 3496-3509.

[5] PAN S, YANG Q. A Survey on Transfer Learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2010, 22(10): 1345-1359.

[6] ZHUANG D, QI Z, DUAN K, et al. A Comprehensive Survey on Transfer Learning[J]. Proceedings of the IEEE, 2021, 109(1): 43-76.

[7] LY V, QUANG U N, DIEP N N, et al. Deep Transfer Learning for IoT Attack Detection[J], IEEE Access, 2020, 8: 107335-107344.

[8] PENG Y, CHEN X, CHEN S, et al. Cross-Domain Anomalous Traffic Detection Based on Transfer Learning[J]. Journal of Beijing University of Posts and Telecommunications. 2021, 44(2): 33-39.

[9] CHRISTIAN D, RAPHAEL L C, JESSICA S, et al. IoT-Botnet Detection and Isolation by Access Routers[C]// 9th International Conference on the Network of the Future. 2018: 88-95.

[10] DOUGLAS H S, KENNETH M Z, CHEN Y. Ultra-lightweight Deep Packet Anomaly Detection for Internet of Things Devices [C]// IEEE 34th International Performance Computing and Communications Conference. 2015: 1-8.

[11] VALERIAN R, PEDRO M S S, ALBERTO H C, et al. Federated Learning for Malware Detection in IoT Devices[J]. Computer Networks, 2022, 204: 108693.

[12] MOJTABA E, ZAFFAR H J, MASSIMO V, et al. Passban IDS: An Intelligent Anomaly-based Intrusion Detection System for IoT Edge Devices[J]. IEEE Internet of Things Journal, 2020, 7(8): 6882-6897.

[13] SARUMATHI M, ABBAS J. A Lightweight Intrusion Detection for Sybil Attack under Mobile RPL in the Internet of Things [J]. IEEE Internet of Things Journal, 2020, 7(1): 379-388.

[14] LI Z, XU C, DENG K, et al. A Subspace-based Few-shot Intrusion Detection System for the Internet of Things[J]. Frontiers of Information Technology & Electronic Engineering, 2025, 26(6): 862-876.

[15] XIE B, LI S, LYU F, et al. A Collaborative Alignment Framework of Transferable Knowledge Extraction for Unsupervised Domain Adaptation[J]. IEEE Transactions on Knowledge and Data Engineering, 2023, 35(7): 6518-6533.

[16] YAO Y, ZHANG Y, LI X, et al. Discriminative Distribution Alignment: A Unified Framework for Heterogeneous Domain

Adaptation[J]. Pattern Recognition, 2020, 101: 107165.
[17] NING J, GUAN G, WANG Y, et al. Malware Traffic Classification Using Domain Adaptation and Ladder Network for Secure Industrial Internet of Things[J]. IEEE Internet of Things Journal, 2022, 9(18): 17058-17069.
[18] HU X, ZHU C, CHENG G, et al. A Deep Subdomain Adaptation Network with Attention Mechanism for Malware Variant Traffic Identification at an IoT Edge Gateway[J]. IEEE Internet of Things Journal, 2022, 10(5): 3814-3826.
[19] WU J, WANG Y, DAI H, et al. Adaptive Bi-Recommendation and Self-Improving Network for Heterogeneous Domain Adaptation-Assisted IoT Intrusion Detection[J]. IEEE Internet of Things Journal, 2023, 10(15): 13205-13220.
[20] LIAM D M, SIAMAK L, WAI W L, et al. FlowTransformer: A Transformer Framework for Flow-based Network Intrusion Detection Systems[J]. Expert Systems with Applications, 2023, 241: 122564.
[21] LI X, HOU Y, WANG P, et al. Trear: Transformer-Based RGB-D Egocentric Action Recognition[J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(1): 246-252.
[22] HU R, AMANPREET S. UniT: Multimodal Multitask Learning with a Unified Transformer[C]// 2021 IEEE/CVF International Conference on Computer Vision. 2021: 1419-1429.
[23] BHARATH B D, BENJAMIN K, RÉMI F, et al. DeepJDOT: Deep Joint Distribution Optimal Transport for Unsupervised Domain Adaptation[C]// 15th European Conference on Computer Vision. 2018: 467-483.
[24] WU J, WANG Y, XIE B, et al. Joint Semantic Transfer Network for IoT Intrusion Detection[J]. IEEE Internet of Things Journal, 2023, 10(4): 3368-3383.
[25] NOUR M, JILL S. UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)[C]// 2015 Military Communications and Information Systems Conference. 2015: 1-6.
[26] IMANS, ARASH H L, GHORBANI A, et al. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization[C]// 4th International Conference on Information Systems Security and Privacy. 2018: 108-116.
[27] NICKOLAOS K, NOUR M, ELENA S, et al. Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset[J]. Future Generation Computer Systems, 2019, 100: 779-796.
[28] BOOIJ T M, IRINA C, ERIK M, et al. ToN_IoT: The Role of Heterogeneity and the Need for Standardization of Features and Attack Types in IoT Network Intrusion Data Sets[J]. IEEE Internet of Things Journal, 2022, 9(1): 485-496.
[29] MOHANAD S, SIAMAK L, MARIUS P. Towards a Standard Feature Set for Network Intrusion Detection System Datasets[J]. Mobile Networks and Applications, 2022, 27: 357-370.
[30] YAN Z, WU Y, LI G, et al. Multi-level Consistency Learning for Semi-supervised Domain Adaptation[C]// 31th International Joint Conference on Artificial Intelligence. 2022: 1530-1536.
[31] YU Y, LIN H. Semi-Supervised Domain Adaptation with Source Label Adaptation[C]// 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 24100-24109.

ZHU Feng, born in 1987, Ph.D, assistant professor, master supervisor. His main research interests include cyberspace security, Internet of Things security and operating system security.

LI Peng, born in 1979, Ph.D, professor, Ph.D supervisor, is a member of CCF (No. 48573M). His main research interests include computer communication networks, clouding computing and information security.

(责任编辑:何杨)